Turning repulsive point processes into subsampling algorithms

Rémi Bardenet

UNIVERSITÉ DE LILLE, FRANCE

DOCTORAL SCHOOL MADIS

Mathematics, digital sciences, and their interactions

Habilitation thesis

Domain: Automatique, Génie Informatique,
Traitement du Signal et Image

prepared by

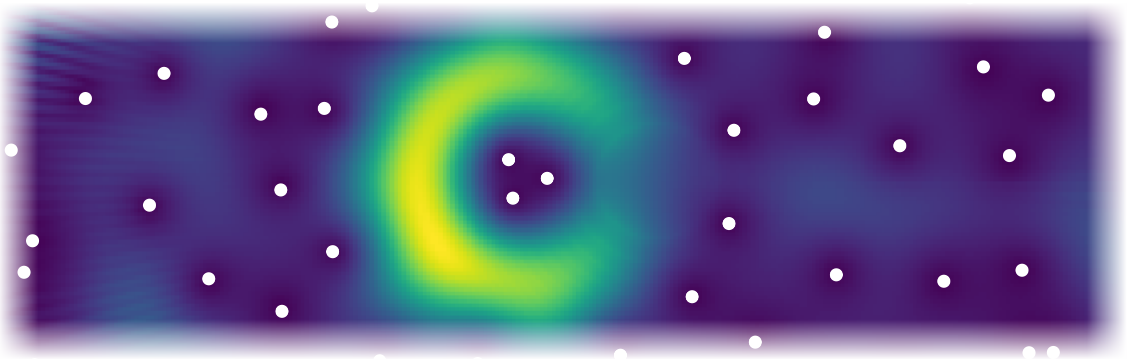
Rémi Bardenet

Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France

Turning repulsive point processes into subsampling algorithms

defended on 12 December 2022

Reviewers	Patrick Flandrin	CNRS & ÉNS Lyon, France
	Arthur Gretton	University College London, UK
	Judith Rousseau	University of Oxford, UK
President	Pierre-Olivier Amblard	CNRS & Univ. Grenoble-Alpes, France
Examiners	Peter Grünwald	CWI Amsterdam, Netherlands
	Joaquim Ortega Cerdà	Universitat de Barcelona, Spain
Guarantor	Pierre Chainais	École Centrale de Lille, France



Acknowledgments

> *'Rain don't fall on a witch if she doesn't want it to, although personally I prefer to get wet and be thankful.' 'Thankful for what?' said Tiffany. 'That I'll get dry later.'*
>
> Terry Pratchett, *A hat full of sky*, 2004

My first thanks go to my reviewers Patrick Flandrin, Arthur Gretton, and Judith Rousseau, for accepting to referee this manuscript, to Pierre-Olivier Amblard, Peter Grünwald, and Joaquim Ortega Cerdà for joining the jury, and to Pierre Chainais for his role as internal referent. It is an honour to have you all around me for this rite of passage.

One of my favourite aspects of academia is its collective nature, with people across the globe teaming up to tackle universal problems. I would thus like to thank all my past and current collaborators, in Lille and around the world. To cite only the most recent, thank you Mylène, Subhro, Meixia, Raphaël (×2), Adam, Pascal, Benjamin, Grégory, Chris, Pierre, Julien, Adrien, Michal. A special thought goes to my past, current, and aspiring PhD students Guillaume, Ayoub, Diala, Yoann, Martin, and Hugo, to my postdocs Arnaud, Barbara, Alexandre, Michaël, Xiaoyi, Juan Manuel, and Thibaut, and to my one past research engineer Guillaume again, who gets to be thanked twice in one sentence. Thanks for all your hard work and the stimulating discussions, on science, on problems of the world, or on the wacky random topics on which we like to debate in SigMA. Where else would I have learnt that you can drown in space, or that 18th century Moroccan monarch Mulay Ismail is allegedly second only to Genghis Khan in the number of children he fathered (868+). I would like to see the face of the future historian who will look into Arnaud's browsing history.

My gratitude extends to the whole SigMA team, among whom Pierre and Patrick stand out as mentors. Working among such friendly and knowledgeable people is a luxury. Beyond SigMA, the CRIStAL lab has been a great work environment, and many interesting discussions have been had with Guille, Charles, Pierre, Rayan, Laurentiu, Odalric, Émilie, Alessandro, Aurélien, and others. Our Esprit building only misses a music room and a rooftop with a barista, but hopefully a suitable commission will make sure we get these at some point. Thanks also to École Centrale, who hosted our team until we had our new building. They did not have a rooftop coffee bar either, but they actually had a music room, even if getting the keys was tricky and the admin department right above were complaining. A big thank you also goes to all the people running structures such as CRIStAL, École Centrale, Université de Lille, and DR18, in particular to all the staff who have been helping with my grants, Fatima, Christine, Myriam, Sonia, and Dominique; I know it is not always easy to administer researchers. I also have kind thoughts for my colleagues from the maths lab Painlevé; long live our joint point process workgroups!

At the national level, I would like to thank INS2I for their support throughout my early career, and all the scientific communities to which I belong, in particular people in stats, signal, stochastic geometry, or physics, around GdRs like ISIS and GeoSto. Interactions with so many smart and friendly people is also what contributes to a researcher's happiness.

Many people above were already or have become friends, some even godfathers to my kids. This brings me to the source of my happiness, namely my family, with a special mention for the three people who share my life. First my two wonderful kids: Aliénor, who regularly offers to add some color whenever she sees a point configuration drawn on a draft at home, and Basile, who believes that my job consists in drinking coffee with friends. And, last but not least, my beloved Amélie, who has been continuously supporting me. I hope that she can forgive me for accidentally putting (another) one of her fragile sweaters in the washing machine yesterday evening.

Finally, to keep things relatively short and because I present in the manuscript my activity while in Lille, I have focused here on thanking people related to that period of time. Yet, I certainly do not forget all the wonderful people met in Oxford, and before that in Orsay, to whom I send my warm thoughts if they come across this page.

Contents

1 Introduction **1**
1.1 Notation 4

2 Repulsive point processes **5**
2.1 Correlation functions 6
2.2 Determinantal point processes (DPPs) 7
2.2.1 Projection DPPs 9
2.2.2 A geometric chain rule for projection DPPs 11
2.2.3 DPPs with Hermitian kernels 12
2.2.4 Sampling DPPs 15
2.2.5 DPPs are (usually) repulsive 15
2.3 Zeros of Gaussian analytic functions 16
2.3.1 A constructive definition of GAFs 16
2.3.2 Three canonical GAFs 17

3 Monte Carlo integration with determinantal point processes **19**
3.1 A central limit theorem from random matrix theory 20
3.2 From random matrices to numerical integration 22
3.2.1 Multivariate orthogonal polynomial ensembles 22
3.2.2 A stochastic Gaussian quadrature 22
3.2.3 Our generalization of Johansson's CLT 23
3.2.4 An importance sampling version 24
3.3 Soviets already did it 25
3.4 Interpolation and quadrature in RKHSs 26
3.4.1 Reproducing kernel Hilbert spaces 27
3.4.2 Quadrature in a RKHS with a projection DPP 28
3.4.3 Quadrature in an RKHS with continuous volume sampling 29
3.5 Closing remarks 30

4 Feature selection with determinantal point processes **33**
4.1 Apples and oranges: column subsets and PCA 34
4.2 Multinomial sampling with and without SVD 35
4.3 Repulsive sampling needs fewer columns 36
4.4 DPPs take the best of leverage scores and repulsiveness 37
4.5 A guarantee on the excess risk in sketched linear regression 39
4.6 Closing remarks 41

5 Time-frequency filtering with zeros of GAFs **43**
5.1 From the STFT to the planar GAF 45
5.2 From a DPP to analytic wavelets 47
5.2.1 The space of signals 48
5.2.2 The time-scale transform 49
5.2.3 The white noise 49
5.3 Mapping signals to GAFs: a generic template 50
5.4 From the spherical GAF to the Kravchuk transform 52
5.5 Opening remarks 54
5.5.1 Removing the analyticity constraint 54
5.5.2 Making the Kravchuk transform a practical tool 54
5.5.3 The point process of maxima 55
5.5.4 Ridges and level sets 55
5.6 Despite appearances, this chapter fits in my narrative 55

6 Other work and perspectives **57**
6.1 Past work on repulsive point processes 57
6.2 Other work and a few perspectives 59

Appendices 61

A Teaching, supervision, grant management **63**
A.1 Teaching 63
A.2 Supervision 64
A.2.1 Postdocs 64
A.2.2 PhD students 65
A.3 Grant management 65

Bibliography **67**

Chapter 1
Introduction

> *Million-to-one chances... crop up nine times out of ten.*
>
> Terry Pratchett, *Equal Rites*, 1987

This document is the manuscript of my *habilitation à diriger des recherches (HDR)*, described in the official texts[1] as *validating the high scientific level of the candidate, the originality of his scientific programme, his ability to master a research strategy in a sufficiently vast scientific or technological domain, and his ability to supervise young researchers.* The typical form of the manuscript is a high-level summary of one's work after the PhD, emphasizing the research programme and directions, and I will follow that template.

My thesis at the departments of computer science (LRI) and particle physics (LAL) of Univ. Paris-Sud[2] dates from December 2012. After that, I spent two years as a postdoc at the department of statistics of the University of Oxford, working on Monte Carlo methodology for big data (Bardenet, Doucet, and Holmes, 2014, 2017a), as well as Bayesian inference for cardiac electrophysiology (Johnstone et al., 2016a,b; Beattie et al., 2018; Ridder et al., 2020). In early 2015, I crossed back the Channel, and started as a CNRS junior permanent researcher (CR) at CRIStAL (short for *research centre on computer science, signal processing, and automatic control of Lille*). The major part of my work since 2015 has dealt with using repulsive point processes as computational tools, on which I now focus.

Many tasks in statistics, machine learning, or signal processing, can be cast as sampling, i.e., summarizing a large (possibly infinite) collection of items by a finite subset of these items. Variable selection in regression, for instance, aims at selecting a few representative features to reduce the dimensionality of a dataset. Monte Carlo integration aims at approximating integrals over a measure space by weighted sums of evaluations of the integrand at a handful of points in the support of the target measure. Finally, digital signal processing requires first to encode a square-integrable function by its orthogonal projections onto a finite number of reference functions. In all these examples, random sampling, i.e., drawing at random either features, quadrature nodes, or the reference functions on which to project, has a role to play. Loosely said, one motivation for *random* instead of *deterministic* sampling is to obtain simpler guarantees on the performance of the algorithm of interest under weaker assumptions. The output of standard Markov chain Monte

[1] https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000298904/

[2] Now Paris-Saclay.

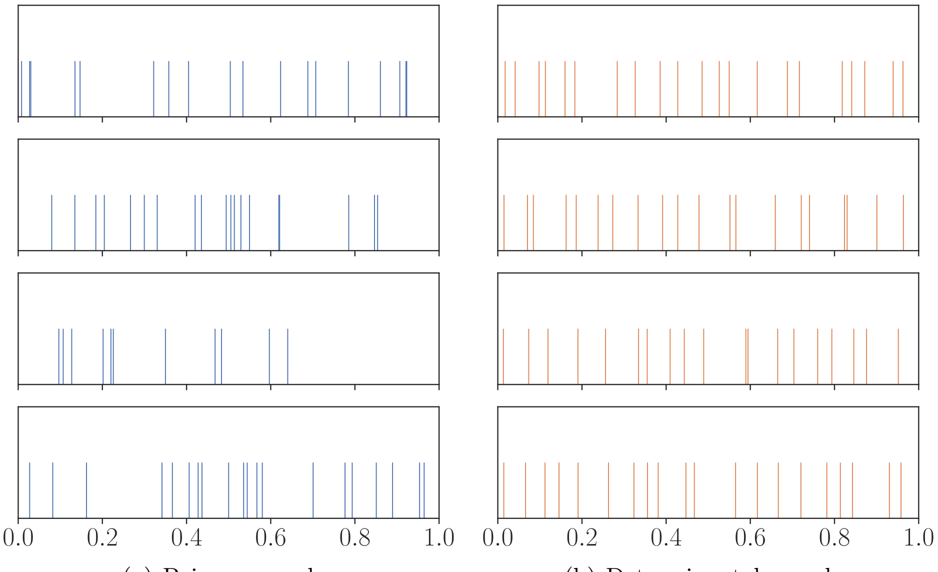
Carlo algorithms, for instance, often satisfies a central limit theorem, with the classical Monte Carlo rate, as soon as the integrand is square-integrable. In contrast, deterministic quadrature guarantees in large dimension typically involve very smooth integrands and dimension-dependent rates; see e.g. quasi-Monte Carlo methods (Dick and Pillichshammer, 2010). It is also common that, for performance guarantees of the same order, the best known randomized algorithms are computationally cheaper than their deterministic counterparts. For instance, we shall see in Chapter 4 that for feature selection with both Frobenius and spectral guarantees, we do not know of a better deterministic algorithm yet than enumerating all subsets of columns of a (fat) matrix, when randomized feature selection algorithms can simultaneously provide both types of guarantees in polynomial time. That is not to say that randomized algorithms are necessarily the final answer to every, and actually any, problem. Whether randomness actually helps is actually a deep and open question in theoretical computer science (Impagliazzo, 2002). Moreover, it is sometimes possible to de-randomize a randomized algorithm, while transferring the bounds obtained with probabilistic tools to the obtained deterministic algorithm; see (Deshpande and Rademacher, 2010) for an example in column subset selection. To put it briefly, randomized sampling algorithms are a useful object of study, whether you accept them as a solution to your problem or only as an intermediate step.

If you agree with me that random sampling is an interesting research area,[3] then my objective in this manuscript is to convince you that *negatively dependent* random sampling is also a promising one. In essence, drawing items from a large set *independently* leaves holes, and the corresponding sample looks far from anything regular. One illustrative example is shown in Figure 1.1. The left panel shows four draws from a homogenous Poisson process on the real line, intersected with $[0, 1]$. Equivalently, each row is independently obtained by first drawing a Poisson random variable N with some large enough mean, and conditionally on N, drawing N locations independently from the uniform distribution on $[0, 1]$. In each row, it is striking how independent locations fail to fill the unit interval evenly. If we evaluated a function $f : [0, 1] \to \mathbb{R}$ at the samples shown in any of the rows of this left panel, the set of evaluations would poorly represent f, in that any oscillation of f in the observed holes would go unnoticed. We argue that by introducing dependence between the sampled items, and actually *negative* dependence or repulsiveness, we can obtain random samples that represent the initial set of items much more accurately. For instance, the right panel of Figure 1.1 shows samples of a translation-invariant determinantal point process.[4] We have been able to make more precise points on how negative dependence can solve sampling tasks at polynomial expense.

In Monte Carlo integration, we have shown that a specific class of negatively dependent yet computationally tractable distributions, called *determinantal point processes*, can provide central limit theorems with faster-than-square-root rates (Bardenet and Hardy, 2020). We have even shown arbitrarily fast rates of decay for the mean square error of a related estimator under additional assumptions on the integrand (Belhadji et al., 2019, 2020a). Moreover, determinantal point processes, while providing tunable repulsiveness,

[3] And if you don't, I hope the Pratchett quote opening this chapter tips the balance in my favour.

[4] Actually approximate samples of the sine process; see Example 2



(a) Poisson samples

(b) Determinantal samples

Figure 1.1: The intersection with $[0, 1]$ of independent samples of (a) a homogenous Poisson point process, and (b) a determinantal point process, more precisely a scaled version of the circular unitary ensemble to approximate the sine process; see Chapter 2 for details.

can be sampled in polynomial time, which has made them a workhorse in my research programme. In feature selection for linear regression, we have shown that, again with a carefully designed determinantal point process, one could obtain a tight bound on the reconstruction error of a data matrix, as well as guarantees for linear regression based on the obtained column sample (Belhadji, Bardenet, and Chainais, 2020b). Finally, in signal processing, we have studied the repulsive point processes that appear as the zeros of specific random analytic functions, which result from applying time-frequency transforms to white noise (Bardenet, Flamant, and Chainais, 2018; Bardenet and Hardy, 2020). One of my objectives in this manuscript, is to present these selected works under a common umbrella, namely random sampling with negative dependence.

In Chapter 2, I present some background material on repulsive point processes, i.e., random configurations of items in which similar or close-by items have little chance of simultaneously appearing. My flagship examples will be *determinantal* point processes and zeros of Gaussian analytic functions. My first own contributions appear in Chapter 3, where I discuss our results on Monte Carlo integration with determinantal point processes. In Chapter 4, I develop our results on column-subset selection and variable selection in linear regression. In Chapter 5, I explain our results on the zeros of time-frequency transforms of white noise, and discuss how they fit in my narrative of negatively dependent sampling. Each of Chapters 3 to 5 ends with chapter-specific open questions. I conclude in Chapter 6 by taking a step back and presenting higher-level open questions, as well as shortly discussing the part of my work since 2013 that I chose to omit in the manuscript for conciseness. Finally, Appendix A contains a history of my teaching, supervision, and grant management since 2013.[5]

1.1 Notation

This section is here for reference.

We shall constantly work with a complete metric space $\mathbb{X}$, equipped with μ a Borel measure on the Borel sets of $\mathbb{X}$. Our point processes will be random subsets of that space $\mathbb{X}$. For our purpose, it is enough to think of either $\mathbb{X} \subset \mathbb{R}^d$ equipped with the Lebesgue measure, or $\mathbb{X} = [n] := \{1, \dots, n\}$ a finite set, equipped with its counting measure.

We shall also often consider submatrices. We write matrices in bold font ($\mathbf{K}$, $\mathbf{L}$, etc.). For a matrix $\mathbf{K}$, a subset A of $|A| = a$ indices of rows of $\mathbf{K}$, and a subset B of $|B| = b$ indices of columns of $\mathbf{K}$, we write $\mathbf{K}_{A,B}$ for the $a \times b$ submatrix of $\mathbf{K}$ formed by the rows indexed by A and the columns indexed by B. We also write $\mathbf{K}_A$ instead of $\mathbf{K}_{A,A}$, and $\mathbf{K}_{:,A}$ when we keep all rows but only the columns indexed by A.

Chapter-specific notation will be introduced on-the-fly. Note that because Chapters 3, 4, and 5 correspond to work published in different communities, some conventional notations are bound to conflict. For instance, k is the kernel of an RKHS in Chapter 3, while in Chapter 4, $k \in \mathbb{N}$ is the dimension of the reference principal subspace. These notations are so classical in the respective literatures that I refrained from changing, hoping that the context will keep it clear what k is in which chapter.

[5] For a hierarchical list of publications, a more detailed track record, etc. please refer to my CV at http://rbardenet.github.io/pdf/cvFull_bardenet.pdf.

Chapter 2

Repulsive point processes

> *'Y'know,' he said, 'it's very hard to talk quantum using a language originally designed to tell other monkeys where the ripe fruit is.'*
>
> Terry Pratchett, *Night Watch*, 2002

Contents

2.1 Correlation functions . . . 6

2.2 Determinantal point processes (DPPs) . . . 7

2.2.1 Projection DPPs . . . 9

2.2.2 A geometric chain rule for projection DPPs . . . 11

2.2.3 DPPs with Hermitian kernels . . . 12

2.2.4 Sampling DPPs . . . 15

2.2.5 DPPs are (usually) repulsive . . . 15

2.3 Zeros of Gaussian analytic functions . . . 16

2.3.1 A constructive definition of GAFs . . . 16

2.3.2 Three canonical GAFs . . . 17

Let us fix a complete metric space $\mathbb{X}$, with μ a Borel measure on the Borel sets of $\mathbb{X}$. For this manuscript, it is enough to bear in mind the two cases of $\mathbb{X} \subset \mathbb{R}^d$ equipped with the Lebesgue measure, and a finite set $\mathbb{X}$ equipped with its counting measure. Define the space of configurations as

$$\operatorname{Conf}(\mathbb{X}) := \Big\{ S \subset \mathbb{X} : \ |S \cap A| < \infty \text{ for all compact } A \subset \mathbb{X} \Big\},$$

where $|B|$ stands for the cardinality of set B. In other words, configurations are locally finite subsets of $\mathbb{X}$.

Definition 1 *A point process S on $\mathbb{X}$ is a random configuration*[1] *of points in $\mathbb{X}$.*

[1] *Stricto sensu*, we are defining here *simple* point processes, i.e., such that with probability 1, S does not contain a given point of $\mathbb{X}$ more than once. Since all point processes in this work are simple, we take this shortcut and identify a (simple) point process with a random configuration.

2.1 Correlation functions

Correlation functions are a common way of defining and studying point processes. For any $k \geqslant 1$, the k-point correlation function $\rho_k : \mathbb{X}^k \to [0, \infty]$ satisfies, when it exists,

$$\mathbb{E}\left[\sum_{\substack{x_1,\dots,x_k \in S \\ x_i \neq x_j \text{ if } i \neq j}} f(x_1, \dots, x_k)\right] = \int_{\mathbb{X}^k} f(x_1, \dots, x_k)\, \rho_k(x_1, \dots, x_k) \prod_{i=1}^{k} \mathrm{d}\mu(x_i) \tag{2.1}$$

for any bounded (or positive) and measurable test function f. Here

$$\mathbb{X}^k = \overbrace{\mathbb{X} \times \cdots \times \mathbb{X}}^{k}$$

and the symbol $\mathbb{E}$ in (2.1) stands for the expectation under the law $\mathbb{P}$ of the random variable S. Thus, the k-point correlation function ρ_k encodes the distribution of k-tuples of points from S. Indeed, an informal rewriting of (2.1) reads

$$\rho_k(x_1, \dots, x_k)\mathrm{d}\mu(x_1) \dots \mathrm{d}\mu(x_k) = \mathbb{P}\begin{pmatrix} \text{There are at least } k \text{ points in } S, \\ \text{one around each of } x_1, \dots, x_k. \end{pmatrix}. \tag{2.2}$$

It is common to actually define a point process S by a sequence (ρ_k) of *compatible* correlation functions; see (Daley and Vere-Jones, 2003). By *compatible*, we mean that not every sequence (ρ_k) actually defines a point process, and that usually a mathematical argument for existence is necessary.

Of particular practical significance in the description of point processes are the one- and two-point correlation functions. The one-point correlation function ρ_1 describes marginal information about the particles, and is called the *intensity* of the point process. In particular, (2.2) applied to $f = \mathbf{1}_A$ implies that the average number of points falling in A is $\int_A \rho_1 \mathrm{d}\mu$. The second correlation function ρ_2 describes pairwise interactions, and is often discussed in its normalized form

$$g(x, y) = \frac{\rho_2(x, y)}{\rho_1(x)\rho_1(y)}. \tag{2.3}$$

Finally, when $\mathbb{X} = \mathbb{R}^d$ and the distribution of S is translation- and rotation-invariant, g only depends on $r = \|x - y\|$, so that we write $g(x, y) = g_0(r)$. The function g_0 is called the *pair correlation function* of S.

Example 1 *Let $\lambda : \mathbb{X} \to \mathbb{R}_+$ be locally integrable, that is, $\int_B \lambda(x)\mathrm{d}\mu(x) < \infty$ for every bounded $B \subset \mathbb{X}$. Further assume for simplicity that the measure $\lambda \mathrm{d}\mu$ has no atom.*[2] *The point process with correlation functions*

$$\rho_k(x_1, \dots, x_k) = \lambda(x_1) \dots \lambda(x_k), \quad k \geqslant 1, \tag{2.4}$$

always exists and is called the Poisson point process *with parameter function λ.*

[2] Without this assumption, a Poisson point process would not necessarily be a *simple* point process.

For a Poisson process with parameter function λ, the one-point correlation function is by definition $\rho_1 = \lambda$. The pair correlation function (2.3) is constant, $g(x, y) = 1$, which indicates that the relative positions of x and y do not alter the marginal probability of seeing them co-occur. In a sense, the factorized form of (2.4) indicates the lowest level of correlation among points.

When $g(x, y) > 1$, Equation (2.2) indicates that pairs are more likely to occur around (x, y) than under a Poisson process with the same intensity. Similarly $g(x, y) < 1$ indicates that pairs are less likely to occur. Consequently, one way to define a *repulsive* point process[3] is to require that $g(x, y) < 1$ for all x, y.

Figure 2.1 shows samples of three translation-invariant point processes in $\mathbb{R}$, along with their pair correlation functions. The three point processes share the same (constant) one-point correlation function, and the left panel corresponds to a Poisson point process. Note how the pair correlation function in Figure 2.1(f) keep below 1, and even close to zero around zero, indicating fewer small pairwise distances than the reference Poisson point process in Figure 2.1(d). This is a sign of a very regular, more grid-like distribution of the points, as seen in Figure 2.1(c). On the contrary, the pair correlation function of Figure 2.1(e) shows more small pairwise distances than the reference Poisson process: points are lumped together, as confirmed by Figure 2.1(b). The latter process is actually a permanental point process, and was introduced by Macchi (2017) is her 1972 thesis.[4] It is a Poisson process with a random (smooth) parameter function, which is why we see lumps. We refer the reader interested in permanental point processes to our recent survey on point processes in quantum optics (Bardenet et al., 2022).

We now define a flexible class of repulsive point process, of which the right-hand panel of Figure 2.1 shows an instance: determinantal point processes.

2.2 Determinantal point processes (DPPs)

DPPs were also introduced by Macchi (2017) in her 1972 thesis, and first featured in (Bénard and Macchi, 1973; Macchi, 1975), this time to model detection times in quantum beams of fermions.

Definition 2 *A point process is determinantal if there exists $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ or $\mathbb{C}$ such that the n-point correlation function* (2.2) *exists for every n and reads*

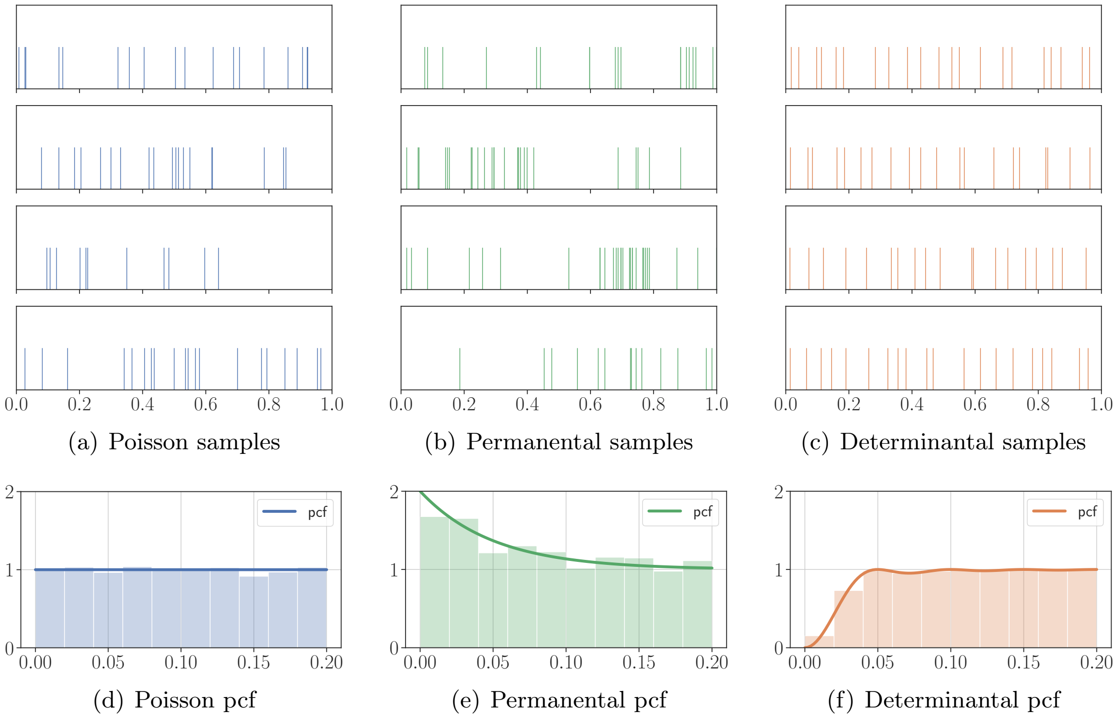
$$\rho_n(x_1, \dots, x_n) = \det \Big[K(x_k, x_\ell) \Big]_{k,\ell=1}^{n}, \qquad x_1, \dots, x_n \in \mathbb{X}. \tag{2.5}$$

We say that K is the kernel of the DPP, μ its reference measure, and we denote the point process by $\mathrm{DPP}(K, \mu)$.

Just like a covariance kernel is enough to describe a zero-mean Gaussian process, the kernel K of a DPP thus encodes how the points in the random configurations interact.

[3] While we simply use $g < 1$ as a token of repulsiveness in this manuscript, much more can be said on quantifying repulsiveness in point processes (Biscio and Lavancier, 2016; Møller and O'Reilly, 2021).

[4] The date mismatch with the reference (Macchi, 2017) comes from the fact that the latter is a translated reprint.



(a) Poisson samples (b) Permanental samples (c) Determinantal samples

(d) Poisson pcf (e) Permanental pcf (f) Determinantal pcf

Figure 2.1: (a, b, c) The intersection with $W = [0, 1]$ of four independent samples of each of the following types of point processes: Poisson, permanental, and determinantal. All three point processes are translation-invariant and are scaled to have the same expected number of points falling in W. (d, e, f) The pair correlation function (2.3) of the three point processes, and histograms of pairwise distances from 100 independent samples.

However, the existence of a point process with (2.5) as its correlation functions is, in general, a difficult question. It is easy to see that the kernel has to be positive definite, so that the right-hand side of (2.5) is always non-negative. Henceforth, all DPP kernels are thus assumed to be positive definite. But non-negativity is not sufficient for (2.5) to consistently define a point process.

2.2.1 Projection DPPs

A canonical way to ensure existence of a DPP is to restrict oneself to so-called *projection DPPs*, which generate configurations of a constant number $N < \infty$ of points $\mathbb{P}$-almost surely, i.e. $S = \{x_1, \dots, x_N\}$. More precisely, consider N orthonormal functions $\varphi_0, \dots, \varphi_{N-1}$ in $L^2(\mu)$, that is, $\int \varphi_k(x)\varphi_\ell(x)\mu(\mathrm{d}x) = \delta_{k\ell}$, and take for kernel

$$K(x, y) = \sum_{k=0}^{N-1} \varphi_k(x)\overline{\varphi_k(y)}. \tag{2.6}$$

In this setting, it turns out that the (permutation invariant) random variables $x_1, \dots, x_N$ with joint probability distribution

$$\frac{1}{N!} \det \Big[K(x_i, x_\ell)\Big]_{i,\ell=1}^{N} \times \prod_{i=1}^{N} \mathrm{d}\mu(x_i) \tag{2.7}$$

generate a DPP with kernel $K(x, y)$. (Hough et al., 2006, Section 2) gives an elementary proof that (2.7) satisfies (2.5). The resulting DPP is called a projection DPP, in reference to the fact that its kernel defines a projection operator.

Example 2 (CUE) *Take* $\mathbb{X} = [0, 1]$*,* μ *to be the uniform distribution, and the orthonormal functions* φ_ℓ *to be the Fourier basis* $\varphi_\ell(x) = \mathrm{e}^{2\mathrm{i}\pi\ell x}$*,* $\ell \geqslant 0$*. The projection DPP with kernel*

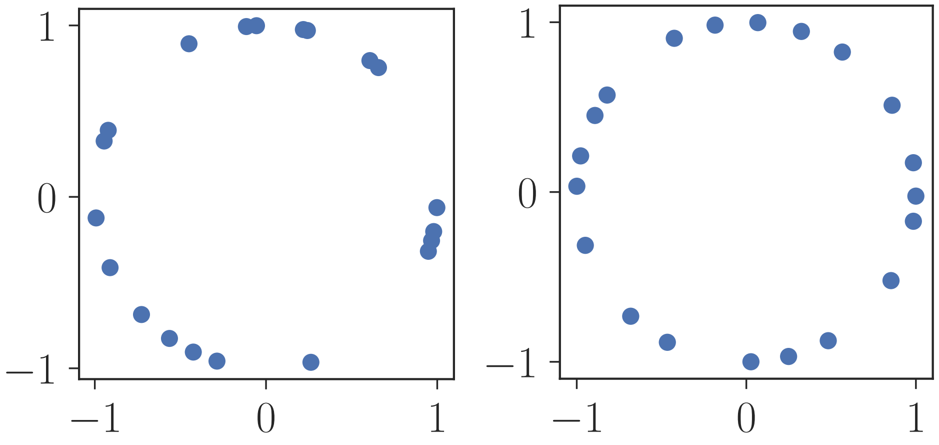
$$x, y \mapsto \sum_{\ell=0}^{N-1} \mathrm{e}^{2\mathrm{i}\pi\ell(x-y)} = \mathrm{e}^{\mathrm{i}\pi(N-1)(x-y)} \frac{\sin \pi N(x-y)}{\sin \pi(x-y)}$$

given by (2.6) *is called the circular unitary ensemble (CUE).*

Note that from the definition (2.5) of the correlation functions of a DPP, for any valid kernel and any non-vanishing function f, the kernel $x, y \mapsto f(x)K(x, y)/f(y)$ defines the same DPP as K. The kernel of the CUE in Example 2 can thus be simplified to

$$x, y \mapsto \frac{\sin \pi N(x-y)}{\sin \pi(x-y)}. \tag{2.8}$$

A CUE sample is shown in the right panel of Figure 2.2. Note how regularly spread the points of each realization of the CUE are on the circle. In comparison, i.i.d. draws from $\mu = \mathrm{d}x$ in the left panel seem less well organized, creating clusters and holes. Historically, the CUE was introduced as the law of the eigenvalues of a Haar-distributed random unitary matrix, as we shall recall in Section 3.1. Since the Haar distribution on unitary matrices



(a) Uniform (b) Haar

Figure 2.2: (a) An i.i.d. sample of the uniform distribution on the circle, compared to (b) a sample of the eigenvalues of a Haar-distributed unitary matrix. The arguments of the complex variables in (b) form a draw of the CUE.

is actually easy to sample, this connection with random matrices can actually be seen as a $\mathcal{O}(N^3)$ exact sampling algorithm. Thanks to a scaling argument, this also provides an approximate sampling algorithm for another well-known DPP, the sine process. Indeed, if $(\theta_i)_{i=1}^N \sim$ CUE, then the average spacing between the points of (θ_i) is of order $1/N$. Scaling the point process to increase this mean distance to 1, the random measure $\sum \delta_{N\theta_i}$ converges in distribution (a.k.a. weakly) to the DPP on $\mathbb{R}$ with kernel

$$K_{\text{sine}}(x, y) = \frac{\sin(x - y)}{x - y}, \tag{2.9}$$

a DPP called the sine process. Approximate samples obtained by scaling CUE samples are shown in Figure 2.1(c). The sine process is known for its universality: it appears as the scaling limit not only of the CUE, but of many DPPs. Another reason for its fame is its link with the zeros of the Riemann zeta function; see e.g. (Rudnick and Sarnak, 1996) and references therein. We refer to (Lambert, 2016, Section 3.4) for a survey of the properties of the sine process and precise statements on scaling limits of DPPs. We now switch to another reference DPP from random matrix theory.

Example 3 (Ginibre ensemble) *Take $\mathbb{X} = \mathbb{C}$, the reference measure μ to be the Gaussian measure $\pi^{-1}\mathrm{e}^{-|z|^2}$, and the orthonormal functions to be $\varphi_\ell(z) := (k!)^{-1/2}z^k$. The projection DPP with kernel* (2.6) *is called the Ginibre ensemble of size N.*

Like the CUE, the Ginibre ensemble actually rose to attention in random matrix theory, this time as the distribution of the eigenvalues of an $N \times N$ random matrix filled with i.i.d. unit complex Gaussians; see e.g. (Anderson et al., 2010). One can actually let $N \to \infty$ and define a limiting DPP, called the Ginibre ensemble, see (Hough et al., 2006, Example 15).

2.2.2 A geometric chain rule for projection DPPs

Besides their constant cardinality, one key property for computational applications of projection DPPs is that (2.7) can easily be rewritten as a chain rule, and this chain rule brings insight on the link between DPPs and geometry in Hilbert spaces.

We first illustrate our point by doing explicit computations when the rank of the kernel (2.6) is $N = 2$, and then discuss the chain rule for arbitrary N. Let thus $\varphi_0, \varphi_1 \in L^2(\mu)$ be two orthonormal functions. Consider the projection kernel (2.6), which simply reads

$$K(x, y) = \varphi_0(x)\overline{\varphi_0(y)} + \varphi_1(x)\overline{\varphi_1(y)}.$$

Samples from the corresponding DPP will thus have $N = 2$ points almost surely, drawn from (2.14), which becomes

$$\frac{1}{2}\det\begin{pmatrix} K(x,x) & K(x,y) \\ K(y,x) & K(y,y)\end{pmatrix} \mathrm{d}\mu(x)\mathrm{d}\mu(y) = \frac{1}{2}K(x,x)\mathrm{d}\mu(x) \times \left[K(y,y) - \frac{K(x,y)^2}{K(x,x)}\right]\mathrm{d}\mu(y) \tag{2.10}$$

Now, it is routine to check that both terms in the product in the RHS of (2.10) integrate to 1. This implies that (2.10) is a chain rule. In other words, to obtain $\{x, y\} \sim \mathrm{DPP}(K, \mu)$, it is enough to first draw $x \sim \frac{1}{2}K(x,x)\mathrm{d}\mu(x)$, and then

$$y|x \sim \left[K(y,y) - \frac{K(x,y)^2}{K(x,x)}\right]\mathrm{d}\mu(y).$$

As long as the two sampling steps can be implemented, e.g. using rejection sampling, this is already an interesting remark. But an important remark is that this chain rule is actually a simple area computation using the *base times height* formula. Indeed, consider the parallelotope formed in $\mathrm{Span}(\varphi_0, \varphi_1)$ by the so-called *feature functions* $K(x, \cdot)$ and $K(y, \cdot)$. The squared length of $K(x, \cdot)$ (the *base*) is

$$\begin{aligned} \int K(x,\cdot)\overline{K(x,\cdot)}\mathrm{d}\mu(x) &= |\varphi_0(x)|^2 \int |\varphi_0|^2 \mathrm{d}\mu + |\varphi_1(x)|^2 \int |\varphi_1|^2\mathrm{d}\mu & (2.11) \\ &= K(x,x), & (2.12) \end{aligned}$$

while the squared length of the projection of $K(y, \cdot)$ onto the orthogonal of $\mathrm{Span}(K(x, \cdot))$ (the *height*) is

$$K(y,y) - \left[\int K(y,\cdot)\frac{\overline{K(x,\cdot)}}{\sqrt{K(x,x)}}\mathrm{d}\mu(x)\right]^2 = \left[K(y,y) - \frac{K(x,y)^2}{K(x,x)}\right]. \tag{2.13}$$

From (2.12) and (2.13), it should now be clear that the chain rule (2.10) computes the squared volume of a parallelotope.

For a projection kernel K of rank $N \geqslant 1$, similar computations lead to an N-dimensional volume computation through a generalized "base times height” formula, namely

$$(2.7) = \prod_{i=1}^{N} \frac{1}{N-i+1} \left\| P_{H_{i-1}} K(x_i, \cdot)\right\|^2_{L^2(\mu)} \mathrm{d}\mu(x_i). \tag{2.14}$$

In (2.14), P_H is the orthogonal projection onto a subspace H of $L^2(\mu)$,

$$H_0 = \mathrm{Span}(\varphi_0, \dots, \varphi_{N-1}),$$

and for every $i > 1$, H_{i-1} is the orthocomplement in H_0 of

$$\mathrm{Span}\left(K(x_\ell, \cdot),\ 1 \leqslant \ell \leqslant i-1\right).$$

Like (2.10), (2.14) is a chain rule since each term in the product is a probability measure (Hough et al., 2006, Proposition 19).

Beyond sampling algorithms, we believe that the geometric interpretation of the chain rule (2.14) is paramount to understanding DPPs: the favoured point configurations in a projection DPP are those that correspond to feature vectors that span a large volume. Several contributions described in this manuscript come from understanding what feature vectors to choose in a particular problem involving linear structures, so that large volumes correspond to a small value of the problem's objective.

2.2.3 DPPs with Hermitian kernels

The projection kernels of Section 2.2.1 are a special case of Hermitian kernels, that is $K(x,y) = \overline{K(y,x)}$. For Hermitian kernels, the Macchi-Soshnikov theorem gives a necessary and sufficient condition for the existence of the associated DPP. The Hermiticity of the kernel guarantees that the integration operator $\Sigma : L^2(\mathrm{d}\mu) \to L^2(\mathrm{d}\mu)$ defined by

$$\Sigma : f \mapsto \int K(\cdot, y) f(y) \mathrm{d}\mu(y)$$

is self-adjoint.

Theorem 1 (Macchi-Soshnikov; Macchi, 1975; Soshnikov, 2000) *If Σ is self-adjoint and trace-class, then existence of the associated DPP is equivalent to all the eigenvalues of Σ lying is $[0,1]$.*

Being trace-class is a technical, but unrestrictive assumption, which guarantees in particular that Σ is compact, so that we can talk about its eigenvalues. Actually, a slight modification of Theorem 1 allows treating locally trace-class operators, i.e., such that $P_C \Sigma P_C$ is trace-class for any compact $C \subset \mathbb{X}$, with P_C the multiplication operator by the indicator function $\mathbf{1}_C$. Checking that a given operator is locally trace-class can be hard work, but there are useful criteria in the literature; see e.g. (Gohberg et al., 1990). For instance, since our kernels in this section are positive definite and Hermitian, a sufficient condition for being locally trace-class is that K is continuous and $\int_C K(x,x)\mathrm{d}\mu(x) < \infty$ for any compact $C \subset \mathbb{X}$. In particular, the sine kernel defined in (2.9) is locally trace-class. This is also the criterion used in spatial statistics by Lavancier et al. (2014), who show how covariance functions for Gaussian processes give rise to DPPs in $\mathbb{X} = \mathbb{R}^d$ using the Macchi-Soshnikov theorem. The constraint in Theorem 1 on the spectrum of the integration operator then further constrains the parameters of the covariance function. A simple such example is the Gaussian DPP.

Example 4 (Gaussian kernel; Lavancier et al., 2014) *Take μ to be the Lebesgue measure on $\mathbb{X} = \mathbb{R}^d$ and*

$$K(x,y) = \alpha \exp(-\|x/\ell\|^2).$$

Provided the parameters are restricted to $\alpha \leqslant (\sqrt{\pi}\ell)^{-d}$, $\mathrm{DPP}(K,\mu)$ exists (Lavancier et al., 2014, Theorem 2.5).

It is interesting to note that the locally trace-class condition guarantees that the average number of points that fall in a bounded $A \subset \mathbb{X}$ is finite, since the definition of the one-point correlation function of a DPP leads to

$$\mathbb{E}\left[\sum_{x\in S} \mathbf{1}_A(x)\right] = \int_A K(x,x)\mathrm{d}\mu(x) < \infty. \tag{2.15}$$

In particular, for the Gaussian DPP of Example 4, the expected number of points in a given compact C is α times the Lebesgue measure of C. The parameter α thus controls the intensity of the point process. Meanwhile, the pair correlation function reads

$$g(x,y) = 1 - \exp\left[-2\|x-y\|^2/\ell^2\right],$$

and thus only depends on ℓ, which controls repulsiveness. The fact that parameters of K can be easily interpreted is part of the appeal of DPPs in statistical modeling.

For the rest of this section, we turn to examining Hermitian DPPs in the special case when the cardinality $|\mathbb{X}| = n$ is finite. The kernel is now automatically trace-class, and the Macchi-Soshnikov theorem thus only requires $\mathrm{Spec}(\mathbf{K}) \in [0,1]$, where the *kernel matrix* is defined as

$$\mathbf{K} = ((\,K(i,j)\,))_{1\leqslant i,j\leqslant n}.$$

Example 5 (L-ensemble) *Take $\mathbb{X} = \{1,\dots,n\}$ and μ to be the uniform distribution over $\mathbb{X}$. Let $\mathbf{L} \in \mathbb{R}^{n\times n}$ be symmetric and positive definite, then*

$$\mathbf{K} = (\mathbf{I} + \mathbf{L})^{-1}\mathbf{L} \tag{2.16}$$

is the kernel matrix of a DPP.

The construction in Example 5 satisfies the Macchi-Soshnikov conditions by design, and it is convenient to define finite DPPs through any positive definite matrix $\mathbf{L}$. Indeed, by construction, (2.16) ensures that $\mathbf{K}$ is symmetric, with eigenvalues in $[0,1]$. Furthermore, in that case, $\mathbf{L}$ plays the role of a *likelihood* kernel, in the sense that if S is drawn from the corresponding $\mathbf{L}$-ensemble,

$$\mathbb{P}(S = A) = \frac{\det \mathbf{L}_A}{\det \mathbf{I} + \mathbf{L}}, \quad \forall A \subset \{1,\dots,n\}, \tag{2.17}$$

where $\mathbf{L}_A$ is the $|A| \times |A|$ submatrix of $\mathbf{L}$ indexed by the rows and columns in A; see e.g. (Kulesza and Taskar, 2012). This is to be compared to $\mathbf{K}$ being the kernel that encodes the correlation functions, since in the finite setting of Example 5, (2.5) is equivalent to

$$\mathbb{P}(A \subset S) = \det \mathbf{K}_A, \quad \forall A \subset \{1,\dots,n\}.$$

$\mathbf{L}$-ensembles are popular DPPs in machine learning, where a typical construction is as follows (Kulesza and Taskar, 2012, Section 4.2.1). Consider a large ground set of n objects, indexed by $\mathbb{X} = \{1, \dots, n\}$. Say, n is the number of sentences in a large text, and you want to extract a few (random) sentences from the text that are representative of the whole text.

Example 6 (Summary extraction; Kulesza and Taskar, 2012) *In the finite setting of Example 5, let $\mathbf{L}_{ij} = q_i \Phi_i^T \Phi_j q_j$, $1 \leqslant i, j \leqslant n$, where*

- *$q_i \in \mathbb{R}$ measures the marginal relevance of sentence i. For instance, it can be chosen as an increasing function of the number of nouns and a decreasing function of the number of pronouns in the sentence. A sentence with many pronouns is unlikely to be useful in a summary.*

- *$\Phi_i \in \mathbb{R}^p$ is a normed vector of features of sentence i, such that two sentences can be judged dissimilar if and only if $\Phi_i^T \Phi_j$ is close to zero. For instance, take Φ_i to have one entry per word w in the whole text, and $\Phi_i(w)$ to be an increasing function of how frequent word w is in sentence i and how rare it is in the text. Pairs of features with a large inner product will likely indicate sentences with similar content.*

A draw from the corresponding DPP will be a set of indices of sentences that are marginally relevant, but jointly dissimilar; in other words, a potentially good summary.

To help with choosing the model, it is natural to actually parametrize (q_i) and (Φ_i) and learn them from a set of human-made summaries; see (Kulesza and Taskar, 2012, Section 4.2.1). Learning can be achieved by likelihood-based methods, since $\mathbf{L}$ gives the likelihood (2.17), although one has to care for the cost of evaluating the normalization constant in (2.17). However, in my opinion, a more severe downside of modeling $\mathbf{L}$ directly is that the parameters of $\mathbf{L}$ are usually hard to interpret. In particular, the number of items in a realization of S is hard to control. In that respect, working with $\mathbf{K}$ is more natural, as we noticed when interpreting the parameters of the Gaussian DPP in Example 4. One way to recover interpretability and a control over the number of points in a realization is to work with projection DPPs as in Gautier, Bardenet, and Valko (2017), taking the number p of features in Example 6 to be the number of desired items in a summary. If the features are linearly independent, the resulting projection kernel has rank p. This construction using projection DPPs has however not been favoured in machine learning so far. MLers typically prefer using k-DPPs (Kulesza and Taskar, 2012).

For a given $k \in \mathbb{N}$, a k-DPP is a DPP conditioned to have cardinality k. While in general a k-DPP is not a DPP, a projection DPP is a k-DPP, since it has constant cardinality almost surely. Usually, k-DPPs are built by conditioning an $\mathbf{L}$-ensemble, i.e., replacing (2.17) by

$$\mathbb{P}(S = A) \propto \det \mathbf{L}_A \, \mathbf{1}_{|S|=k}. \tag{2.18}$$

In summary extraction applications such as (Kulesza and Taskar, 2012, Section 4.2.1), which inspired Example 6, the $\mathbf{L}$-ensemble is usually replaced by a k-DPP with the same $\mathbf{L}$ matrix. Interestingly, k-DPPs are statistical mixtures of projection DPPs; see e.g.

(Kulesza and Taskar, 2012). Combined to the chain rule for projection DPPs in (2.14), this observation yields gives a generic sampling algorithm for k-DPPs. We end this section with a fundamental example of k-DPP.

Example 7 (Volume sampling; Deshpande and Rademacher, 2010) *Let* $\mathbf{X}$ *be an* $n \times d$ *data matrix, with* $d \geqslant n$*, take* $\mathbf{L} = \mathbf{X}^T\mathbf{X}$*, and let* $k \geqslant n$*. The* k*-DPP* (2.18) *on* $\{1, \dots, d\}$ *is called volume sampling.*

Volume sampling was introduced by Deshpande and Vempala (2006) as a distribution for randomized column subset selection. Indeed $\mathbf{L}_{ij}$ is the inner produce of columns i and j of $\mathbf{X}$, and intuitively, sampling S proportionally to size-k subdeterminants of $\mathbf{L}$ should choose a diverse set of columns; see Chapter 4 for more details.

2.2.4 Sampling DPPs

To conclude on Hermitian kernels, still assuming that K is Hermitian and that Σ is trace-class, one can write $\mathrm{DPP}(K, \mu)$ as a statistical mixture of projection DPPs (Hough et al., 2006, Theorem 7), with the mixture weights being simple functions of the eigenvalues of the kernel. Together with the chain rule for projection DPPs (2.14), this provides a generic abstract algorithm to sample DPPs with Hermitian kernels. It is generic because it applies widely. Is it abstract because sampling from the conditionals in the chain rule is not straightforward in general, and rejection sampling routines have to be carefully designed for each kernel; see e.g. (Gautier, Bardenet, and Valko, 2019b).

2.2.5 DPPs are (usually) repulsive

As statistical models, DPPs are often thought of as repulsive point processes. By definition (2.7), the pair correlation function (2.3) of a DPP reads

$$g(x, y) = 1 - \frac{K(x, y)K(y, x)}{K(x, x)K(y, y)}.$$

If K is Hermitian and does not vanish, this indeed implies that $g < 1$, so that the DPP is repulsive in the sense of Section 2.1. Note however that many DPP kernels in random matrix theory are not Hermitian (Anderson et al., 2010, Section 4.2.9), and thus should not be thought of as repulsive. Moreover, non-Hermitian DPPs are starting to be used as statistical models that capture attractiveness as well as repulsive behaviour, e.g. in recommendation tasks such as basket completion in online retail (Gartrell et al., 2019). To explain the idea, if you just put a coffee machine in your basket, then you should be recommended items such as coffee grains (attractiveness), but the seller should avoid offering a second coffee machine (repulsiveness).

On a historical note, DPPs were first formalized in the 1972 thesis of Macchi (2017). Her initial objective was a point process description of a beam of fermions (Bénard and Macchi, 1973), using correlation functions (2.2) and the by-then recent quantum coherence formalism of Glauber (1970). Importantly, the repulsive behaviour of the resulting (Hermitian) DPPs corresponds to a physical phenomenon for fermions known as *anti-bunching*.

We refer to our survey (Bardenet et al., 2022) for more on the physical roots of DPPs. The connections between DPPs and quantum (electronic) optics have been more or less dormant since the 70s, and DPPs had to wait for the 2000s to be developed and popularized by random matrix theory (Anderson et al., 2010). After that, they have been in particular investigated as repulsive statistical models in spatial statistics (Lavancier et al., 2015; Bardenet et al., 2017b) and machine learning (Kulesza and Taskar, 2012). For more information on determinantal point processes, we refer the reader to (Macchi, 1975; Hough et al., 2006; Johansson, 2006; Soshnikov, 2000; Lyons, 2002; Kulesza and Taskar, 2012; Lavancier et al., 2015), as well as to the theses of our students Gautier (2020), Belhadji (2020), and the HDR manuscript of Hardy (2020). In the rest of this section, we introduce some examples of DPPs that will play a role in this manuscript.

2.3 Zeros of Gaussian analytic functions

Besides DPPs, a natural way to define a repulsive point process on $\mathbb{X} = \mathbb{C} \approx \mathbb{R}^2$ equipped with the Lebesgue measure is by looking at the zeros of a random analytic function. Intuitively, an analytic function is very smooth, so that its zeros should be far from each other. Restricting oneself to analytic-valued *Gaussian* processes further allows to describe the correlation functions (2.2) of the zeros using the covariance kernel of the Gaussian process.

In the following, $N_{\mathbb{C}}(0, \sigma^2)$ refers to the law of a centered Gaussian variable on $\mathbb{C}$ with variance σ^2, namely the law of $Z := \sigma(X + \mathrm{i}Y)/\sqrt{2}$ with X, Y identically distributed and independent (i.i.d.) real standard $N(0,1)$ variables; note that the $\sqrt{2}$ is here to have $\mathbb{V}\mathrm{ar}(Z) := \mathbb{E}|Z|^2 - |\mathbb{E}(Z)|^2 = \sigma^2$. More generally, given $m \geqslant 1$, a complex Gaussian vector of mean $\mu \in \mathbb{C}^m$ and covariance matrix $\Sigma := AA^*$, with $A \in \mathcal{M}_n(\mathbb{C})$, is a random variable with the same distribution as $AX + \mu$, where $Z := (Z_1, \dots, Z_d)$ with $(Z_i)_{i=1}^m$ i.i.d. $N_{\mathbb{C}}(0,1)$ random variables. Given an open subset $\Lambda \subset \mathbb{C}$, we let $\mathcal{A}(\Lambda)$ be the space of analytic functions on Λ.

2.3.1 A constructive definition of GAFs

A *Gaussian analytic function* (GAF) on an open subset $\Lambda \subset \mathbb{C}$ is a random variable GAF taking its values in the space $\mathcal{A}(\Lambda)$ of analytic functions such that, for any $m \geqslant 1$ and $z_1, \dots, z_m \in \Lambda$, the vector $(\mathsf{GAF}(z_1), \dots, \mathsf{GAF}(z_m))$ is a complex Gaussian vector. In this work, we actually slightly restrict the definition.

Definition 3 *All GAFs in this work are defined by*

$$\mathsf{GAF}(z) = \sum_{k=0}^{\infty} \xi_k \, \Psi_k(z), \tag{2.19}$$

where $(\xi_k)_{k\in\mathbb{N}}$ is a sequence of i.i.d. $N_{\mathbb{C}}(0,1)$ random variables and $(\Psi_k)_{k\in\mathbb{N}}$ is a sequence in $\mathcal{A}(\Lambda)$ satisfying

$$\sup_{z\in K} \sum_{k=0}^{\infty} |\Psi_k(z)|^2 < \infty \tag{2.20}$$

for any compact $K \subset \Lambda$.

Condition (2.20) ensures that GAF converges almost surely on every compact subset of Λ, and thus that GAF is a well-defined analytic function, see (Hough et al., 2009, Lemma 2.2.3). As a side note, the proof of the latter lemma, while elementary, is a good example of the beautiful interplay of analytic and probabilistic notions that is behind GAFs.

As a centered Gaussian process on the space of analytic functions $\mathcal{A}(\Lambda)$, GAF is completely characterized by its covariance structure, namely

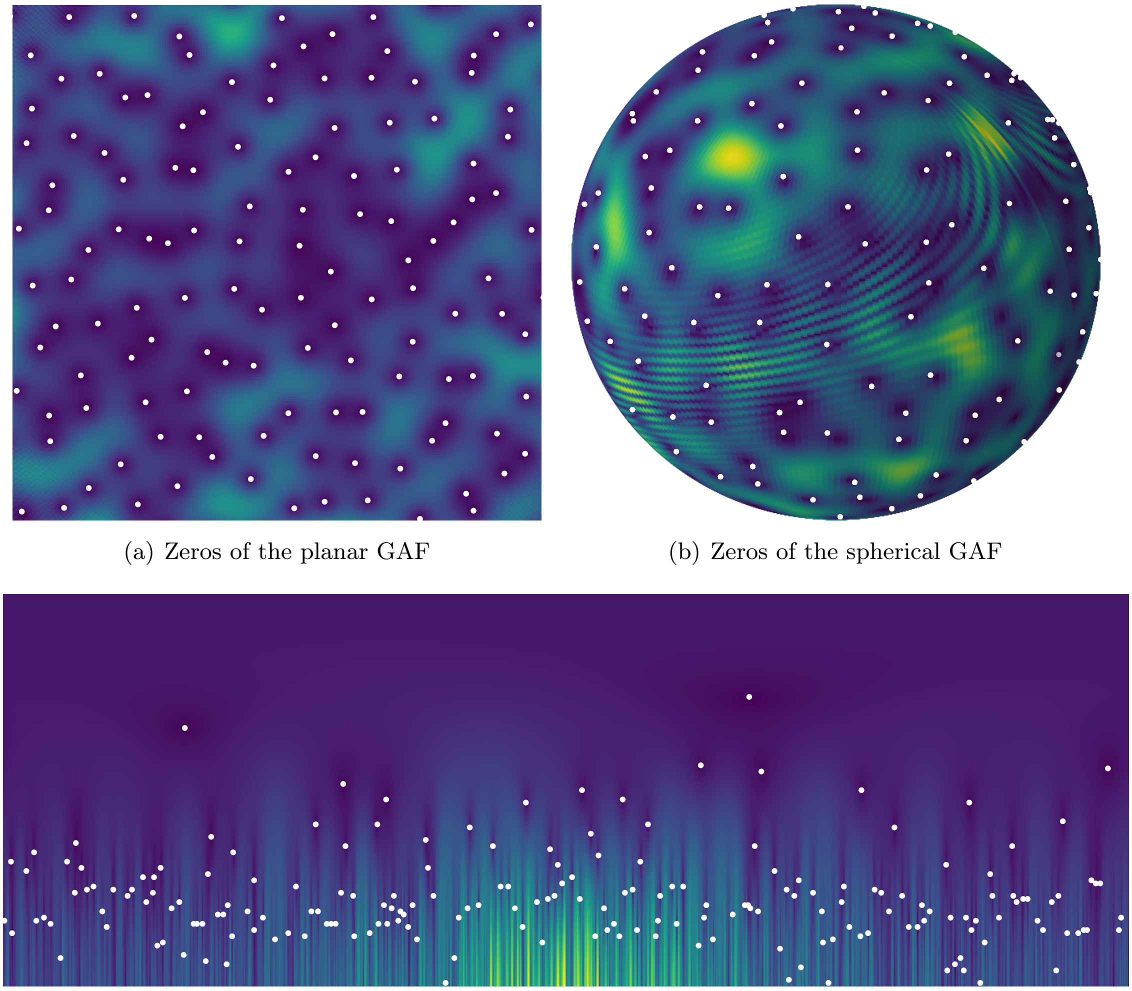
$$C(z,w) := \mathbb{E}\left[\mathsf{GAF}(z)\overline{\mathsf{GAF}(w)}\right] = \sum_{k=0}^{\infty} \Psi_k(z)\overline{\Psi_k(w)}, \qquad z, w \in \Lambda. \tag{2.21}$$

Our particular interest lies in the random set $\mathsf{GAF}^{-1}(0)$ of zeros of GAF. More precisely, our interest lies in the *random* zeros of GAF, that is, zeros that are not present with probability one. Indeed, $z \mapsto (z-1)\mathsf{GAF}(z)$ is still a GAF, but the (deterministic) zero in 1 is not of interest to us. (Hough et al., 2009, Lemma 2.4.1) states that the random zeros of any nonzero GAF are pairwise distinct with probability one. Thus, the random zeros form a (simple) point process in the sense of Definition 1. Like the DPPs of Section 2.2, the correlation functions (2.2) of the zeros of a GAF are known in closed form (Hough et al., 2009, Corollary 3.4.2).

GAFs are very constrained models. For instance, it is known that the law of the point process $\mathsf{GAF}^{-1}(0)$ is characterized by its first intensity ρ_1 solely and moreover that, if two GAFs, say GAF and $\widetilde{\mathsf{GAF}}$, share the same zero set distribution, namely $\mathsf{GAF}^{-1}(0) \overset{\text{law}}{=} \widetilde{\mathsf{GAF}}^{-1}(0)$, then there exists a non-vanishing, deterministic $\varphi \in \mathcal{A}(\Lambda)$ such that $\mathsf{GAF} = \varphi \times \widetilde{\mathsf{GAF}}$, see (Hough et al., 2009, Theorem 2.5.2).

2.3.2 Three canonical GAFs

There are three prototypical spaces in complex geometry: the complex plane $\mathbb{C}$, the hyperbolic plane $\mathbb{H}$, and the sphere $\mathbb{S}$. For instance, the so-called uniformization theorem says that any simply connected Riemann surface is conformally equivalent to $\mathbb{C}$, $\mathbb{H}$ or $\mathbb{S}$. The complex plane is a Riemannian manifold with null curvature when equipped with the Euclidean metric g_{flat}. The hyperbolic plane $\mathbb{H}$ can be modeled as the unit disc $\mathbb{D} := \{z \in \mathbb{C} : |z| < 1\}$ with the hyperbolic metric $(1-|z|^2)^{-2} g_{\text{flat}}$, giving a Riemannian manifold of constant negative curvature. It can alternatively be modeled as the upper-half plane $\mathbb{C}_+ := \{z \in \mathbb{C} : \mathfrak{Im}(z) > 0\}$ with the metric $\mathfrak{Im}(z)^{-1} g_{\text{flat}}$. The sphere $\mathbb{S}$ can be modeled as the extended complex plane $\mathbb{C} \cup \{\infty\}$ with the spherical metric $(1+|z|^2)^{-2} g_{\text{flat}}$, yielding a Riemannian manifold of constant positive curvature. This metric is the image by the stereographic projection of the natural metric on the unit sphere of $\mathbb{R}^3$ induced by the Euclidean structure of $\mathbb{R}^3$. The group of isometries for each of these three spaces is well-known, and it is natural to ask which GAFs have their zeros invariant under these isometries. It turns out that, up to multiplication by a non-vanishing and deterministic analytic function, there is only a one-parameter family of GAFs satisfying this condition for each space; see (Hough et al., 2009).



(a) Zeros of the planar GAF

(b) Zeros of the spherical GAF

(c) Zeros of the hyperbolic GAF

Figure 2.3: White dots are sample zero sets of (a) the planar GAF, (b) the spherical GAF, and (c) the hyperbolic GAF.

We shall meet these so-called *planar, hyperbolic, and spherical* GAFs in Chapter 5. Meanwhile, a realization of the zero sets of each of the three invariant GAFs is shown in Figure 2.3 as white dots. For readers unfamiliar with complex geometry, Figure 2.3 can also serve as an illustration of the previous paragraph. Roughly speaking, on all three subfigures, the white dots are uniformly spread, with roughly equal distance between each pair of nearest neighbors. Figure 5.2 represents a portion of the complex plane, and the flat metric corresponds to the intuitive idea of a uniform distribution of zeros on this portion. Figure 5.3 represents a portion of $\mathbb{C}_+$ with the hyperbolic metric $\mathfrak{Im}(z)^{-1}g_{\text{flat}}$. Again, the white dots are uniformly spread for this metric, with distances between nearest neighbors roughly equal. Figure 5.4 similarly represents the sphere with the spherical metric.

Chapter 3

Monte Carlo integration with determinantal point processes

> *God does not play dice with the universe; He plays an ineffable game of His own devising, which might be compared, from the perspective of any of the other players [i.e. everybody], to being involved in an obscure and complex variant of poker in a pitch-dark room, with blank cards, for infinite stakes, with a Dealer who won't tell you the rules, and who* smiles all the time.
>
> Terry Pratchett and Neil Gaiman, *Good Omens*, 1990

Contents

3.1 A central limit theorem from random matrix theory 20

3.2 From random matrices to numerical integration 22

3.2.1 Multivariate orthogonal polynomial ensembles 22

3.2.2 A stochastic Gaussian quadrature 22

3.2.3 Our generalization of Johansson's CLT 23

3.2.4 An importance sampling version 24

3.3 Soviets already did it 25

3.4 Interpolation and quadrature in RKHSs 26

3.4.1 Reproducing kernel Hilbert spaces 27

3.4.2 Quadrature in a RKHS with a projection DPP 28

3.4.3 Quadrature in an RKHS with continuous volume sampling 29

3.5 Closing remarks 30

My initial motivation to look into DPPs came from discussions around a swimming pool with my long-time friend Adrien Hardy in the Summer holidays of 2013, while we were both postdocs, respectively in Oxford and Stockholm. Adrien was trying to give me an overview of the big problems of random matrix theory, while I was trying to do the same with Monte Carlo methodology. What caught our attention is that fundamental theorems about the asymptotic behaviour of the spectrum of certain random matrices could be interpreted as yielding fast Monte Carlo estimators in $\mathbb{R}$ and $\mathbb{C}$. By *fast*, I mean estimators with mean square error decreasing faster than the inverse of the number of evaluations of the integrand, that is, faster than the usual Monte Carlo error. One natural question was then to understand what variance cancellation mechanisms made this fast decay possible, and try to construct quadrature nodes in arbitrary dimension that featured the same limiting behaviour. This turned out to be possible, though quite technically involved, using specific DPPs. It gave (Bardenet and Hardy, 2020), which is the starting point of my work on repulsive point processes, although the publication date does not reflect it. Section 3.1 describes a classical theorem by Johansson (1997) that provided motivation for our work, and our main results are given in Section 3.2.

After I presented the results of Bardenet and Hardy (2020) at a seminar in Bristol, UK, Mathieu Gerber, who was in the audience, thought that he had seen something similar somewhere. In the afternoon following the seminar, we engaged in serious scientific archaeology to find out the source of the remark he remembered. Following a chain of progressively more obscure references, we ended up with an English translation of a paper by Ermakov and Zolotukhin (1960), two Soviet mathematicians that actually had proposed using a projection DPP for Monte Carlo, almost 60 years before us, and some 15 years before DPPs where even formally defined by Macchi (1975). Of course, Ermakov and Zolotukhin (1960) did not have the mathematical toolbox back then that we had at our disposal in (Bardenet and Hardy, 2020), and did not prove a CLT, for instance, but they already had the intuition to sample quadrature nodes proportionally to the square of some determinant, motivated by the fact that it gave good interpolation nodes. For a discussion of their results in modern parlance, see (Gautier, Bardenet, and Valko, 2019c) or Section 3.3 below. Revisiting this connection with interpolation with Pierre Chainais and our PhD student Ayoub Belhadji, we came up with a dual approach to (Bardenet and Hardy, 2020). Instead of trying to find a DPP yielding a good estimator for as wide a function class as possible, we asked the following question: if I give you a DPP with kernel K, is that a good interpolating set of nodes for some function space, and what functions is this DPP supposed to integrate well? The approach naturally connects with recent work on quadrature in reproducing kernel Hilbert spaces (Bach, 2017), and gave (Belhadji et al., 2019, 2020a); see Section 3.4.

To sum up, this chapter is based on our works (Bardenet and Hardy, 2020; Belhadji et al., 2019; Gautier et al., 2019b; Belhadji et al., 2020a).

3.1 A central limit theorem from random matrix theory

Much of random matrix theory deals with the behaviour of the eigenvalues of certain random matrices, as can be gathered from the reference book (Anderson, Guionnet, and Zeitouni, 2010). These eigenvalues are highly correlated random variables, and they can have surprising properties to the scientist more used to independence. The following example is due to Johansson (1997).

Let $\mathrm{U}_N \subset \mathcal{M}_N(\mathbb{C})$ be the group of $N \times N$ unitary matrices, that is, $A \in \mathrm{U}_N \Leftrightarrow A\bar{A}^T = \bar{A}^T A = I_N$. There is a unique probability measure μ_H on U_N such that

$$\mu_H(A) = \mu_H(\{ga; a \in A\}) \tag{3.1}$$

for any $g \in \mathrm{U}_N$ and any Borel set $A \subset \mathrm{U}_N$. This measure is called *the Haar measure* on U_N, see for instance (Anderson et al., 2010, Theorem F.13). In other words, the Haar measure μ_H plays on U_N the role of the Lebesgue measure on the additive group $\mathbb{R}$, so that μ_H is often thought of as the natural equivalent of the uniform probability distribution. Now, let $U \sim \mu_H$, and since the eigenvalues of a unitary matrix are complex numbers of unit modulus, let us further denote the N eigenvalues of U by their arguments $(\theta_i) \in [0, 2\pi]^N$.

Johansson (1997) used a result called the *strong Szegő* theorem to prove that if $g : [0, 2\pi] \to \mathbb{R}$ is in L^1, and its Fourier coefficients satisfy

$$\sigma^2 \triangleq 2 \sum_{k=1}^{\infty} k |\hat{g}_k|^2 < \infty, \tag{3.2}$$

then

$$\sum_{i=1}^{N} g(\theta_i) - N \int_0^{2\pi} g(\theta) \frac{\mathrm{d}\theta}{2\pi} \to \mathcal{N}(0, \sigma^2), \tag{3.3}$$

where the convergence is in distribution when $N \to \infty$. The central limit theorem (CLT) in Equation (3.3) is surprising by its lack of normalization: the standard CLT for i.i.d. variables would have a factor $\sqrt{N}$ in the left-hand side. In other words, if you think of $I_N \triangleq N^{-1} \sum_{i=1}^{N} g(\theta_i)$ as a Monte Carlo estimator of $(2\pi)^{-1} \int g(\theta) \mathrm{d}\theta$, then (3.3) implies that the mean square error of I_N is asymptotically equivalent to $1/N^2$. This contrasts with traditional Monte Carlo variance of order $1/N$.

Efficient cancellation happens in the variance of the left-hand side of (3.3). Actually, the arguments (θ_i) of U form a DPP on $[0, 2\pi]$, namely the circular unitary ensemble defined in Example 2. One way to get some intuition for the fast convergence is as follows. The eigenvalues of U tend to be very regularly spaced on the unit circle compared to uniform samples, as if they "repelled" each other; see Figure 2.2 If we think of the eigenvalues as a random grid on which the signal g is measured, repulsiveness is akin to the grid being rigid. Rigidity has two positive features. First, the number of points that fall in any fixed Borel subset of $[0, 2\pi]$ does not vary much from one realization of U to the other, hence the small variance of $\sum g(\theta_i)$. Second, g essentially contains low frequencies by assumption (3.2), so that repulsiveness ensures both that $\sum g(\theta_i)$ has small bias and that we make the most of the smoothness of g by spreading out samples (θ_i).

So how do we take a result like Johansson's in (3.3) and make it a general numerical integration result? In other words, we want to reverse-engineer (3.3) and assess whether repulsive random variables as seen in random matrix theory can be a general device for variance reduction in Monte Carlo computation. There are two issues to consider. First, eigenvalues of random matrices are bound to be in $\mathbb{R}$ or $\mathbb{C}$, while we would like to integrate over $\mathbb{R}^d$ for arbitrary $d \geqslant 1$. Second, we would like to integrate against an arbitrary measure, whereas random matrix results such as (3.3) implicitly choose one measure (here, uniform over $[0, 2\pi]$). We add to these requirements that we need to preserve both the fast convergence compared to traditional Monte Carlo methods, and the interpretability of the asymptotic variance (3.2) as a measure of the decay of the Fourier coefficients of the integrand.

3.2 From random matrices to numerical integration

Let μ be a positive Borel measure supported in $[-1,1]^d$. We want a quadrature

$$\int f \mathrm{d}\mu \approx \sum_{i=1}^{N} w_i f(x_i) \tag{3.4}$$

for a wide class of functions f. Johansson's result (3.3) does this for $d=1$ and μ the uniform measure on $[0,2\pi]$, using nodes (x_i) drawn from a DPP with kernel the projection kernel onto the first N Fourier exponentials, which are orthonormal w.r.t. the uniform measure on S^1. Betting that this structure is what yields (3.3), and focussing now on a generic measure $\mu(\mathrm{d}x) = \omega(x)\mathrm{d}x$ supported in the cube $[-1,1]^d$, we look for a projection kernel onto N functions that are orthonormal w.r.t. the target measure μ in (3.4).

3.2.1 Multivariate orthogonal polynomial ensembles

When $d=1$, letting φ_k in (2.6) be the orthonormal polynomials w.r.t. μ results in a DPP called an *orthogonal polynomial ensemble* (OPE, König (2004)). When $d>1$, orthonormal polynomials can still be uniquely defined by applying the Gram-Schmidt procedure to a set of monomials, provided the base measure is not pathological. However, unlike for $d=1$, there is no natural order on multivariate monomials, so we pick an order $\mathfrak{b} : \mathbb{N} \to \mathbb{N}^d$ before we apply Gram-Schmidt to the first N monomials in that order. In (Bardenet and Hardy, 2020) we use the graded lexicographic order, defined by ordering multi-indices $(\alpha_1,\dots,\alpha_d)$ by their maximum degree $\max \alpha_i$, and for constant maximum degree, by the usual lexicographic order. There is some limited freedom to change the order in our proofs.

Denoting by Φ_k the corresponding multivariate orthonormal polynomials, then by multivariate OPE we mean the DPP with base measure $\mu(\mathrm{d}x) = \omega(x)\mathrm{d}x$ and kernel $K_{\mathrm{OPE}}(x,y) = \sum \Phi_{\mathbf{k}}(x)\Phi_{\mathbf{k}}(y)$, where the sum runs over multi-indices $\mathbf{k}$ such that $0 \leqslant \mathfrak{b}(\mathbf{k}) \leqslant N-1$. This is a projection kernel, and thus a valid DPP kernel; see Section 2.2.1. Note that for a separable base measure, i.e. if

$$\omega(x) = \omega^1(x^1)\dots\omega^d(x^d), \tag{3.5}$$

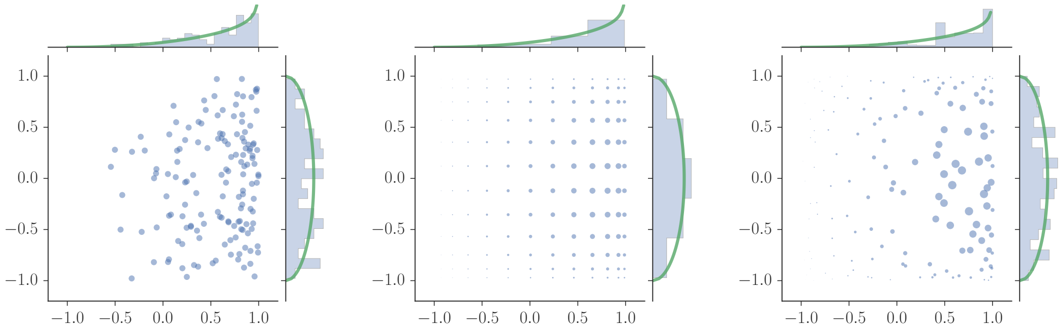
where $x = (x^1,\dots,x^d) \in \mathbb{R}^d$, multivariate orthonormal polynomials are products of univariate ones. In that case,

$$K_{\mathrm{OPE}}(x,y) = \sum_{\mathfrak{b}(\mathbf{k})=\mathbf{0}}^{N-1} \prod_{j=1}^{d} \varphi_{k_j}^j(x^j)\varphi_{k_j}^j(y^j), \tag{3.6}$$

where $(\varphi_k^j)_{k\geqslant 0}$ are the orthonormal polynomials w.r.t. ω^j.

3.2.2 A stochastic Gaussian quadrature

Let μ be separable as in (3.5), and $\{x_1,\dots,x_N\}$ be drawn from the multivariate OPE given by (3.6). If we integrate (2.7) $N-1$ times, we obtain that any point in our DPP



(a) i.i.d.

(b) Gaussian quadrature

(c) DPP

Figure 3.1: (a) I.i.d. sample of a product of beta marginals, (b) Product Gaussian quadrature, (c) bivariate OPE sample.

has marginal pdf $K_{\mathrm{OPE}}(x,x)\omega(x)$. For $f \in L^1(\mathrm{d}\mu)$, we deduce that

$$\hat{I}_N \triangleq \sum_{i=1}^{N} \frac{f(x_i)}{K_{\mathrm{OPE}}(x_i,x_i)} \tag{3.7}$$

is an unbiased estimator of $\int f(x)\omega(x)\mathrm{d}x$. Note that (3.3) does not contain node-dependent weights, since the kernel of the CUE has constant diagonal, $K_{\mathrm{CUE}}(x,x) \propto 1$. Note also that the estimator (3.7) should be familiar to users of Gaussian quadrature. Gauss (1815) indeed showed that if you restrict to $d = 1$, and replace the DPP samples in (3.7) by the zeros of φ_N, then $\hat{I}_N$ is deterministic and has zero error when f is a polynomial of degree less than $2N - 1$. This is remarkable as φ_N has degree N and thus only N zeros, while integrating polynomials of degree twice that. In a sense, our estimator (3.7) is a stochastic relaxation of Gaussian quadrature.

To further visualize the similarity with Gaussian quadrature, we plot in Figure 3.1 three samples. The left plot depicts an i.i.d. sample from a product measure with Jacobi marginals. The marginals are plotted in green on each axis. The middle plot contains the Cartesian product of two Gaussian quadratures, one on each axis with the corresponding green marginal as target measure. Each node x_i is depicted with a circle, the area of which is proportional to the node's weight $1/K_{\mathrm{OPE}}(x_i,x_i)$ in (3.7) and in Gaussian quadrature. The right plot shows a sample from the corresponding multivariate OPE, with the same weights. Well-spread points and large weights accumulate in the bulk of ω in both the middle and right plots.

3.2.3 Our generalization of Johansson's CLT

Theorem 2 (Bardenet and Hardy, 2020) *Let $\mu(\mathrm{d}x) = \omega(x)\mathrm{d}x$ with ω separable, $\mathcal{C}^1$, positive on the open set $(-1,1)^d$, and satisfying a technical regularity assumption. If $x_1, \dots, x_N$ stands for the associated multivariate OPE, then for every g that is $\mathcal{C}^1$ and* *vanishing outside* $[-1+\varepsilon, 1-\varepsilon]^d$ *for some* $\varepsilon > 0$*,*

$$\sqrt{N^{1+1/d}}\left(\hat{I}_N - \int g(x)\mu(\mathrm{d}x)\right) \xrightarrow[N\to\infty]{law} \mathcal{N}\left(0, \Omega^2_{g,\omega}\right),$$

where

$$\Omega^2_{g,\omega} = \frac{1}{2}\sum_{k_1,\dots,k_d=0}^{\infty}(k_1+\dots+k_d)\widehat{\left(\frac{g\omega}{\omega_{\mathrm{eq}}^{\otimes d}}\right)}(k_1,\dots,k_d)^2, \tag{3.8}$$

and $\omega_{\mathrm{eq}}(x) = \pi^{-1}(1-x^2)^{-1/2}$*.*

Looking back at our requirements in Section 3.1, we have extended (3.3) to an arbitrary dimension d, and an arbitrary choice of target measure. We have preserved the fast convergence of (3.3), although the gain w.r.t. the usual Monte Carlo rate decreases with dimension, and we have preserved the interpretability of the asymptotic variance as a measure of decay of the Fourier coefficients of the integrand, including the target pdf. The additional ω_{eq} term is related to the projection of the circle onto the x-axis: ω_{eq} is the "marginal" distribution of the uniform distribution on the circle. As a last comment, we have a CLT that essentially only assumes that the integrand is $\mathcal{C}^1$, while typical faster-than-Monte-Carlo algorithms like quasi-Monte Carlo (Dick and Pillichshammer, 2010) require the smoothness of the integrand to grow with d. To sum up, the fact that the repulsiveness in our DPP is tailored to the integration problem at hand allows for weaker assumptions, and gives a fast CLT with interpretable asymptotic variance.

3.2.4 An importance sampling version

One downside of Theorem 2 is that in practice, one rarely can evaluate the orthogonal polynomials w.r.t. μ. Indeed, this would imply that we know all moments of μ, which is unrealistic in, say, Bayesian inference, where access to μ is limited to pointwise evaluation up to a multiplicative constant (Robert and Casella, 2004). In (Bardenet and Hardy, 2020, Theorem 2.9), we show that if $x_1,\dots,x_N$ is the multivariate OPE associated to a pdf q that satisfies the assumption of Theorem 2, then

$$\tilde{I}_N \triangleq \sum_{i=1}^{N}\frac{g(x_i)}{K_{\mathrm{OPE}}(x_i,x_i)}\frac{\omega(x_i)}{q(x_i)}$$

satisfies the same CLT as in Theorem 2, with the same asymptotic variance. In other words, you do not incur any cost, asymptotically, for having replaced ω by another reference distribution. This is highly unusual from a Monte Carlo perspective, where you would expect the asymptotic variance to contain a term that measures the mismatch between q and ω. From an orthogonal point of view, this is a consequence of the universality of Chebyshev polynomials: whenever the underlying measure satisfies some weak assumptions (technically, is Nevai-class; see Bardenet and Hardy, 2020), the corresponding orthogonal polynomials.

3.3 Soviets already did it

In a classical twist, we later realized that Soviet mathematicians were actually the first to consider drawing quadrature nodes from projection DPPs. Fifteen years before Macchi (1975) even defined DPPs, Ermakov and Zolotukhin (1960) had already proven the following result, which we restate in modern jargon.

Theorem 3 (Ermakov and Zolotukhin, 1960) *Consider $f \in L^2(\mu)$, and N functions $\varphi_0, \dots, \varphi_{N-1} \in L^2(\mu)$ orthonormal w.r.t. μ. Let $\{x_1, \dots, x_N\}$ be drawn from the DPP with reference measure μ and projection kernel (2.6). Consider the linear system*

$$\begin{pmatrix} \varphi_0(x_1) & \dots & \varphi_{N-1}(x_1) \\ \vdots & & \vdots \\ \varphi_0(x_N) & \dots & \varphi_{N-1}(x_N) \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_N) \end{pmatrix}. \tag{3.9}$$

Then the solution of (3.9) is unique, μ-almost surely, with coordinates

$$y_k = \frac{\det \mathbf{\Phi}_{\varphi_{k-1},f}(x_{1:N})}{\det \mathbf{\Phi}(x_{1:N})},$$

where $\mathbf{\Phi}_{\varphi_{k-1},f}(x_{1:N})$ is the matrix obtained by replacing the k-th column of $\mathbf{\Phi}(x_{1:N})$ by $f(x_{1:N})$. Moreover, for all $1 \leqslant k \neq j \leqslant N$, the solution vector satisfies

$$\mathbb{E}[y_k] = \langle f, \varphi_{k-1} \rangle, \qquad \mathbb{V}\mathrm{ar}[y_k] = \|f\|^2 - \sum_{\ell=0}^{N-1} \langle f, \varphi_\ell \rangle^2, \quad \textit{and} \quad \mathbb{C}\mathrm{ov}[y_j, y_k] = 0. \tag{3.10}$$

Of course, back then, Ermakov and Zolotukhin (1960) did not have the DPP machinery that we now have in our toolbox to study, e.g., the asymptotic fluctuations of their estimator, but Theorem 3 is still good food for thought. We investigated some the consequences of Theorem 3 in (Gautier, Bardenet, and Valko, 2019c).

For starters, it provides unbiased estimators of the N Fourier-like coefficients $\langle f, \varphi_k \rangle$, $k \leqslant N-1$, and these N estimators are uncorrelated, with the same variance. Moreover, the faster the decay of the Fourier-like coefficients, the smaller the variance. In particular, if f is in the span of the eigenfunctions of the kernel, the variance is zero. In the setting of multivariate OPEs, the first (in the order prescribed by $\mathfrak{b}$) orthonormal polynomial φ_0 is constant, equal to $\mu\big([-1,1]^d\big)^{-1/2}$. Hence, a direct application of Theorem (3) yields that

$$\widetilde{I}_N(f) \triangleq \mu\big([-1,1]^d\big)^{1/2} \frac{\det \mathbf{\Phi}_{\varphi_0,f}(x_{1:N})}{\det \mathbf{\Phi}(x_{1:N})} \tag{3.11}$$

is an unbiased estimator of $\int_{[-1,1]^d} f \mathrm{d}\mu$. Developing the numerator w.r.t. the column that contains evaluations of f, $\widetilde{I}_N(f)$ is seen to be a quadrature estimator

$$\widetilde{I}_N(f) = \sum_{i=1}^{N} w_i^{\mathrm{EZ}}(x_1, \dots, x_N) f(x_i),$$

which we term the EZ estimator after Ermakov and Zolotukhin (1960). Unlike the variance of our Gaussian quadrature estimator $\hat{I}_N(f)$ in (3.7), the variance of $\widetilde{I}_N(f)$ clearly reflects the accuracy of the approximation of f by its projection onto $\operatorname{Span}(\varphi_1, \dots, \varphi_N)$. In particular, it allows to interpolate and perfectly integrate polynomials up to degree $\mathfrak{b}^{-1}(N-1)$.

Nonetheless, the limiting theoretical properties of $\widetilde{I}_N(f)$, like a CLT, look hard to establish. In particular, unlike our estimator $\hat{I}_N(f)$ in (3.7), the EZ estimator is not a linear statistic of the underlying DPP, since each weight $w_i^{\mathrm{EZ}}(x_1, \dots, x_N)$ depends on all nodes. Thus, it does not fit the assumptions of traditional CLTs for DPPs (Soshnikov, 2000). More general CLTs are available (Błaszczyszyn et al., 2019), but only for DPPs with sufficiently short-range correlation.

Experiments in (Gautier et al., 2019c) suggest that the EZ estimator is an excellent replacement for $\widehat{I}_N(f)$ when the residual of the projection of f onto $\operatorname{Span}(\varphi_1, \dots, \varphi_N)$ is small. In other words, if you can interpolate the integrand with a few basis functions, then you should form a projection kernel using these basis functions, draw from the corresponding DPP, and use the EZ estimator. However, the performance of EZ rapidly decreases once the integrand spreads even a small part of its norm on all functions φ_k, $k > N$. In particular, the variance of the EZ estimator on a non-polynomial but still C^∞ function is seen in (Gautier et al., 2019c) to quickly fall down to the classical $\mathcal{O}(1/N)$ as d grows, while the variance of our estimator $I_N(f)$ steadily decreases as the predicted rate $\mathcal{O}(1/N^{1+1/d})$ from Theorem 2.

3.4 Interpolation and quadrature in RKHSs

To take the best of both our stochastic Gaussian quadrature estimator (Bardenet and Hardy, 2020) and the interpolative quadrature of Ermakov and Zolotukhin (1960), it is natural to ask what functions can be well interpolated using their evaluation on a DPP sample. Intuitively, since repulsiveness in DPPs is measured by a kernel, one would like to consider functions, the smoothness of which is measured by the same kernel. This is naturally embodied by requiring the function to belong to a reproducing kernel Hilbert space (RKHS), for which interpolation and integration are tightly linked.

Like the EZ estimator of Theorem 3, our quadrature scheme shall come from interpolating a certain function on a DPP sample. But the interpolated function will not directly be the integrand, rather a representative of the target measure in the space of functions. Finally, relating the RKHS kernel (governing the smoothness of the integrand) and the DPP kernel (governing the repulsiveness of the points) will lead us to an algorithm that realizes tight interpolation and quadrature rates in RKHSs (Belhadji, Bardenet, and Chainais, 2019, 2020b).

3.4.1 Reproducing kernel Hilbert spaces

Let $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}_+$ be a positive definite kernel, and consider the set of real linear combinations of the so-called *feature functions*

$$\mathcal{F}_0 = \mathrm{Span}\big(k(x, \cdot), x \in \mathbb{X}\big).$$

Now let $\mathcal{F}$ be the completion of $\mathcal{F}_0$ for the norm derived from $\langle k(x, \cdot), k(y, \cdot)\rangle := k(x, y)$. The resulting $\mathcal{F}$ is a Hilbert space of functions such that

$$\langle f, k(x, \cdot)\rangle = f(x), \tag{3.12}$$

for all $f \in \mathcal{F}$. $\mathcal{F}$ is called the reproducing kernel Hilbert space of kernel k (RKHS; Berlinet and Thomas-Agnan, 2011). Two famous examples are the RKHS with Gaussian kernel, and Sobolev spaces of sufficiently high smoothness; see again (Berlinet and Thomas-Agnan, 2011).

The RKHS kernel k itself might not be a valid DPP kernel. One natural, "PCA-flavoured" strategy to build a valid DPP kernel from k is to write k in diagonal form (also called a *Mercer decomposition*), and build a projection kernel onto the span of some of its eigenfunctions, as in Section 2.2.1. To do that, we need a few assumptions. Still assuming that $\mathbb{X}$ is equipped with a Borel measure μ, we further assume that it is supported on all of $\mathbb{X}$, and that

$$\int_{\mathbb{X}} k(x, x)\mathrm{d}\mu(x) < +\infty. \tag{3.13}$$

As a consequence, we get that the embedding operator $I_{\mathcal{F}} : \mathcal{F} \mapsto L^2(\mu)$ is compact and injective (Steinwart and Christmann, 2008), and that the operator

$$\mathbf{\Sigma} f(\cdot) = \int_{\mathbb{X}} k(\cdot, y)f(y)\mathrm{d}\mu(y), \quad f \in L^2(\mu) \tag{3.14}$$

is self-adjoint, positive semi-definite, and trace-class (Simon, 2005). Thus, there exists an orthonormal family (e_m) of $L^2(\mu)$ formed by eigenfunctions of $\mathbf{\Sigma}$. The basis is ordered by requiring that the corresponding eigenvalues (σ_m) are non-decreasing. Our assumptions imply a Mercer-type decomposition of k,

$$k(x, y) = \sum_{m \in \mathbb{N}^*} \sigma_m e_m(x) e_m(y), \tag{3.15}$$

where $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$ and the convergence is pointwise (Steinwart and Scovel, 2012).

Moreover, for $m \in \mathbb{N}^*$, we write $e_m^{\mathcal{F}} = \sqrt{\sigma_m} e_m$. Since $I_{\mathcal{F}}$ is injective, $(e_m^{\mathcal{F}})_{m \in \mathbb{N}^*}$ is an orthonormal basis of $\mathcal{F}$ (Steinwart and Scovel, 2012). Unless explicitly stated, we assume that $\mathcal{F}$ is dense in $L^2(\mathrm{d}\mu)$, so that $(e_m)_{m \in \mathbb{N}^*}$ is an orthonormal basis of $L^2(\mathrm{d}\mu)$. For more intuition, under our assumptions,

$$f \in \mathcal{F} \quad \Leftrightarrow \quad \sum_m \sigma_m^{-1} \langle f, e_m\rangle^2_{L^2(\mathrm{d}\omega)} < \infty. \tag{3.16}$$

Again, this shows that the smoothness in an RKHS is governed by the kernel, in the sense that the modulus $|\langle f, e_m\rangle|$ of the Fourier-like coefficients of $f \in \mathcal{F}$ has to decrease fast enough compared to the eigenvalues of the kernel to make (3.16) converge.

3.4.2 Quadrature in a RKHS with a projection DPP

Now let

$$K(x,y)=\sum_{m=1}^{N} e_m(x)e_m(y) \tag{3.17}$$

be the projection kernel onto the first N eigenfunctions of the RKHS kernel k. Unlike k, the projection kernel K is a valid DPP kernel, so that we can consider quadrature nodes $\boldsymbol{x}=\{x_1,\dots,x_N\}$ drawn from the projection DPP with reference measure μ and kernel (3.17). To get a quadrature scheme, it remains to define the weights in (3.4).

To this end, we note that, given $f\in\mathcal{F}$ and $g\in L^2(\mu)$, Cauchy-Schwarz in $\mathcal{F}$ yields

$$\left|\int_{\mathbb{X}} fg\,\mathrm{d}\mu-\sum_{i\in[N]} w_i f(x_i)\right| \leqslant \|f\|_{\mathcal{F}}\times\left\|\nu_g-\sum_{i=1}^{N} w_i k(x_i,\cdot)\right\|_{\mathcal{F}}, \tag{3.18}$$

where $\nu_g=\int_{\mathbb{X}} g(x)k(x,.)\mathrm{d}\mu(x)=\boldsymbol{\Sigma}g$ is the so-called *embedding* of g in the RKHS $\mathcal{F}$. From now on, we fix an arbitrary $g\in L^2(\mu)$.

By (3.18), an upper bound on the approximation error of ν_g in $\mathcal{F}$ implies an upper bound on the integration error of fg that is uniform over any bounded subset of $\mathcal{F}$. This observation has sparked intense research on the kernel approximation of embeddings ν_g. On our end, we note that thanks to the reproducing property (3.12), the approximation error in the RHS of (3.18) rewrites

$$\left\|\nu_g-\sum_{i=1}^{N} w_i k(x_i,\cdot)\right\|_{\mathcal{F}} = \|\nu_g\|^2-2\sum_{i=1}^{N} w_i\nu_g(x_i)+\sum_{i,j=1}^{N} w_iw_jk(x_i,x_j). \tag{3.19}$$

If the nodes $x_1,\dots,x_N$ are fixed and in general position, i.e. $\det((k(x_i,x_j)))>0$, then (3.19) entails that there is a unique set of weights that minimizes the upper bound (3.18), namely

$$\begin{pmatrix}\hat{w}_1\\ \vdots\\ \hat{w}_N\end{pmatrix}=\big((k(x_i,x_j))\big)^{-1}\begin{pmatrix}\nu_g(x_1)\\ \vdots\\ \nu_g(x_N)\end{pmatrix}. \tag{3.20}$$

Note that these optimal weights also guarantee that $\sum_{i=1}^{N}\hat{w}_i k(x_i,\cdot)$ interpolates ν_g at the nodes x_i.

Theorem 4 (Belhadji, Bardenet, and Chainais, 2019) *Under the assumptions of Section 3.4, let $(x_1,\dots x_N)$ be drawn from the projection DPP with reference measure μ and kernel K in (3.17). Then det$((k(x_i,x_j))>0$ almost surely, so that we can consider the optimal weights (3.20).*

Let $r_N=\sum_{m\geqslant N+1}\sigma_m$, then

$$\mathbb{E}\sup_{\|g\|_\mu\leqslant 1}\left\|\nu_g-\sum_{i=1}^{N} w_i k(x_i,\cdot)\right\|_{\mathcal{F}}^2 \leqslant 2\sigma_{N+1}+2N^2 r_N. \tag{3.21}$$

This version of the theorem is actually a slight improvement over our the original result in (Belhadji et al., 2019); see the PhD manuscript of Belhadji (2020). Note that I present a uniform bound in g for simplicity, but we have more precise g-dependent bounds; see (Belhadji, 2020, Chapter 4) for the sharpest version. The important thing to notice here is that if σ_n decreases fast enough, say geometrically as for the Gaussian kernel, then both terms in the bound are of the same order. But if the spectrum decays polynomially, like for Sobolev RKHSs, then $N^2 r_N$ dominates. In experiments, however, we have actually observed that the expected approximation error decays as σ_{N+1} for any classical kernel, so that it should be possible to prove a bound in $\mathcal{O}(\sigma_{N+1})$. While we haven't been able to prove it under a DPP, we have obtained this faster rate for a mixture of projection DPPs called *volume sampling*.

3.4.3 Quadrature in an RKHS with continuous volume sampling

In Section 3.4.2, we built a DPP kernel K from the RKHS kernel k. If we momentarily forget about DPPs, it is also natural to define a (permutation invariant) distribution over nodes that simply measures the squared volume in $\mathcal{F}$ of the parallelotope built onto the feature functions $k(x_i, \cdot)$, i.e.

$$x_1, \dots, x_N \sim \frac{1}{Z} \det \big((k(x_i, x_j)) \big) \, \mathrm{d}\mu(x_1) \dots \mathrm{d}\mu(x_N). \tag{3.22}$$

Hadamard's inequality and the finite trace condition (3.13) guarantee that the normalization constant Z is finite, so that (3.22) is a *bona fide* probability distribution. By the Cauchy-Binet identity, it turns out that (3.22) actually defines a statistical mixture of projection DPPs.

Proposition 1 *Recall the Mercer decomposition $k(x,y) = \sum_{n\geqslant 1} \sigma_m e_m(x) e_m(y)$. For $U \subset \mathbb{N}^*$, let us further define the projection kernel*

$$K_U(x,y) = \sum_{u\in U} e_u(x) e_u(y). \tag{3.23}$$

For $N \in \mathbb{N}^*$, we have

$$\det \big((k(x_i, x_j)) \big) \propto \sum_{|U|=N} \left[\prod_{u\in U} \sigma_u \right] \frac{1}{N!} \det \big((K_U(x_i, x_j)) \big). \tag{3.24}$$

Each component in the mixture (3.24) is a projection DPP onto a choice of N eigenfunctions of k, with weight proportional to the product of the corresponding eigenvalues. The component with the largest weight is thus precisely the DPP that we use in Section 3.4.2. We can thus expect similar performances for kernel quadrature, at least when the spectrum of k has a sharp cutoff. Actually, we can even improve on the bound in Theorem 4.

Theorem 5 (**Belhadji, Bardenet, and Chainais, 2020b**) *Under the assumptions of Section 3.4, let $(x_1, \dots x_N)$ be drawn from the volume sampling distribution (3.22) with* *reference measure μ and kernel k. Then* $\det\big((k(x_i, x_j)\big) > 0$ *almost surely, so that we can consider the optimal weights* (3.20)*. Moreover,*

$$\sup_{\|g\|_\mu \leqslant 1} \mathbb{E} \left\| \nu_g - \sum_{i=1}^{N} \hat{w}_i k(x_i, \cdot) \right\|_{\mathcal{F}}^2 \leqslant \sigma_N \left(1 + \beta_N\right), \tag{3.25}$$

where $\beta_N = \min_{M \in [2:N]} \left[(N - M + 1)\sigma_N\right]^{-1} \sum_{m \geqslant M} \sigma_m$.

We note that β_N is bounded as soon as the spectrum (σ_n) decays polynomially, so that the bound in Theorem 5 is quite generic. It is actually the first such generic bound on interpolation and quadrature in RKHSs: give it any kernel, the theorem gives you a sharp bound. Indeed, a result by (Pinkus, 2012, Theorem 2.2, Chapter 4) shows that the worst-case (in g) approximation error is lower-bounded by a constant times σ_{N+1}; in that sense, our bound is sharp. Note also that, unlike Theorem 4, the supremum is outside of the expectation in (3.25); this comes from the fact that a direct computation of the expected squared error is possible under volume sampling, so that bounds appear later in the proof. We refer to the discussion in (Belhadji et al., 2020b) for details, and numerical experiments that illustrate Theorem 5.

3.5 Closing remarks

I think that the bound in Theorem 5 is a solid milestone. It demonstrates what DPPs and mixtures thereof can reach in terms of quadrature and interpolation error in a favourable setting. By favourable, I mean that the integrands live in a Hilbert space, that their smoothness is controlled by the spectral properties of a known kernel, and that the performance measure is the residual of an orthogonal projection.

Indeed, in Theorems 4 and 5, the weighted sum $\sum_{1 \leqslant i \leqslant N} \hat{w}_i k(x_i, \cdot)$ is the orthogonal projection of ν_g onto the subspace $\mathrm{Span}(k(x_i, \cdot), 1 \leqslant i \leqslant N)$ formed by linear combinations of feature functions anchored at our quadrature design. In words, the setting is favourable to DPPs because sampling quadrature nodes proportionally to the determinant of $\det\big((K(x_i, x_j))\big)$ in Theorem 4 guarantees both that the functions $k(x_i, \cdot)$ are almost orthogonal in $\mathcal{F}$, and that their span is close to that of the first N eigenfunctions of the integration operator $\mathbf{\Sigma}$, thus capturing a large part of the norm of any $f \in \mathcal{F}$. More technically, the alignment between the two spans is controlled by the so-called principal angles between the two subspaces. Taking expectations under our DPP yields tractable upper bounds on symmetric quantities of these angles; see the detailed discussion in the thesis of (Belhadji, 2020, Section 4.4). In particular, there is room for improvement in controlling *non-symmetric* quantities of the principal angles, which would likely lead to a similar sharp bound as for volume sampling.

Rather than controlling principal angles, volume sampling allows for a more direct approach. The proof of Theorem 5 works by decomposing the interpolation error onto the eigenfunctions of the integration operator $\mathbf{\Sigma}$, with the coefficients giving a prominent role to the continuous equivalent of the leverage scores from regression analysis. Volume sampling controls these leverage scores in expectation, thus yielding our Theorem 5. I don't think there is much room for improvement here.

My current belief is that DPPs and volume sampling are the right way to perform quadrature as long as the target functions are in an RKHS. The randomization helps in two ways. Algorithmically, on the one hand, one can sample the nodes in polynomial time as long as one knows the Mercer decomposition of the kernel, while maximizing the kind of Gramian determinants appearing in the densities of our nodes is a hard non-convex optimization problem. Mathematically, on the other hand, as often in Monte Carlo versus, say, quasi-Monte Carlo (Dick and Pillichshammer, 2010), it is easier to derive and interpret bounds on the mean square error of an estimator than on the error of a deterministic quadrature. I believe that Theorem 5 is a solid translation of these beliefs, and that we should now turn to methodology. In terms of Bayesian inference applications, for instance, it is not clear what role the function $g \in L^2(\mu)$ in Theorem 5 should play: should it be part of the loss function, part of the posterior pdf? How can we use that degree of freedom to further optimize the constants in the (g-dependent) version of our bound?

Similarly, in Bayesian practice, we rarely know the kernel of an RKHS to which our integrands belong. Can we learn one as we sample nodes using some MCMC sampler, and progressively turn on repulsiveness in sampling? What is a parametric class of kernels that is large enough to capture the integrands in a particular Bayesian application? How robust are our bounds to misspecifying the kernel? Reversing the viewpoint, can we force the integrand to belong to a pre-specified RKHS by limiting the choices of the modeler when designing her likelihood, so that she can trade off modeling and integration accuracy?

On the other hand, how do we actually sample from the point processes involved in our theorems, if we *don't* have access to the Mercer decomposition of the kernel? The advantage of volume sampling over any projection DPP like (3.17) is that evaluating the density (3.22) of volume sampling does not require a spectral decomposition, but only pointwise evaluations of the kernel. MCMC approaches thus seem natural (Rezaei and Gharan, 2019), but one would ideally need a sampler that reaches an error in total variation of the order of σ_N in a reasonable number of iterations, say quadratic in N. I believe that this requires a dedicated sampler that uses the kernel in a smart way.

Chapter 4

Feature selection with determinantal point processes

> *'Their magic's all numbers and angles and edges and what the stars are doing, as if that really mattered. It's all power. It's all-' Granny paused, and dredged up her favourite word to describe all she despised in wizardry, '-jommetry.'*
>
> Terry Pratchett, *Equal Rites*, 1987

Contents

4.1 Apples and oranges: column subsets and PCA 34
4.2 Multinomial sampling with and without SVD 35
4.3 Repulsive sampling needs fewer columns 36
4.4 DPPs take the best of leverage scores and repulsiveness 37
4.5 A guarantee on the excess risk in sketched linear regression . . 39
4.6 Closing remarks 41

In Chapter 3, we saw how projection DPPs can select a finite number of feature functions among $\{k(x,\cdot), x \in \mathbb{X}\}$, so as to yield a small average reconstruction error of some target function belonging to a Hilbert space. We shall see here the same general principle in action. The role of the feature functions is played by the columns of a large matrix $\boldsymbol{X} \in \mathbb{R}^{N\times d}$, $d \gg N$, of which we want to select $s \ll d$ columns. This is the general problem of *feature selection*, or *variable selection*, in statistics and learning.

One common motivation is to reduce dimensionality while preserving the interpretability of the selected features, unlike principal component analysis (PCA), which defines new mixed features. If the dimensionality reduction is meant as a preprocessing for, say, a linear regression only on the p selected columns, then one potential performance metric is the excess risk of regressing onto the p columns, compared to the original regression vector. However, if one wants to have a performance metric for selecting a subset S of column indices that is independent of the regression method, another natural choice is some matrix norm

$$\|\boldsymbol{X} - \Pi_S^\nu \boldsymbol{X}\|_\nu \tag{4.1}$$

of the residual of the projection of $\boldsymbol{X}$ onto the span of the selected columns. The index ν in (4.1) denotes the matrix norm that is used in measuring the size of the residual and in defining the projection Π_S^ν; see Section 4.1. If the residual is measured in terms of Frobenius or spectral norm, i.e. $\nu \in \{\mathrm{Fr}, 2\}$, then the rank-$k$ projection Π that minimizes the norm of the residual $\|\boldsymbol{X} - \Pi\boldsymbol{X}\|_\nu$ is the projection Π_k onto the subspace spanned by the first k principal components of $\boldsymbol{X}$.

After quickly surveying the state-of-the-art in column subset selection in Sections 4.2 and 4.3, we shall first show in Section 4.4 that a suitably chosen projection DPP yields a reconstruction error (4.1) in both Frobenius and spectral norm that is as close as possible to the PCA gold standard. In that sense, DPPs yield a variable selection that is as efficient as PCA, while preserving the semantics of the original features, since –unlike PCA– the new features (columns) are a subset of the original ones. Then, we will show that the same DPP allows to control the excess risk in linear regression; see Section 4.5. As in Chapter 3, we shall finish by underlining the limitations of the approach and current work.

Finally, this chapter is based on part of Ayoub Belhadji's PhD thesis, in particular the paper (Belhadji, Bardenet, and Chainais, 2020a), to which I refer the reader for details. On top of all proofs and illustrative experiments, the paper contains a detailed discussion of past work on column-subset selection that is of independent interest.

4.1 Apples and oranges: column subsets and PCA

The result of column subset selection will usually be compared to the result of projecting onto the first k principal components of $\boldsymbol{X}$, and in this chapter, we reserve the notation $k \in \mathbb{N}$ for that reference size. We denote by $\Pi_k\boldsymbol{X}$ the best rank-k approximation to $\boldsymbol{X}$. The sense of *best* can be understood in either Frobenius or spectral norm, as both give the same result. But this coincidence is lost when projecting onto the span of a subset of columns of $\boldsymbol{X}$, and some care must be taken in defining the quantities to be compared in a fair way.

For a given subset $S \subset \{1, \dots, d\}$ of size $|S| = s$ and $\nu \in \{2, \mathrm{Fr}\}$, let

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

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 $\mathrm{rk}\boldsymbol{B} \leqslant k$; 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 $\Pi_S^\nu\boldsymbol{X} = \Pi_{S,k}^\nu\boldsymbol{X}$. In practice, the Frobenius projection can be computed as $\Pi_S^{\mathrm{Fr}}\boldsymbol{X} = \boldsymbol{C}\boldsymbol{C}^+\boldsymbol{X}$, yet there is no simple expression for $\Pi_S^2\boldsymbol{X}$. However, $\Pi_S^{\mathrm{Fr}}\boldsymbol{X}$ can be used as a proxy for $\Pi_S^2\boldsymbol{X}$ since

$$\|\boldsymbol{X} - \Pi_S^2\boldsymbol{X}\|_2 \leqslant \|\boldsymbol{X} - \Pi_S^{\mathrm{Fr}}\boldsymbol{X}\|_2 \leqslant \sqrt{2}\|\boldsymbol{X} - \Pi_S^2\boldsymbol{X}\|_2, \tag{4.2}$$

see (Boutsidis et al., 2011, Lemma 2.3).

4.2 Multinomial sampling with and without SVD

Note that because my overall narrative is about explaining when and where DPPs can help, I take it for granted that we want to apply *randomized* algorithms. Randomized approaches to numerical linear algebra tend to be more scalable, and come with more interpretable error bounds. For instance, naively optimizing a criterion like $\|\boldsymbol{X}-\Pi_{S,k}\|^{\nu}\boldsymbol{X}$ over subsets of $\{1,\dots,d\}$ requires exhaustive enumeration. This is prohibitive when $d \gg 1$, while randomized approaches typically have polynomial costs. Less naive deterministic column subset selection methods have been investigated, in particular rank-revealing QR factorizations in the vein of (Golub, 1965), but even there, randomized approaches (Boutsidis et al., 2009) can match the performance and cost of state-of-the-art factorizations (Gu and Eisenstat, 1996) in spectral norm, while also providing sharp guarantees in Frobenius norm.

The first randomized algorithms to have been investigated are naturally independent draws from the set of columns (Deshpande and Vempala, 2006; Drineas et al., 2007; Boutsidis et al., 2011). More specifically, a prototypical 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), defined as follows.

Definition 4 (k-leverage scores) *Let $\boldsymbol{X} = \boldsymbol{U\Sigma V}^T$ be the singular value decomposition of the original matrix $\boldsymbol{X}$. The k-leverage scores, denoted by $\ell_1^k,\dots,\ell_d^k$, are the squared Euclidean norms of the d columns of the first k rows of $\boldsymbol{V}^T$.*

This definition is further illustrated in Figure 4.2. Geometrically, ℓ_i^k is the squared cosine of the angle between the i-th vector of the canonical basis of $\mathbb{R}^d$ and the k-principal subspace, i.e., the subspace spanned by the vectors with canonical coordinates the first k columns of $\boldsymbol{V}$. We refer the reader to (Belhadji et al., 2020a, Section 3.4) for a detailed discussion of k-leverage scores and principal angles. It thus makes intuitive sense to select columns marginally according to their k-leverage score if one wants to approximate the k-principal subspace. Note however in passing that there may be redundancy here, and that it would be even more natural to select k columns *jointly*, according to how close their span is to the k-principal subspace. This is precisely what the DPP in Section 4.4 shall do.

The k-leverage scores are expensive to evaluate, since they call for a truncated SVD of order k, but they come with the best known bound for multinomial sampling on the ratio of their expected approximation error over that of PCA.

Theorem 6 (Drineas et al., 2007, Theorem 3) *Let $\varepsilon \in (0,1]$. If the number s of sampled columns satisfies*

$$s \geqslant \frac{3200k^2}{\varepsilon^2}, \tag{4.3}$$

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

$$\mathbb{P}\left(\|\boldsymbol{X}-\Pi_{S,k}^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}^2 \leqslant (1+\varepsilon)\|\boldsymbol{X}-\Pi_k\boldsymbol{X}\|_{\mathrm{Fr}}^2\right) \geqslant 0.7. \tag{4.4}$$

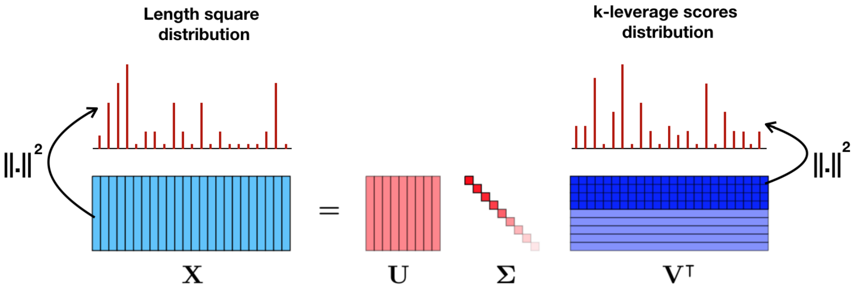


Figure 4.1: Two multinomial distributions that have been investigated for column subset selection: the squared norms of the columns of $\boldsymbol{X}$, here denoted as *length square distribution*, and the k-leverage scores. Figure taken from the PhD thesis of Belhadji (2020).

Drineas et al. (2007) also considered replacing multinomial with Bernoulli sampling, still using the k-leverage scores. The expected number of columns needed for (4.4) to hold is then lowered to $\mathcal{O}(\frac{k\log k}{\varepsilon^2})$. A natural question is then to understand how low the number of columns can be, while still guaranteeing a multiplicative bound like (4.4). A partial answer has been given by Deshpande and Vempala (2006).

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

$$\|\boldsymbol{X}^\varepsilon - \Pi^{\mathrm{Fr}}_{S,k}\boldsymbol{X}^\varepsilon\|^2_{\mathrm{Fr}} \geqslant (1+\varepsilon)\|\boldsymbol{X}^\varepsilon - \boldsymbol{X}^\varepsilon_k\|^2_{\mathrm{Fr}}. \tag{4.5}$$

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

4.3 Repulsive sampling needs fewer columns

A multiplicative bound on the expected approximation error using only k columns was obtained by Deshpande, Rademacher, Vempala, and Wang (2006) using so-called *volume sampling*, which we introduced in Example 7.

Theorem 7 (Deshpande et al., 2006) *Let S be a random subset of $\{1,\dots,d\}$, chosen with probability*

$$\mathbb{P}(S) \propto \det(\boldsymbol{X}^T_{:,S}\boldsymbol{X}_{:,S})\,\mathbf{1}_{\{|S|=k\}}. \tag{4.6}$$

Then

$$\mathbb{E}\|\boldsymbol{X} - \Pi^{\mathrm{Fr}}_S\boldsymbol{X}\|^2_{\mathrm{Fr}} \leqslant (k+1)\|\boldsymbol{X} - \Pi_k\boldsymbol{X}\|^2_{\mathrm{Fr}} \tag{4.7}$$

and

$$\mathbb{E}\|\boldsymbol{X} - \Pi^2_S\boldsymbol{X}\|^2_2 \leqslant (d-k)(k+1)\|\boldsymbol{X} - \Pi_k\boldsymbol{X}\|^2_{\mathrm{Fr}}. \tag{4.8}$$

Sampling according to (4.6) can be done in polynomial time; see (Deshpande and Rademacher, 2010) or the mixture argument for k-DPPs in Chapter 2. Using a worst case example, Deshpande et al. (2006) proved that the $k+1$ factor in (4.7) cannot be improved.

Proposition 3 (Deshpande et al., 2006) *Let $\varepsilon > 0$. There exists a $(k+1) \times (k+1)$ matrix $\boldsymbol{X}^{\varepsilon}$ such that for every subset S of k columns of $\boldsymbol{X}^{\varepsilon}$,*

$$\|\boldsymbol{X}^{\varepsilon} - \Pi_S^{\mathrm{Fr}} \boldsymbol{X}^{\varepsilon}\|_{\mathrm{Fr}}^2 > (1-\varepsilon)(k+1)\|\boldsymbol{X}^{\varepsilon} - \Pi_k \boldsymbol{X}^{\varepsilon}\|_{\mathrm{Fr}}^2. \tag{4.9}$$

4.4 DPPs take the best of leverage scores and repulsiveness

In (Belhadji et al., 2020a), our contribution is to introduce the DPP on $\{1, \dots, d\}$ with kernel

$$\boldsymbol{K} = \boldsymbol{V}_{:,1:k} \boldsymbol{V}_{:,1:k}^T, \tag{4.10}$$

and reference measure the counting measure on $[d] := \{1, \dots, d\}$. Note that this kernel matrix is a projection ($\boldsymbol{K}^2 = \boldsymbol{K}$) and has rank k, so that the corresponding DPP is a projection DPP, and samples S thus satisfy $|S| = k$ almost surely; see Chapter 2. We are thus on par with volume sampling in Theorem 7, in the sense that we sample only k columns. We shall see that our DPP reaches the $(k+1)$ factor of volume sampling in Theorem 7, and even improves over volume sampling under a certain sparsity assumption.

To see why such a DPP is natural, note that the diagonal of $\boldsymbol{K}$ is formed by the k-leverage scores of Definition 4. Since, by definition (2.5) of the one-point correlation function of a DPP, the diagonal of a DPP kernel is proportional to the marginal probability of inclusion of each element, the DPP with kernel $\boldsymbol{K}$ is thus an extension of k-leverage score sampling. Moreover, the DPP enforces diversity among the selected columns, in the sense of favouring sets of k columns $S \subset \{1, \dots, d\}$ such that the k feature vectors $\boldsymbol{V}_{i,1:k}$, $i \in S$, span a large volume in $\mathbb{R}^k$. Note that the square of this volume is also the product of the squared cosines of the principal angles between the span of the selected columns and the k-principal subspace; see (Belhadji et al., 2020a, Appendix C). Methodologically, this makes sure that we have a set of columns that jointly approximate the k-principal subspace. Mathematically, this is all the more interesting that performance metrics that compare residuals of projections to that of PCA involve the same set of principal angles.

This is precisely the same argument as for the DPP we used in Section 3.4.2 to perform quadrature in an RKHS. For this reason, the proofs of our results follow the same abstract layout. First, we decompose the error as the contribution of the principal subspace and a mismatch term that measures the alignment between the selected columns and that principal subspace. Second, we show that the mismatch term is bounded by a quantity that is symmetric in the principal angles between the two subspaces; taking expectations, we get a simple closed-form bound.

To present our results, we first denote the number of nonzero k-leverage scores by

$p \leqslant d$, and we quantify the decay of the singular values of $\boldsymbol{X}$ by the flatness parameter

$$\beta = \sigma_{k+1}^2 \left(\frac{1}{d-k} \sum_{j \geqslant k+1} \sigma_j^2 \right)^{-1}. \tag{4.11}$$

In words, $\beta \in [1, d-k]$ measures the flatness of the spectrum of $\boldsymbol{X}$ above the cut-off at k. Indeed, (4.11) 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 (4.11). At the extreme, $\beta = d-k$ when $\sigma_{k+1}^2 > 0$ while $\sigma_j^2 = 0$, $\forall j \geqslant k+2$.

Proposition 4 (Belhadji et al., 2020a) *Under the DPP of kernel* (4.10)*,*

$$\mathbb{E}\|\boldsymbol{X} - \Pi_S^2 \boldsymbol{X}\|_2^2 \leqslant (1 + k(p-k))\|\boldsymbol{X} - \Pi_k \boldsymbol{X}\|_2^2 \tag{4.12}$$

and

$$\mathbb{E}\|\boldsymbol{X} - \Pi_S^{\mathrm{Fr}} \boldsymbol{X}\|_{\mathrm{Fr}}^2 \leqslant \left(1 + \beta \frac{p-k}{d-k} k\right) \|\boldsymbol{X} - \Pi_k \boldsymbol{X}\|_{\mathrm{Fr}}^2. \tag{4.13}$$

The bound in (4.12) compares favorably with volume sampling (4.8) since the dimension d has been replaced by the sparsity level p. For β close to 1, the bound in (4.13) is better than the bound (4.7) of volume sampling since $(p-k)/(d-k) \leqslant 1$. Again, the sparser the k-leverage scores, the smaller the bounds. Finally, if needed, bounds in high probability easily follow from Proposition 4 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 introduce two ideas to further tighten the bounds of Proposition 4. First, we define an effective sparsity level in the vein of (Papailiopoulos et al., 2014) as follows. Let π be a permutation of $\{1, \dots, d\}$ such that leverage scores are reordered

$$\ell_{\pi_1}^k \geqslant \ell_{\pi_2}^k \geqslant \dots \geqslant \ell_{\pi_d}^k. \tag{4.14}$$

For $\delta \in \{1, \dots, d\}$, let $T_\delta = \{\pi_\delta, \dots, \pi_d\}$. Let $\theta \geqslant 1$ and

$$p_{\mathrm{eff}}(\theta) = \min \left\{ q \in \{1, \dots, d\} \Big/ \sum_{i \leqslant q} \ell_{\pi_i}^k \geqslant k - 1 + \frac{1}{\theta} \right\}. \tag{4.15}$$

Note that by definition, the k-leverage scores sum to k, so that the partial sum defining $p_{\mathrm{eff}}(\theta)$ should be close to the total sum. Second, we condition the DPP on a favourable event with controlled probability.

Theorem 8 (Belhadji et al., 2020a) *Let $\mathcal{A}_\theta$ be the event $\{S \cap T_{p_{\mathrm{eff}}(\theta)} = \varnothing\}$. Then, under the DPP of kernel* (4.10)*, on the one hand,*

$$\mathbb{P}(\mathcal{A}_\theta) \geqslant \frac{1}{\theta}, \tag{4.16}$$

and, on the other hand,

$$\mathbb{E}\left[\|\boldsymbol{X} - \Pi_S^2 \boldsymbol{X}\|_2^2 \,\middle|\, \mathcal{A}_\theta\right] \leqslant (1 + (p_{\mathrm{eff}}(\theta) - k + 1)(k - 1 + \theta))\|\boldsymbol{X} - \Pi_k \boldsymbol{X}\|_2^2 \tag{4.17}$$

and

$$\mathbb{E}\left[\|\boldsymbol{X} - \Pi_S^{\mathrm{Fr}}\boldsymbol{X}\|_{\mathrm{Fr}}^2 \,\middle|\, \mathcal{A}_\theta\right] \leqslant \left(1 + \beta\frac{(p_{\mathrm{eff}}(\theta) + 1 - k)}{d-k}(k - 1 + \theta)\right)\|\boldsymbol{X} - \Pi_k\boldsymbol{X}\|_{\mathrm{Fr}}^2. \tag{4.18}$$

In Theorem 8, the effective sparsity level $p_{\mathrm{eff}}(\theta)$ replaces the sparsity level p of Proposition 4. The key is to condition on S not containing any index of 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 $S \sim \mathrm{DPP}(\boldsymbol{K})$ until $S \cap T_{p_{\mathrm{eff}}}(\theta) = \varnothing$. The caveat of any rejection sampling procedure is a potentially large number of samples required before acceptance. But in the present case, Equation (4.16) 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 (4.15), and the complexity of rejection sampling. The approximation bounds suggest picking θ close to 1, which implies a compromise with the value of $p_{\mathrm{eff}}(\theta)$ 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.

Finally, we have observed that the bound in Theorem 8 is representative of the actual behaviour of the mean error, and that many real datasets from natural language processing had sparse leverage scores, so that our DPP is a practical improvement over volume sampling in that setting; see the numerical experiments in (Belhadji et al., 2020a, Section 6).

4.5 A guarantee on the excess risk in sketched linear regression

So far, we have focused on approximation bounds in spectral or Frobenius norm for the residual $\boldsymbol{X} - \Pi_{S,k}^{\nu}\boldsymbol{X}$. 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 $i \in \{1, \dots, N\}$, assume $y_i = \boldsymbol{X}_{i,:}\boldsymbol{w}^* + \xi_i$, where $\boldsymbol{w}^*$ is some ground truth vector of regression coefficients, and the noise variables ξ_i are i.i.d. real variables with mean 0 and variance v. For a given estimator $\boldsymbol{w} = \boldsymbol{w}(\boldsymbol{X}, \boldsymbol{y})$, its excess risk is defined as

$$\mathcal{E}(\boldsymbol{w}) = \mathbb{E}_{\boldsymbol{\xi}}\left[\frac{\|\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{X}\boldsymbol{w}\|_2^2}{N}\right]. \tag{4.19}$$

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

$$\mathcal{E}(\hat{\boldsymbol{w}}) = v \times \frac{\mathrm{rk}(\boldsymbol{X})}{N}. \tag{4.20}$$

For convenience, for $S \subset \{1, \dots, d\}$ of cardinality k, we introduce the matrix $\boldsymbol{S} \in \{0,1\}^{d\times k}$ such that $\boldsymbol{A}\boldsymbol{S} = \boldsymbol{A}_{:,S}$ for any $\boldsymbol{A} \in \mathbb{R}^{N\times d}$. 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 5 (Theorem 9, Mor-Yosef and Avron, 2019) *Let $S \subset [d]$, such that $|S| = k$. Let $(\theta_i(S))_{i\in\{1,\dots,k\}}$ be the principal angles between* $\mathrm{Span}\boldsymbol{S}$ *and* $\mathrm{Span}\boldsymbol{V}_k$*, see (Belhadji et al., 2020a, Appendix C). Then*

$$\mathcal{E}(\hat{\boldsymbol{w}}_S) \leqslant \frac{1}{N}\left(1 + \max_{i\in\{1,\dots,k\}} \tan^2\theta_i(S)\right)\|\boldsymbol{w}^*\|^2\sigma_{k+1}^2 + \frac{vk}{N}. \tag{4.21}$$

Compared to the excess risk (4.20) of the OLS estimator, the second term of the right-hand side of (4.21) replaces $\mathrm{rank}\boldsymbol{X}$ by k. But the price is the first term of the right-hand side of (4.21), which we loosely term *bias*. To interpret this bias term, we first look at the excess risk of the principal component regressor

$$\boldsymbol{w}_k^* \in \mathop{\arg\min}_{\boldsymbol{w}\in\mathrm{Span}\boldsymbol{V}_{:,k}} \mathbb{E}_\xi\left[\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|^2/N\right]. \tag{4.22}$$

Proposition 6 (Corollary 11, Mor-Yosef and Avron, 2019)

$$\mathcal{E}(\boldsymbol{w}_k^*) \leqslant \frac{\|\boldsymbol{w}^*\|^2\sigma_{k+1}^2}{N} + \frac{vk}{N}. \tag{4.23}$$

The right-hand side of (4.23) is almost that of (4.21), except that the bias term in the CSS risk (4.21) 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.

Again, everything is prepared for DPPs to yield a small bound in expectation, since the right-hand side of (4.21) is a symmetric function of the principal angles between the span of the selected columns and the k-principal subspace. Jumping directly to the *effective sparsity* setting of Theorem 8, we use the same rejection sampling trick, and arrive at the following.

Theorem 9 (Belhadji et al., 2020a) *Using the notation of Theorem 8 for leverage scores and their indices, it holds*

$$\mathbb{E}\big[\mathcal{E}(\hat{\boldsymbol{w}}_S)\,\big|\,\mathcal{A}_\theta\big] \leqslant \big[1 + \big(k - 1 + \theta\big)\big(p_{\mathrm{eff}}(\theta) - k + 1\big)\big]\frac{\|\boldsymbol{w}^*\|^2\sigma_{k+1}^2}{N} + \frac{vk}{N}. \tag{4.24}$$

The price of interpretable dimensionality reduction w.r.t. principal component regression is thus the same $\mathcal{O}(1 + k(p-k))$ prefactor as in (4.17) in front of the PCA bias term from (4.23). The price goes down as the k-leverage scores get sparser. Finally, to the best of our knowledge, bounding the excess risk in linear regression has not been investigated under volume sampling.

4.6 Closing remarks

Seeing a projection DPP as a random subspace with an alignment constraint is maybe the most intuitive viewpoint to understand how to build a kernel for a given subsampling application, in my opinion, as long as the target performance measure is somehow a function of that alignment. By expressing alignment in terms of principal angles, it becomes obvious that the determinant of the Gram matrix somehow controls how close a data-aligned subspace (the span of selected columns in this chapter, the span of the feature functions $k(x, \cdot)$ anchored at x a quadrature node in Chapter 3) is to a predefined subspace (the k-principal subspace in this chapter, the subspace of polynomials of bounded degree in Chapter 3, the span of the first eigenfunctions of the RKHS kernel in Section 3.4). We refer to (Kassel and Lévy, 2019) for a more abstract take on this geometric view of DPPs.

In particular, we have been using the same alignment technique to provide a supervised version of the DPP-based column selection presented in this chapter; see (Mai and Bardenet, 2022) for preliminary results. The idea is to sample columns from a DPP which aligns the span of the selected columns with a Krylov subspace, which can be interpreted as a supervised counterpart to PCA.

At a lower level, sampling our PCA-based DPP requires computing cross-leverage scores, i.e. our kernel, which requires running a PCA. If we could lower that cost, our algorithms would become more than a replacement for PCA. There are workarounds to computing leverage scores, like estimating them by random projections (Drineas et al., 2012; Boutsidis et al., 2011). There are also double-phase techniques, where one first preselects marginally relevant columns before finding a diverse subset of the selected ones. The latter family of methods has given algorithms that are extremely efficient in practice (Boutsidis et al., 2009), yet come with loose bounds. It would be interesting to find the right way to pre-screen items that does not impact too much the theoretical performance of a DPP; see e.g. (Dereziński et al., 2019) for results on DPP sampling that relate to this idea.

Chapter 5

Time-frequency filtering with zeros of GAFs

> *Silence fell like a hammer made of feathers. It left holes in the shape of the sound of the sea.*
>
> Terry Pratchett, *Nation*, 2008

Contents

5.1 From the STFT to the planar GAF 45

5.2 From a DPP to analytic wavelets 47

5.2.1 The space of signals 48

5.2.2 The time-scale transform 49

5.2.3 The white noise 49

5.3 Mapping signals to GAFs: a generic template 50

5.4 From the spherical GAF to the Kravchuk transform 52

5.5 Opening remarks 54

5.5.1 Removing the analyticity constraint 54

5.5.2 Making the Kravchuk transform a practical tool 54

5.5.3 The point process of maxima 55

5.5.4 Ridges and level sets 55

5.6 Despite appearances, this chapter fits in my narrative 55

When I arrived in Lille at the beginning of 2015, I joined a signal processing team. From coffee break to lunch time at the *restaurant universitaire* with my excellent colleagues, I traded long monologues on DPPs against a crash course in signal processing. In particular, I developed an interest in time-frequency and time-scale analysis listening to Pierre Chainais and Julien Flamant. When we all went to GRETSI'15 in Lyon, the main French-speaking event in signal and image processing, I saw a plenary talk by Patrick Flandrin, who presented his recent interest in the zeros of spectrograms (Flandrin, 2015), as opposed to the traditional focus on spectrogram maxima.

As an aside, spectrograms are a cornerstone of time-frequency analysis (Flandrin, 1998). They are quadratic time-frequency representations of a signal $f : \mathbb{R} \to \mathbb{C}$

(Gröchenig, 2001, Chapter 4), associating to each time and frequency a real number that measures the energy content of a signal at that time and frequency, unlike global-in-time tools such as the Fourier transform. Since it is natural to expect that there is more energy where there is more information or signal, most methodologies have focused on detecting and processing the local maxima of the spectrogram; see e.g. (Flandrin, 1998). Usual techniques include *ridge extraction*, e.g., to identify chirps, *reassignment* and *synchrosqueezing*, to better localize the maxima before further quantitative analysis. Using the zeros of a spectrogram instead of higher-level sets thus may sound counter-intuitive.

Yet, at GRETSI'15, Patrick Flandrin explained that the locations of the zeros of a spectrogram in the time-frequency plane almost completely characterize the spectrogram, and he proposed to use the point pattern formed by the zeros in tasks such as filtering and reconstruction of signals in noise. This proposition stems from the empirical observation that the zeros of the spectrogram of white noise are uniformly spread over the time-frequency plane, and tend not to cluster, as if they repelled each other. In the presence of a signal, zeros are absent in the time-frequency support of the signal, thus creating large holes that appear to be very rare when observing pure white noise. This leads to testing the presence of signal by looking at statistics of the point pattern of zeros, and trying to identify holes. Patrick's talk connected to all my pet subjects of the time: time-frequency analysis, repulsive point processes, and even reproducing kernel Hilbert spaces. This was the trigger for what has become a large part of my research programme.

The similarity of the mathematical backbone of DPPs with the analytical tools behind time-frequency analysis (Gröchenig, 2001) made me conjecture during GRETSI'15 that the point process formed by the zeros of the spectrogram of white Gaussian noise was a DPP. As we shall see in Section 5.1 and detailed in (Bardenet, Flamant, and Chainais, 2018), this is not the case; the zeros rather follow the distribution of the zeros of the planar Gaussian analytic function introduced in Section 2.3. By the time we finished working on (Bardenet et al., 2018), I was invited to give a talk at GRETSI'19 on DPPs and signal processing. How frustrating that my flagship time-frequency example was *not* a DPP! However, once all the maths were laid out to characterize the zeros of random spectrograms, I realized that I could have obtained a well-known DPP if I had replaced the Hermite functions by the inverse Fourier transform of the Laguerre functions in the proofs. A crucial discussion with Patrick Flandrin helped me realize that my modified proof corresponded to characterizing the zeros not of the classical spectrogram of white noise, but of the analytic wavelet transform (Daubechies and Paul, 1988) of a particular white noise. This is sketched in Section 5.2. A few days later, Adrien Hardy, passing though my office for some other reason (most likely for one of his usual pranks), listened to my musings of the moment and immediately realized the key role implicitly played by generating functions of orthogonal polynomials. This allowed us to further generalize our results on the identification of zeros to more time-frequency transforms, and cleanly define all sorts of white noises (Bardenet and Hardy, 2019). Section 5.3 gives this construction, of which Sections 5.1 and 5.2 can be seen as special cases. In Section 5.4, I show how this generic construction led us to define a new time-frequency transform, acting on finite-dimensional vectors (Bardenet and Hardy, 2019; Pascal and Bardenet, 2022). The topic of zeros of spectrograms has now gained momentum, and there are many open questions, some of which I list in Section 5.5.

Finally, while this chapter may seem out of my initial narrative for this manuscript (i.e., *repulsive point processes provide small residuals of orthogonal projections in expectation*), I show in Section 5.6 that this is not the case, and that we are somehow pursuing the same goal as in Chapters 3 and 4, only in a different form.

For reference, this chapter is based on our works (Bardenet, Flamant, and Chainais, 2018; Bardenet and Hardy, 2019), and (Pascal and Bardenet, 2022).

5.1 From the STFT to the planar GAF

Let $f \in L^2(\mathbb{R})$ and g be the pdf of a centered Gaussian, with variance normalized so that its L^2 norm is $\|g\|_2 = 1$. The short-time Fourier transform (STFT) $V_g f$ of f with Gaussian window g, is defined by

$$V_g f(u,v) = \int f(t)\overline{g(t-u)}e^{-2\mathrm{i}\pi tv}\mathrm{d}t = \langle f, M_v T_u g\rangle, \quad (u,v) \in \mathbb{R}^2, \tag{5.1}$$

with $\langle \cdot, \cdot \rangle$ denoting the inner product in $L^2(\mathbb{R})$, $M_v f = e^{2\mathrm{i}\pi v \cdot} f(\cdot)$ and $T_u f = f(\cdot - u)$. The spectrogram of f is simply the squared modulus of its STFT, i.e.

$$u, v \mapsto |V_g f(u,v)|^2.$$

From (5.1), it is intuitive that the spectrogram is large in (u,v) if frequency v is contributing to the signal around time u. An example spectrogram is shown in Figure 5.1(a). The signal is a so-called *linear chirp*, as can be seen guessed from the linear relation between time and frequency. The (numerical) zeros of the spectrogram look well-spread, with a slight accumulation around the signal support, and a complete absence in the support, which the spectrogram takes large values.

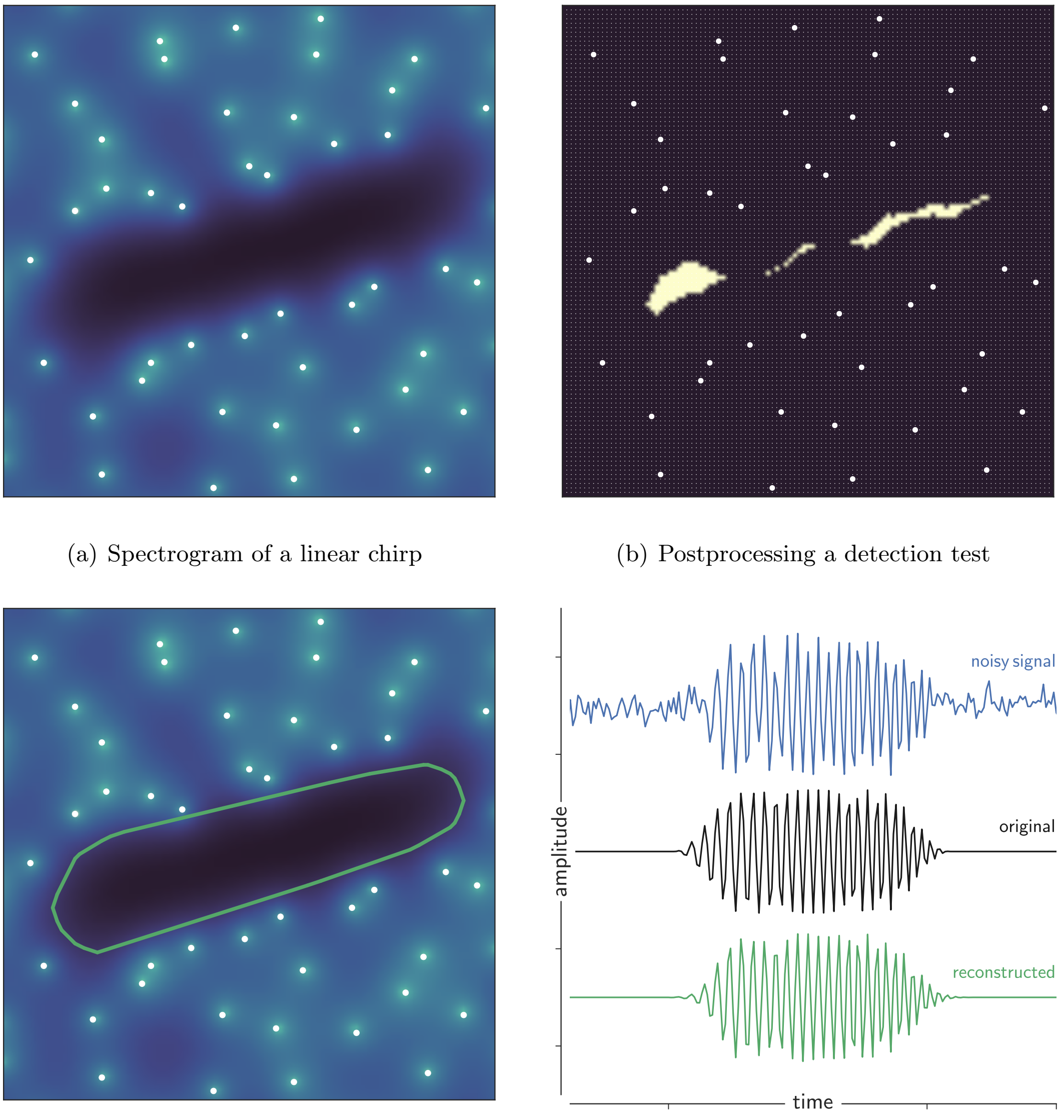
To investigate the zeros of the STFT of white noise, our first step was to rigorously define the STFT of complex white Gaussian noise (WGN), in a manner that both satisfied the intuition we may have of white noise, and allowed us to speak of the zeros of the spectrogram. In (Bardenet et al., 2018, Section 3.1), we used a classical definition of real WGN as a random tempered distribution with Gaussian characteristic function. A complex WGN ξ is then defined as having independent white Gaussian noises ξ_1, ξ_2 as its real and imaginary parts.

This allowed us to *pointwise* compute $\langle \xi, M_v T_u g\rangle$ since $M_v T_u g$ is smooth. By pointwise, we mean for a given time u and frequency v. Then in (Bardenet et al., 2018, Proposition 3), we stitched all these pointwise evaluations together and concluded that with probability 1 on the white noise, the zeros of the function

$$z = u + \mathrm{i}v \mapsto \langle \xi, M_v T_u g\rangle$$

are those of so-called *planar* GAF,

$$\mathsf{GAF}_{\mathbb{C}}(z) := \sum_{k=0}^{\infty} a_k \frac{\pi^{k/2} z^k}{\sqrt{k!}}, \tag{5.2}$$



(a) Spectrogram of a linear chirp

(b) Postprocessing a detection test

(c) Identification of the signal support

(d) Signal reconstruction

Figure 5.1: The different steps of a zero-based reconstruction algorithm, applied to a noisy signal. The darker the blue, the larger the spectrogram value, while zeros are shown as white dots.

where $a_k = (\langle \xi_1, h_k\rangle + \mathrm{i}\langle \xi_2, h_k\rangle)/\sqrt{2}$ are i.i.d. unit complex Gaussians, and h_k are the Hermite functions. To read this chapter, it is enough to know that Hermite functions are a basis of $L^2(\mathbb{R})$ built using orthogonal polynomials with respect to the Gaussian g.

All the above was the formalisation of the following heuristic. A complex WGN should have unit complex Gaussian coefficients in any basis of $L^2(\mathbb{R})$. Among such bases, we pick the Hermite basis, because Hermite functions essentially map to complex monomials through the STFT (Gröchenig, 2001). We thus expect an analytic function of the form (5.2) as the STFT of ξ.

Identifying the zeros of the STFT of white noise with the zeros of a random entire function on the time-frequency plane is satisfactory, since entire functions have isolated zeros. It thus makes sense to speak of the point process of the zeros. Furthermore, the zeros of (5.2) are actually a well-studied stationary point process (Hough et al., 2009), called the zeros of the planar Gaussian analytic function (GAF), and which we introduced in Section 2.3. We show a sample of these zeros as white dots in Figure 5.2(a), with the heatmap showing the corresponding spectrogram realization.[1] Combining known probabilistic results on this point process and standard statistical methodology for spatial point processes, we obtained in (Bardenet et al., 2018, Section 5) powerful signal detection procedures that strengthen the seminal work of Flandrin (2015). Without entering into details, spatial statistics provides a sophisticated toolbox for one-sample testing with stationary point processes (Baddeley et al., 2015). In Figure 5.1(b), the beige pixels correspond to the centers of unusually large Euclidean balls that do not contain any zero of the spectrogram. By *unusual*, I mean that under the hypothesis that we are observing the zeros of the spectrogram of pure white noise, the corresponding empty balls appear with very small probability. These empty balls have caused to reject the hypothesis of pure white noise. One can then apply any heuristic to extract an envelope of these beige points; see Figure 5.1(c), and invert the STFT after multiplication with a smooth indicator function of the support. This yields a denoised, or "filtered", signal; see Figure 5.1(d).

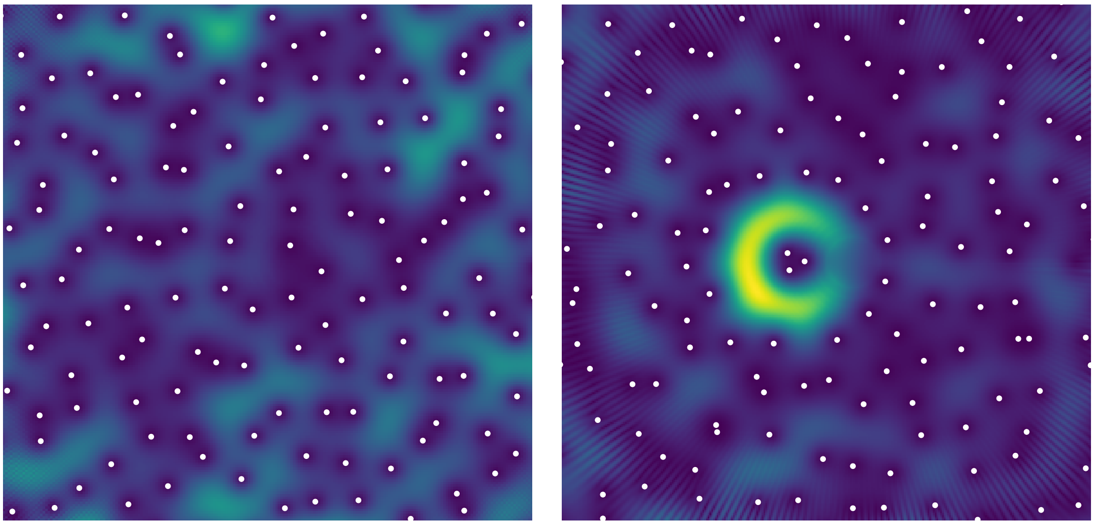
To conclude, one can show (Hough et al., 2009, Section 5.1) that the zeros of the planar GAF are not a DPP with Hermitian kernel. Indeed, the pair correlation function of the zeros of the planar GAF is known in closed form, and it takes values above 1. For a DPP with Hermitian kernel, this is forbidden; see Section 2.2.5. Strictly speaking, they could still be a DPP with non-Hermitian kernel.

5.2 From a DPP to analytic wavelets

We were slightly frustrated of not finding a well-known DPP behind the STFT of white noise in (Bardenet et al., 2018). In particular, for $\alpha > -1$, consider the random analytic function

$$\mathsf{GAF}_{\mathbb{D}}^{(\alpha)}(z) := \sum_{k\in\mathbb{N}} a_k \sqrt{\frac{\Gamma(k+\alpha+1)}{k!}} z^k, \quad |z| < 1, \tag{5.3}$$

[1] Actually, this specific figure was obtained through a different time-frequency transform giving the same distribution; see (Bardenet and Hardy, 2019).



(a) White noise

(b) White noise with a Hermite function

Figure 5.2: White dots are the zeros of the spectrogram of (a) a realization of white noise, and (b) a realization of white noise plus an Hermite function.

where the a_i are again i.i.d. unit complex Gaussians. The series (5.3) defines the so-called *hyperbolic GAF* (Hough et al., 2009). When $\alpha = 0$, it simply reads $\sum a_k z^k$ and its zeros are a DPP. A natural[2] question is then to know what signal processing problem we should have tried to solve in order to end up with the zeros of the hyperbolic GAF instead of the planar GAF. The gamma coefficients are a hint that the Hermite functions should somehow be replaced by Laguerre functions. This backward reasoning is the starting point of (Bardenet and Hardy, 2019).

The same heuristic as in Section 5.1 holds: we need a basis of some Hilbert space of signals, a white noise that can be decomposed onto that basis, and a time-frequency transform that sends elements of the basis onto the right monomials.

5.2.1 The space of signals

Laguerre functions are a basis of $L^2(\mathbb{R}_+)$ built using the orthogonal polynomials with respect to the pdf of a gamma random variable. Thus, we are looking for a space of signals where the half-line $\mathbb{R}_+$ plays a role. A natural[3] example is the space of "analytic" signals

$$H^2(\mathbb{R}) := \{f \in L^2(\mathbb{R}, \mathbb{C}) : \operatorname{Supp}(\hat{f}) \subset \mathbb{R}_+\}.$$

The inverse Fourier transforms (f_k) of the Laguerre functions form a basis of $H^2(\mathbb{R})$.

[2] At least to DPP aficionados

[3] This time, to signal processers!

5.2.2 The time-scale transform

Now, we need a transform that maps the basis (f_k) to monomials. Let $\beta > -1/2$ and set

$$\psi_\beta(t) := \frac{1}{(t+\mathrm{i})^\beta}, \quad t \in \mathbb{R}.$$

The Fourier transform of ψ_β is essentially a gamma pdf

$$\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}_+} x^\beta \mathrm{e}^{-x} \mathrm{e}^{\mathrm{i}tx} \mathrm{d}x = -\frac{1}{\sqrt{2\pi}} \frac{(-\mathrm{i})^{\beta+1}}{(t+\mathrm{i})^{\beta+1}} \Gamma(\beta+1),$$

so that ψ_β is intuitively a fine window if we want to use Laguerre calculus in the Fourier domain. The Daubechies-Paul wavelet transform of $f \in H^2(\mathbb{R})$ is defined as

$$W_\beta f(u,s) := \langle f, T_u D_s \psi_\beta \rangle, \tag{5.4}$$

where $u \in \mathbb{R}$ and $s \in \mathbb{R}_+^*$ are thought of as time and scale, $T_u f(t) := f(t-u)$ and $D_s f(t) := s^{-1/2} f(t/s)$ are the usual translation and dilation operators.

We show in (Bardenet and Hardy, 2019, Theorem 2.3) that for $f \in H^2(\mathbb{R})$ and up to a non-vanishing term,

$$W_\beta f(-u,s) \propto \mathscr{L}_{\mathbb{D}}^{(\beta)} f(\varphi(u+\mathrm{i}s)), \tag{5.5}$$

where $u + \mathrm{i}s \in \mathbb{C}_+$ is in the time-scale half-plane, $\varphi(w) := (2w-\mathrm{i})/(2w+\mathrm{i})$ is a conformal mapping from $\mathbb{C}_+$ to $\mathbb{D}$, and $\mathscr{L}_{\mathbb{D}}^{(\beta)}$ is a transform that precisely sends the inverse Fourier transforms of the Laguerre functions onto the right monomials

$$\mathscr{L}_{\mathbb{D}}^{(\beta)} f_k(z) = \sqrt{\frac{\Gamma(k+2\beta+1)}{k!}} z^k. \tag{5.6}$$

Everything is in place but the white noise. Heuristically, if we were able to somehow decompose a white noise ξ on $H^2(\mathbb{R}_+)$ as $\sum_{k \geqslant 0} a_k f_k$, where a_k are i.i.d. unit complex Gaussians, then (5.5) and (5.6) should help us conclude that

$$W_\beta \xi(z) \propto \sum_{k \geqslant 0} a_k \sqrt{\frac{\Gamma(k+2\beta+1)}{k!}} z^k = \mathsf{GAF}_{\mathbb{D}}^{(2\beta)}(z)$$

up to non-vanishing terms. This would characterize the zeros of the analytic wavelet transform of white noise.

5.2.3 The white noise

The series $\sum_{k \geqslant 0} a_k f_k$ almost surely diverges in $H^2(\mathbb{R})$. To remedy this, we follow Gross (1967) in his formalisation of abstract Wiener spaces. First, we define a slightly weaker norm on $H^2(\mathbb{R})$ that ensures the series converges

$$\|f\|_\Theta^2 := \sum_{k \in \mathbb{N}} \frac{1}{1+k^2} \left|\langle f, f_k \rangle\right|^2. \tag{5.7}$$

Indeed, it follows that

$$\left\| \sum_{k\in\mathbb{N}} a_k\, f_k \right\|_{\Theta}^2 = \sum_{k\in\mathbb{N}} \frac{|\xi_k|^2}{1+k^2} < \infty \quad \text{a.s.} \tag{5.8}$$

since the right-hand side has finite expectation. Now define Θ as the completion of $H^2(\mathbb{R})$ for the norm (5.7). Leonard Gross' construction (Gross, 1967) then gives a unique probability measure on Θ, called white Gaussian noise, such that

$$\int_{\Theta} \mathrm{e}^{\mathrm{i}\lambda(\theta)} \mu(\mathrm{d}\theta) = \mathrm{e}^{-\|h_\lambda\|^2/2}, \qquad \lambda \in \Theta^*, \tag{5.9}$$

where Θ^* is the topological dual of Θ and $h_\lambda \in H^2(\mathbb{R})$ is associated to $\lambda \in \Theta^* \subset H^2(\mathbb{R})$ by Riesz' representation theorem. Equation (5.9) is to be interpreted as specifying the Fourier transform of the white noise on $H^2(\mathbb{R})$. In particular, it leads to $\sum a_k f_k$ converging almost surely in Θ to μ, which is our goal. We can even give rates of convergence (Bardenet and Hardy, 2019, Section 5). We also note in passing that μ is a rigorous definition of an analytic white noise.

With the white noise now defined, we prove in (Bardenet and Hardy, 2019, Theorem 2.3) that $\mathscr{L}_{\mathbb{D}}^{(\beta)}\xi$ can be unambiguously defined as a random analytic function in the unit open disk $\mathbb{D}$. Furthermore, it is equal in distribution to $\mathsf{GAF}_{\mathbb{D}}^{(2\beta)}$. In particular, the case $\beta = 0$ yields zeros that are a DPP! A sample of this DPP obtained as the numerical zeros of a Paul-Daubechies scalogram is shown in Figure 5.3.

Finally, we note that Abreu, Haimi, Koliander, and Romero (2018) have also independently proved that the zeros of $W_\beta\xi$ are the zeros of the hyperbolic GAF, with a slightly different definition of white noise.

5.3 Mapping signals to GAFs: a generic template

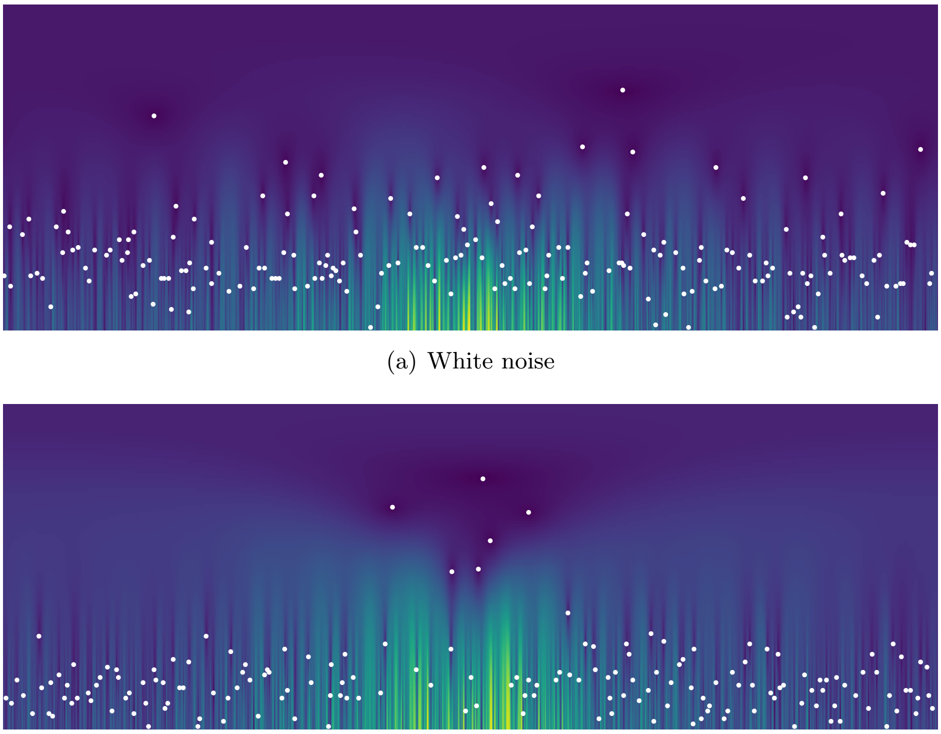
Having extended the STFT case to the analytic wavelet transform, we then realized that the mathematical construction in Section 5.2 was much more general than these two cases. The whole idea is to first fix a Hilbert space $\mathcal{H}$ of signals, fix an orthonormal basis (f_k) of $\mathcal{H}$, and complete $\mathcal{H}$ into Θ as in Section 5.2.3, so as to be able to define a white Gaussian noise on Θ. Then pick a sequence $(\Psi_k)_{k\in\mathbb{N}}$ of holomorphic functions on a domain Λ satisfying

$$\sup_{z\in K} \sum_{k\in\mathbb{N}} |\Psi_k(z)|^2 < \infty \tag{5.10}$$

for any compact $K \subset \Lambda$. Now define a "time-frequency" transform $\mathscr{L}$ acting on $\mathcal{H}$ by

$$\mathscr{L}f(z) := \sum_{k=0}^{\infty} \langle f_k, f\rangle \Psi_k(z), \qquad z \in \Lambda. \tag{5.11}$$

In signal processing terms, the signal $f \in \mathcal{H}$ is analyzed by $\mathscr{L}$ in the basis (f_k), and reconstructed as an analytic function on Λ using the dictionary Ψ_k. Note that dominated



(a) White noise

(b) An inverse Fourier transform of Laguerre function in white noise

Figure 5.3: White dots are a sample zero set of the hyperbolic GAF, superimposed on an analytic scalogram of an analytic white noise.

convergence allows us to rewrite $\mathscr{L}$ in kernel form,

$$\mathscr{L}f(z) := \langle T(\cdot, z), f\rangle, \quad T(x,z) := \sum_{k=0}^{\infty} f_k(x)\Psi_k(z). \tag{5.12}$$

The construction of white noise in Section 5.2.3 guarantees that $\mathscr{L}$ extends uniquely to Θ and that $\mathscr{L}\xi$ has the same law as the GAF $\sum a_k\Psi_k$. This is the main result of Bardenet and Hardy (2019).

The STFT with Gaussian window and the analytic wavelet transform of Daubechies and Paul (1988) are particular cases of $\mathscr{L}$, where the representation (5.12) is provided by identities for the generating functions T of classical orthogonal polynomials. This prompted us to plug other generating functions for orthogonal polynomials as kernels, and check whether the resulting transform had a time-frequency/scale interpretation.

5.4 From the spherical GAF to the Kravchuk transform

Armed with the general pattern in Section 5.3, it is easy to find, given a GAF, a space $\mathcal{H}$ of signals, an analysis basis (f_k) and a reconstruction family (ψ_k) to map white Gaussian noise in $\mathcal{H}$ to any GAF. To the signal processer, the interesting question is whether the resulting transform has any time-frequency interpretation.

Take for instance the *spherical* GAF

$$\mathsf{GAF}_{\mathbb{S}}^{(N)}(z) := \sum_{k=0}^{N} \xi_k \sqrt{\binom{N}{k}}\, z^k, z \in \mathbb{C}. \tag{5.13}$$

We plot in Figure 5.4 a sample of the spherical GAF and its zeros, mapped onto a sphere through stereographic projection. This rendering highlights the invariance of the set of zeros to isometries of the sphere, just like the zeros of the planar and hyperbolic GAFs are invariant to isometries of the plane and the Poincaré half-plane, respectively.

Following Section 5.3, we propose to take $\mathcal{H} = \mathbb{C}^N$ the space of finite signals, along with the basis of Kravchuk functions. In words, Kravchuk functions are to the binomial distribution what Hermite functions are to the Gaussian distribution, and what Laguerre functions are to the gamma distribution. Taking

$$\psi_k(z) = \sqrt{\binom{N}{k}}\, z^k$$

then leads to a transform $\mathscr{L}$ in (5.11) that maps the white noise on $\mathbb{C}^N$ to the spherical GAF. Moreover, the kernel T in (5.12) of this transform has a closed form, due again to a generating identity for Kravchuk polynomials; see the discussion in (Bardenet and Hardy, 2019). Yet, it is not easy to find a time-frequency interpretation of such a transform. To that end, in (Pascal and Bardenet, 2022), we further modified the Kravchuk transform $\mathscr{L}$ to

$$T\boldsymbol{y}(z) = \sqrt{(1+|z|^2)^{-N}} \times \mathscr{L}\boldsymbol{y}(z). \tag{5.14}$$

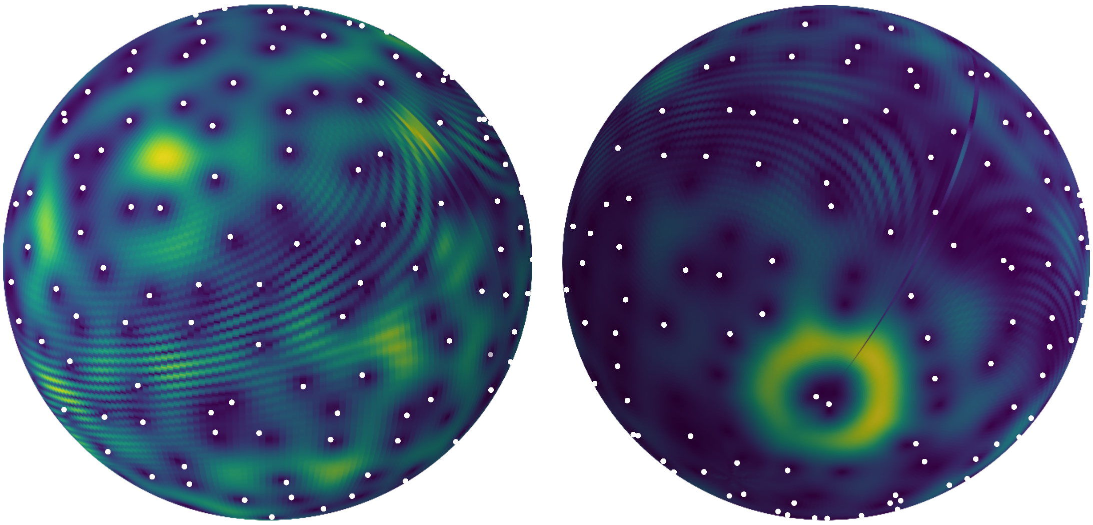
Note that $T\boldsymbol{y}$ is not analytic anymore, but that it has the same zeros as $\mathscr{L}\boldsymbol{y}$. Moreover, T is an isometry in the sense that

$$\|\boldsymbol{y}\|_2^2 = (4\pi)^{-1} \int_{S^2} |T\boldsymbol{y}(\vartheta,\varphi)|^2 \,\mathrm{d}\mu(\vartheta,\varphi), \tag{5.15}$$

where $(\vartheta,\varphi) \in [0,\pi] \times [0,2\pi]$ are the spherical coordinates parametrizing the sphere S^2, $\mathrm{d}\mu(\vartheta,\varphi) = \sin(\vartheta)\mathrm{d}\vartheta\mathrm{d}\varphi$ is the uniform measure on S^2, and we write $T\boldsymbol{y}(\vartheta,\varphi)$ for $T\boldsymbol{y}$ evaluated at the stereographic projection $z = \cot(\vartheta/2)\mathrm{e}^{\mathrm{i}\varphi}$. More importantly, using the formalism of *coherent states*, we show in (Pascal and Bardenet, 2022) that T is *covariant* under the action of the group of rotations SO(3), meaning that

$$T[\boldsymbol{R_u}\boldsymbol{y}](\vartheta,\varphi) = T\boldsymbol{y}(R_{\boldsymbol{u}}(\vartheta,\varphi)), \tag{5.16}$$

where $\boldsymbol{R_u}$ (resp. $R_{\boldsymbol{u}}$) denotes a particular action of a rotation parametrized by the unitary vector $\boldsymbol{u} \in \mathbb{R}^3$ on vectors of size $N+1$ (resp. on points of the unit sphere). We refer the reader to the supplementary material to (Pascal and Bardenet, 2022) for a concise introduction to the representation theory of SO(3).



(a) White noise

(b) A Kravchuk function in white noise

Figure 5.4: White dots are a sample zero set of the spherical GAF, superimposed on a Kravchuk spectrogram, see (Bardenet and Hardy, 2019; Pascal and Bardenet, 2022) for definitions.

This covariance property (5.16) is meaningful to the signal processer. The classical STFT is covariant with respect to translations in the time-frequency plane (Gröchenig, 2001). In words, translating a signal in time and frequency before computing the STFT yields the same result as evaluating the STFT of the original signal at a translated time and frequency. Even if the new variables (ϑ,φ) do not have as intuitive a meaning as time and frequency, they are still variables in a phase space (the sphere, instead of the plane), and there is a group of isometries (rotations, instead of time-frequency translations, a.k.a.

the Weyl-Heisenberg group), with respect to which the Kravchuk transform is covariant. That makes the Kravchuk transform close in spirit to the classical short-time Fourier transform.

5.5 Opening remarks

We close this section with a few thoughts, and remarks on ongoing work.

5.5.1 Removing the analyticity constraint

There are some fundamental limitations to the correspondence between GAFs and time-frequency transforms. The three GAFs we introduced – planar, hyperbolic, and spherical – are canonical in the sense that they are the only GAFs that are invariant to isometries of the complex plane, the Poincaré half-plane, and the sphere. From a statistical point of view, such invariances are crucial, as they make the point process of the zeros stationary and allow estimating functional statistics, a key step in designing detection/reconstruction algorithms (Bardenet et al., 2018, Section 5). Extending the correspondence will require either more exotic geometries, or dropping the Gaussian/analytic aspects on the GAF side.

Another point in favour of removing analyticity is that the Gaussian window is the only window that makes the STFT map $L^2(\mathbb{R})$ to analytic functions (Ascensi and Bruna, 2009). When the window is not Gaussian but still Schwartz, the STFT maps to polyanalytic functions, and Haimi et al. (2020) have studied the zeros of the STFT of white noise in that case. Interestingly, the analytic wavelet transform of Section 5.2 is also the only continuous wavelet transform that maps to analytic functions (Holighaus et al., 2019). Besides working with polyanalytic functions, discrete transforms such as the Kravchuk transform (Bardenet and Hardy, 2019; Pascal and Bardenet, 2022) do bypass constraints on the window while still mapping to analytic functions.

5.5.2 Making the Kravchuk transform a practical tool

We believe that the Kravchuk transform of Section 5.4 has the potential of becoming a standard item in the practitioner's toolbox. When describing the classical STFT, we have swept under the rug the fact that any practical computation is done by approximating the involved integrals by Riemann sums, and then applying fast Fourier transforms. When Riemann sums are a poor approximation to the corresponding integrals, say when the sampling frequency is lower than typical frequencies in the signal, the numerical zeros of the spectrograms we compute in practice are likely to be far from the zeros of the underlying GAFs. In contrast, the Kravchuk transform takes as inputs *finite* signals, i.e. (finite-length) vectors of complex numbers. If we accept to work with the sphere as a phase space, the Kravchuk transform avoids any sophisticated approximation procedure of functions by finite sequences of numbers. We refer to (Pascal and Bardenet, 2022) for numerical experiments that demonstrate that signal detection tests for small signals are more powerful when using the Kravchuk transform than the STFT.

This situation is reminiscent of what the finite Fourier transform is to the continuous Fourier transform, where replacing the latter by the former requires some intellectual gymnastics (e.g., consider only band-limited functions, use Shannon's sampling theorem, and truncate the coefficients) and edge effects that are not easy to quantify. We are also currently working with Barbara Pascal and Julien Flamant on fast algorithms to compute and invert the Kravchuk transform, leveraging analogies with spin spherical harmonics.

5.5.3 The point process of maxima

Maxima of spectrograms are also the zeros of a Gaussian process, but with non-analytic samples. Abreu (2022) shows that the maxima still form a point process, are stationary, and that their intensity if 1/3 that of zeros, in line with a conjecture of Flandrin (2015). A question of practical importance is if, when, and how there is an interest in using maxima, zeros, or both for a given signal processing task. Preliminary experiments suggest that maxima alone are more efficient than zeros alone at, say, estimating the instantaneous frequency of a simple signal with large signal-to-noise ratio and little interference with other signal components. On the contrary, the more rigid structure of zeros (due to being the zeros of an analytic function) makes them more robust than maxima at low signal-to-noise ratio, or when there is strong interference caused by multiple signals with overlapping time-frequency supports. This is the topic of ongoing work with my former postdoc Arnaud Poinas, and my current postdoc Juan Manuel Miramont, along with co-supervisors Pierre Chainais and François Auger; see (Miramont et al., 2022) for preliminary results on a numerical benchmark where we compare maxima-based and zero-based detection and reconstruction.

5.5.4 Ridges and level sets

For estimation of instantaneous frequencies, it is actually customary to work with the maxima not of the spectrogram, but of each restriction of the spectrogram to a constant-time line in the time-frequency plane; sort of *synchrosqueezed* maxima. This gives curves of maxima, so-called *ridges*, and the random set of ridges of a (random) spectrogram would be an interesting object to study. Similarly, many geometric properties of the spectrogram could be used for filtering or instantaneous frequency estimation. Ghosh et al. (2022), for instance, showed that level sets of spectrograms can be used to detect signals made of single Hermite functions with a quantifiable testing power.

5.6 Despite appearances, this chapter fits in my narrative

My narrative is that functions can be efficiently reconstructed if sampled at carefully chosen repulsive point processes. In particular, functions that belong to a reproducing kernel Hilbert space (RKHS), sampled at a DPP or a mixture of DPPs, were shown to yield small residuals in expectation under a suitable DPP and a mixture of DPPs in Chapter 3. We are in a very similar setting here, and we could even make the settings coincide.

Let me consider the STFT V_g from (5.1), with Gaussian window g for concreteness. The range $V_g(L^2)$ of the STFT with Gaussian window is an RKHS – the so-called *Bargmann-Fock space* of quantum physics. Second, time-frequency filtering is usually carried out by projection onto a subset of time-frequency atoms, just like we approximated mean elements by sums of feature functions in Chapter 3. Indeed, as we saw in Figure 5.1, filtering works by localizing a signal, i.e. (i) isolating a subset $\Omega \subset \mathbb{R}^2$ of the time-frequency plane (the assumed "support" of the signal, delimited in green in Figure 5.1(c)), and (ii) reconstructing a signal $\hat{f}$ through the restriction of the STFT of the (noisy) signal f to that subset,

$$\hat{f} = V_g^{-1}\left[\mathbf{1}_\Omega(u,v)V_g f(u,v)\right]. \tag{5.17}$$

In words, we select the (typically uncountable) subset of time-frequency atoms indexed by Ω, and approximate the original signal by a linear combination of these atoms. The proposition of Flandrin (2015) can be paraphrased as defining the subset Ω as a function of the pattern of zeros of $V_g(f)$. Just like in Chapters 3 and 4, the span of the time-frequency atoms indexed by Ω is indeed a random subset that is defined through a repulsive point process. Three differences with Chapter 3 are that the selected random subset is uncountable, that the localization operator in (5.17) is not an orthogonal projection, and that the involved point processes are not DPPs.

There are a number of steps that we could take to prove a result like those in Chapter 3 and 4 here. First, we could use the zeros (u_i, v_i) of the spectrogram of white noise to design a (countable) random dictionary $(M_{v_i}T_{u_i}g)$ of time-frequency atoms, and prove that it gives a (random) frame with good properties (Gröchenig, 2001). Second, we could study the frame given by a DPP in the time-frequency plane with kernel that of the Bargmann-Fock space $V_g(L^2)$. It turns out that this DPP is precisely the Ginibre ensemble introduced in Example 3. Note that Ghosh (2015) gave a clever indirect proof that the set of time-frequency atoms indexed by the Ginibre ensemble is a complete set; a natural question is thus to know whether it is a frame with tighter constants than the independent draws of the seminal (Bass and Gröchenig, 2013). Playing with the spectrum of the Ginibre kernel, say using the truncated Ginibre ensemble of Example 3, we should also end up with a *finite* set of time-frequency atoms with properties resembling that of the infinite Ginibre frame. Such a result would be of practical importance, and is the subject of ongoing collaboration with Pierre Chainais and our former PhD student Ayoub Belhadji.

Chapter 6

Other work and perspectives

> *Nanny's philosophy of life was to do what seemed like a good idea at the time, and do it as hard as possible. It had never let her down.*
>
> Terry Pratchett, *Maskerade*, 1995

Contents

6.1 Past work on repulsive point processes 57
6.2 Other work and a few perspectives 59

I start by reviewing work on point processes that I chose not to include in the main body of this manuscript for conciseness, but that still relates to the overall investigation of the computational and statistical properties of repulsive point processes. Then I discuss current and future work, with a few perspectives that progressively go beyond my focus on point processes.

6.1 Past work on repulsive point processes

I have voluntarily focused on ideas and mathematical results in the main body of this manuscript, to sketch the common umbrella above my work. That being said, I have also worked on methodological, algorithmic, statistical, and experimental aspects of repulsive point processes.

On methodology, recent work that could easily have made it to the main body includes (Bardenet, Ghosh, and Lin, 2021). We show there how to use the variance reduction that we obtained for multivariate orthogonal polynomial ensembles in (Bardenet and Hardy, 2020), see also Chapter 3, but this time for subsampling a *finite-but-large* set. Our target application is stochastic gradient learning, where random subsets of a large training set (so-called *minibatches*) are used to estimate a gradient, at each iteration of a gradient descent applied to an empirical loss function. Since the variance of the gradient estimator controls the mean square error of interest (Moulines and Bach, 2011), variance reduction for gradient estimators has been a rich topic of research. We propose in (Bardenet, Ghosh, and Lin, 2021) a DPP-based estimator of the gradient of the empirical loss, with a mean square error that decreases faster than the inverse of the square root of the cardinality of the minibatch. This paper was selected as a spotlight at NeurIPS'21.

On algorithmic aspects of repulsive point processes, we have focused on sampling algorithms. Indeed, for a projection DPP of rank N, the cost of the standard algorithm is at least cubic in N; see Chapter 2. Like for other kernel machines, this cubic cost typically becomes prohibitive as soon as the required N goes beyond 10^4. There are several avenues around this bottleneck. On our side, we have worked on approximate sampling through Markov chain Monte Carlo for specific DPPs in the thesis of Guillaume Gautier (Gautier, 2020). In particular, we have proposed an MCMC algorithm to sample from finite projection DPPs (Gautier, Bardenet, and Valko, 2017), which we have experimentally demonstrated to mix remarkably fast. We have proposed a similar MCMC kernel for β-ensembles (Gautier, Bardenet, and Valko, 2021). Finally, with my postdoc Michaël Fanuel, we have studied a physics-inspired DPP on the edges of a graph (Fanuel and Bardenet, 2022), for which exact sampling is possible without relying on the "base times height" formula of Chapter 2. This DPP is a natural generalization of the edges of a uniform spanning tree, a well-studied DPP (Lyons and Peres, 2016). Our generalization replaces the Laplacian of the graph by a so-called connection Laplacian, and has applications in, e.g., ranking. The reader can also check our recent preprint (Jaquard, Fanuel, Amblard, Bardenet, Barthelmé, and Tremblay, 2022) for an application of that connection-graph DPP to graph signal processing.

On statistical aspects, we have worked first on inference for DPPs, where the latter are used as models for repulsive behaviour. Repulsive point processes are natural models for, e.g., the location of trees in a forest, which compete for nutrients, or sentences in a summary, whose diversity guarantees coverage. Among repulsive point processes, DPPs with Hermitian kernels stand out as exceptionally tractable: sampling is polynomial, and all correlation functions are available in closed form, which allows using them in inference procedures. In particular, we have proposed MCMC and variational algorithms for Bayesian inference in parametric DPP models (Bardenet and Titsias, 2015). More recently, building on a novel representation for non-negative functions in RKHSs (Marteau-Ferey, Bach, and Rudi, 2020), we have proposed the first non-parametric algorithm to learn a continuous kernel from DPP samples (Fanuel and Bardenet, 2021). Still within statistics, but closer to our motto of using DPPs as subsampling tools, we have investigated DPPs for continuous experimental design (Poinas and Bardenet, 2022). Finally, we have also investigated statistical diagnostics of *hyperuniformity* (Hawat et al., 2022). Informally, a hyperuniform point process is one for which the variance of Monte Carlo estimators decreases faster than the classical Monte Carlo rate. Being able to infer that property can help us focus our mathematical efforts on promising point processes.

On experimental and software aspects, we developed a software library[1] intended to collect all known sampling algorithms for DPPs, exact and approximate (Gautier, Bardenet, Polito, and Valko, 2019a). We did the same for estimators of the *structure factor* of a point process,[2] a key quantity to test hyperuniformity, which is especially useful to material scientists. Finally, we are developing a benchmarking suite for signal detection and reconstruction with time-frequency tools.[3] One of our objectives there is to numeri-

[1] https://github.com/guilgautier/DPPy

[2] https://github.com/For-a-few-DPPs-more/structure-factor

[3] https://github.com/jmiramont/benchmark-test

cally assess in what regime algorithms based on the zeros of the spectrogram shine, and how to combine zero- and maxima-based algorithms.

6.2 Other work and a few perspectives

The Monte Carlo algorithms implicitly described in Chapter 3 are, in a sense, importance sampling algorithms with non-i.i.d. proposals. The resulting mean square errors, for volume sampling, or asymptotic variance, in the case of orthogonal polynomial ensembles, decreases fast with the size of the sample, but actually scales exponentially in the dimension of the space over which we wish to integrate. It is thus highly desirable to investigate whether a *dynamic* version of our results, involving e.g. the trajectory of a Markov process leaving a repulsive point process invariant, could lead to better scaling with the dimension. Indeed, having asymptotic variances that grow polynomially with the dimension is one of the theoretical reasons for the success of MCMC algorithms (Chernozhukov and Hong, 2003). With that in mind, Mylène Maida (Laboratoire de mathématiques Paul Painlevé, Univ. Lille) and I are investigating piecewise deterministic Markov processes that leave repulsive point processes invariant. Relatedly, with our new jointly-supervised doctoral student Martin Rouault (2022–), Mylène and I are investigating other types of Gibbs point processes than DPPs for Monte Carlo integration. These other point processes are partly chosen for their amenability to MCMC sampling.

With Subhro Ghosh (Dept. of mathematics and dept. of data science, NUS Singapore) and aspiring PhD student Hugo Simon, we are investigating variance reduction for discrete DPPs in machine learning, in the vein of our work on stochastic gradient (Bardenet, Ghosh, and Lin, 2021). With Subhro again, we are also investigating methodological alternatives to sampling DPPs (Bardenet and Ghosh, 2020), and topological data analysis applied to the zeros of time-frequency transforms.

With PhD student Diala Hawat (2020–) and her co-supervisor Raphaël Lachièze–Rey (Laboratoire de mathématiques appliquées MAP5, Univ. Paris), we are currently investigating how to obtain variance reduction with repulsiveness applied *a posteriori*. Indeed, to bypass the cost of sampling a DPP with large cardinality, one option is to sample points independently, and then push them apart. Diala is currently proving (and is almost there) that a well-chosen gradient step applied to all particles, pushing them away from each other, reduces the variance of Monte Carlo estimators based on these particles.

After reading Odile Macchi's thesis, where she introduced DPPs as models to explain the anti-bunching properties of beams of fermionic particles, I grew very enthusiastic about the connections between DPPs and quantum electronic optics. We organized a two-day event[4] in 2019 in Lille, France, opened by Odile Macchi and featuring both theoretical and experimental physicists exposing their view of fermionic coherence to an audience of DPP users across mathematics, computer science, and signal processing. The workshop having been met with cross-disciplinary enthusiasm, we made two decisions. The first

[4] https://dpp-fermions.sciencesconf.org/

was to organize an ambitious, two-week follow-up to the workshop,[5] which took place in 2022 in Lyon, France. The second decision was to unite forces with physicists and write a survey on the links between point processes and optics, to pin down a common ground for discussions. With my postdoc Alexandre Feller, I led a team of dedicated speakers and organizers of the aforementioned workshops in writing –actually, Part I of– a survey on DPPs and fermions (Bardenet et al., 2022), fifty years after Macchi's 1972 thesis. Part II of the survey is work in progress, and will feature chosen topics, like fermionic techniques in combinatorics, experimental measurements of (anti-)bunching particles, or quantum electronic signal processing. This work has already generated many puzzling questions that could easily become research lines at the intersection of physics, probability, and time-frequency signal processing, see e.g. (Bardenet et al., 2022, Section 6). For instance, physical fermions are described in quantum field theory by both a quantum state and a measurement. From the same state, a lot of point processes can be obtained by varying the measurement process: how do all these point processes relate to each other? How is the amount of repulsiveness in a point process related to the corresponding quantum measurement and state? What are the consequences of observables not commuting, and of Heisenberg's uncertainty bound, on the resulting point processes?

The last research line that I want to mention here has me take a step back. My initial motivation to study repulsive point processes is Monte Carlo integration. In turn, my motivation for Monte Carlo integration is Bayesian statistics, where every question asked by a decision-maker, be they biologists, physicists, or government, is turned into the computation of integrals. But after years of teaching Bayesian learning and reading about the foundations of Bayesian statistics, I have finally grown suspicious of part of the Bayesian thought process. For instance, there is no fully convincing case, in my opinion, for conditioning (i.e., using conditional expectations) to represent the acquisition of knowledge with new observations. In particular in machine learning, when the focus is on prediction accuracy at least as much as on internal coherence, and when models are necessarily misspecified (i.e., incomplete descriptions of the data generating process), I believe that blindly conditioning a large model as one acquires more data is hardly justified, and is not what practitioners do, in any case. I would like to spend time thinking of an axiomatic derivation of how to simultaneously update knowledge and make decisions online, which trades off internal coherence and prediction accuracy. No doubt that my upcoming semester at Univ. Pompeu Fabra Barcelona, amid online learning experts, will fuel my thoughts on this topic.

[5] https://indico.in2p3.fr/event/25182/

Appendices

Appendix A

Teaching, supervision, grant management

> *'If only we had laboratories to produce self-replicating scientists, to explore all the worlds. Ah, but we do! They're called university campuses.'*
>
> Terry Pratchett and Stephen Baxter, *The Long War*, 2013

In the main body of the manuscript, I have focused on scientific content. For an HDR manuscript, I believe that it is also relevant to mention some of my parallel activities since my thesis, say teaching, supervision, and grant management. For more complete information, a detailed CV with track record, hierarchical publication list, awards, editorial duties, etc. can be found on my webpage.[1]

A.1 Teaching

CNRS positions come with no teaching duty. Out of personal inclination, I maintain a small teaching activity at the master level. I find the live interaction with students extremely stimulating, and a great complement to research. On top of this, having to teach a topic is probably the best way to make sure you understand it deeply. More prosaically, teaching at the master-level is also the best way to recruit PhD students. Since the end of my PhD, I have been in charge of the following master-level courses.

2019– At ENS Paris-Saclay, France, I am teaching *Bayesian machine learning*, jointly with Julyan Arbel (Inria Grenoble, France). The master programme is called MVA ("Mathematics for vision and learning"), and it has been so far the most demanded machine learning master nationwide.

2019– At Univ. Lille and École Centrale de Lille, France, I am also teaching *Bayesian machine learning* in the joint master of data science.

2016–2019 I taught *Bayesian nonparametrics* at ENSAE ParisTech, to master students majoring in statistics and econometry. This was part of a joint course given over the years with Anna Simoni, Arnak Dalalyan, and Nicolas Chopin (all ENSAE ParisTech, France).

[1] http://rbardenet.github.com/pdf/cvFull_bardenet.pdf

2015–2019 At École centrale de Lille, France, I taught advanced machine learning topics, with applications to bankruptcy prediction, to master-level engineering students specializing in data science (by then called the "DAD curriculum"). This was a joint course with Éric Séverin, a professor in management at Univ. Lille. I organized this course around a machine learning competition, see e.g. the 2019 edition.[2] The topics of the lectures were partly decided together with students as they progressed through the competition.

2013–2014 At Univ. Oxford, UK, I taught *Advanced simulation* on Monte Carlo methods to 4th year statistics students, jointly with Pierre Jacob (by then Univ. Oxford, now ESSEC Paris, France).

A.2 Supervision

Since my PhD, I have been involved in the supervision of

- 5 theses (3 completed, 1 discontinued, 1 running). Two more PhD theses are starting these days, as I am writing this document, and do not appear here.
- 6 postdocs (3 completed, 3 running). One more is also being signed these days, and does not appear here.
- 1 research engineer.

This amount of supervision is maybe unusual in France before an HDR, but it is a consequence of me having access to substantial funding, and the doctoral school in Lille consequently giving me pre-HDR license to supervise PhD students.

A.2.1 Postdocs

2021– I am supervising Michaël Fanuel's postdoc on *Repulsive point processes for inverse problems.*

2021– I am supervising Alexandre Feller's postdoc on *Determinantal point processes and fermions in quantum optics.*

2021– I am co-supervising Juan Manuel Miramont's postdoc with François Auger (Univ. Nantes), Pierre Chainais (Univ. Lille), and Patrick Flandrin (ENS Lyon), on *Zero-based detection and reconstruction of signals.*

2021–2022 I supervised Xiaoyi Mai's postdoc on *Repulsive point processes in supervised learning.* Xiaoyi then got a permanent associate professorship at the department of mathematics (IMT) of Univ. Toulouse, France.

[2] https://www.kaggle.com/c/dad-bankruptcy-prediction-challenge-2019/

2020–2022 I supervised Barbara Pascal's postdoc on *Repulsive point processes in time-frequency analysis.* Barbara then got a CNRS permanent junior researcher position at the Laboratory of digital sciences of Nantes (LS2N), France.

2019–2021 I supervised Arnaud Poinas's postdoc on *Repulsive point processes in experimental design.* Arnaud was co-supervised by Adrien Hardy (by then at the department of mathematics of Univ. Lille) during his first year of postdoc. Arnaud then got a permanent associate professorship at the department of mathematics of Univ. Poitiers, France.

A.2.2 PhD students

2020– I am co-supervising Diala Hawat's PhD with Raphaël Lachièze–Rey (department of applied mathematics, Univ. Paris), on *Variance reduction with point processes.*

2020–disc. I have co-supervised the first half of Yoann Jayer's Ph.D. with Mylène Maida (department of mathematics, Univ. Lille), on *Monte Carlo integration with repulsive stochastic processes.* Yoann quit before completing his PhD, to pursue a career in the finance industry at Qube research and technologies, London.

2017–2020 I co-supervised Ayoub Belhadji's PhD with Pierre Chainais (Centrale Lille & CRIStAL), on *Determinantal point processes for dimension reduction in signal processing.* Ayoub went on to a postdoc with Rémi Gribonval at the department of computer science (LIP) of ENS Lyon, France.

2016–2020 I co-supervised Guillaume Gautier's Ph.D. with Michal Valko (by then Inria Lille, now Google Deepmind Paris), on *Fast sampling of determinantal point processes.* Guillaume went on to a postdoc with Pierre-Olivier Amblard, Simon Barthelmé, and Nicolas Tremblay at GIPSA-Lab, Univ. Grenoble-Alpes, France. Guillaume then decided to become a research software engineer, and joined my group again as such for a year in 2021–2022.

2014–2018 I have had a minor role in the supervision of Ross Johnstone's PhD at Univ. Oxford, UK, whose main supervisors were Gary Mirams and David Gavaghan (Oxford Computer Science), with industrial collaborators at *Roche labs* (Basel, Switzerland) on *Uncertainty characterisation in action potential modelling for cardiac drug safety.* Ross went on to an industry position in data science at Rakuten in Japan.

A.3 Grant management

Whatever one might think about this evolution, securing individual or project-based grants has become an important part of research management, and is likely to remain so in the near future. I have had the chance of progressively getting access to considerable funds, including the grants below.

2020-2025 I am the PI of ERC starting Grant Blackjack (1.5M€), on *Fast Monte Carlo integration with repulsive processes.*

2020–2024 I am the PI of National individual Artificial intelligence Chair Baccarat (880 k€), on *Bayesian learning of expensive models, with applications to cell biology.* These chairs came from a one-time call for large individual grants, as part of the French government's national *AI plan.*

2016-2020 I was the PI of ANR starting grant ("JCJC") BoB (172 k€), on *Bayesian inference on a budget.* JCJC grants are only for junior researchers, and awarded by the French research funding agency (13% acceptance rate that year).

Bibliography

L. D. Abreu. Local maxima of white noise spectrograms and Gaussian entire functions. *arXiv preprint arXiv:2210.06721*, 2022. (Cited on page 55.)

L. D. Abreu, A. Haimi, G. Koliander, and J. L. Romero. Filtering with wavelet zeros and Gaussian analytic functions. *arXiv preprint arXiv:1807.03183*, (v1), 2018. (Cited on page 50.)

G. W. Anderson, A. Guionnet, and O. Zeitouni. *An introduction to random matrices*, volume 118. Cambridge university press, 2010. (Cited on pages 10, 15, 16, 20 and 21.)

G. Ascensi and J. Bruna. Model space results for the Gabor and Wavelet transforms. *IEEE Transactions on Information Theory*, 55(5):2250–2259, 2009. (Cited on page 54.)

F. Bach. On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751, 2017. (Cited on page 20.)

A. Baddeley, E. Rubak, and R. Turner. *Spatial point patterns : methodology and applications with R.* 2015. ISBN 9781482210200. (Cited on page 47.)

R. Bardenet and S. Ghosh. Learning from DPPs via sampling: Beyond HKPV and symmetry. *arXiv preprint arXiv:2007.04287*, 2020. (Cited on page 59.)

R. Bardenet and A. Hardy. Time-frequency transforms of white noises and Gaussian analytic functions. *Applied and Computational Harmonic Analysis*, 2019. (Cited on pages 44, 45, 47, 48, 49, 50, 52, 53 and 54.)

R. Bardenet and A. Hardy. Monte Carlo with Determinantal Point Processes. *Annals of Applied Probability*, 2020. doi: 10.1214/19-AAP1504. (Cited on pages 2, 3, 20, 22, 23, 24, 26 and 57.)

R. Bardenet and M. Titsias. Inference for determinantal point processes without spectral knowledge. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3393–3401, 2015. (Cited on page 58.)

R. Bardenet, A. Doucet, and C. Holmes. Towards scaling up MCMC: an adaptive subsampling approach. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014. (Cited on page 1.)

R. Bardenet, A. Doucet, and C. Holmes. On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research (JMLR)*, 2017a. (Cited on page 1.)

R. Bardenet, F. Lavancier, X. Mary, and A. Vasseur. On a few statistical applications of determinantal point processes. *ESAIM: Proceedings and Surveys*, 60:180–202, 2017b. (Cited on page 16.)

R. Bardenet, J. Flamant, and P. Chainais. On the zeros of the spectrogram of white noise. *Applied and Computational Harmonic Analysis*, 2018. (Cited on pages 3, 44, 45, 47 and 54.)

R. Bardenet, S. Ghosh, and M. Lin. Determinantal point processes based on orthogonal polynomials for sampling minibatches in SGD. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. (Cited on pages 57 and 59.)

R. Bardenet, A. Feller, J. Bouttier, P. Degiovanni, A. Hardy, A. Rançon, B. Roussel, G. Schehr, and C. I. Westbrook. From point processes to quantum optics and back. *arXiv preprint arXiv:2210.05522*, 2022. (Cited on pages 7, 16 and 60.)

R. F. Bass and K. Gröchenig. Relevant sampling of band-limited functions. *Illinois Journal of Mathematics*, 57(1):43–58, 2013. (Cited on page 56.)

K. A. Beattie, A. P. Hill, R. Bardenet, Y. Cui, J. I. Vandenberg, D. J. Gavaghan, T. P. de Boer, and G. R. Mirams. Sinusoidal voltage protocols for rapid characterization of ion channel kinetics. *Journal of Physiology*, 2018. (Cited on page 1.)

A. Belhadji. *Subspace sampling using determinantal point processes.* PhD thesis, Centrale Lille Institut, 2020. (Cited on pages 16, 29, 30 and 36.)

A. Belhadji, R. Bardenet, and P. Chainais. Kernel quadrature with DPPs. In *Advances in Neural Information Processing Systems 32*, pages 12907–12917. 2019. (Cited on pages 2, 20, 26, 28 and 29.)

A. Belhadji, R. Bardenet, and P. Chainais. Kernel interpolation with continuous volume sampling. In *International Conference on Machine Learning (ICML)*, 2020a. (Cited on pages 2, 20, 34, 35, 37, 38, 39 and 40.)

A. Belhadji, R. Bardenet, and P. Chainais. A determinantal point process for column subset selection. *Journal of Machine Learning Research (JMLR)*, 2020b. (Cited on pages 3, 26, 29 and 30.)

C. Bénard and O. Macchi. Detection and “emission" processes of quantum particles in a “chaotic" state. *Journal of mathematical physics*, 1973. (Cited on pages 7 and 15.)

A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics.* Springer Science & Business Media, 2011. (Cited on page 27.)

C. A. N. Biscio and F. Lavancier. Quantifying repulsiveness of determinantal point processes. *Bernoulli*, 22(4):2001–2028, nov 2016. ISSN 13507265. doi: 10.3150/15-BEJ718. (Cited on page 7.)

B. Błaszczyszyn, D. Yogeshwaran, and J. E. Yukich. Limit theory for geometric statistics of point processes having fast decay of correlations. *The Annals of Probability*, 47(2): 835–895, 2019. (Cited on page 26.)

C. Boutsidis, M. W. Mahoney, and P. Drineas. An improved approximation algorithm for the column subset selection problem. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '09, pages 968–977, Philadelphia, PA, USA, 2009. Society for Industrial and Applied Mathematics. (Cited on pages 35 and 41.)

C. Boutsidis, P. Drineas, and M. Magdon-Ismail. Near optimal column-based matrix reconstruction. In *Proceedings of the 2011 IEEE 52Nd Annual Symposium on Foundations of Computer Science*, FOCS '11, pages 305–314, Washington, DC, USA, 2011. IEEE Computer Society. ISBN 978-0-7695-4571-4. doi: 10.1109/FOCS.2011.21. (Cited on pages 34, 35 and 41.)

V. Chernozhukov and H. Hong. An MCMC approach to classical estimation. *Journal of Econometrics*, 2003. (Cited on page 59.)

D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes. Vol. I.* Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2003. ISBN 0-387-95541-0. Elementary theory and methods. (Cited on page 6.)

I. Daubechies and T. Paul. Time-frequency localization operators – a geometric phase space approach: Ii. the use of dilations. *Inverse problems*, 4:661–680, 1988. (Cited on pages 44 and 52.)

M. Dereziński, D. Calandriello, and M. Valko. Exact sampling of determinantal point processes with sublinear time preprocessing. In H. W. Garnett, H. Larochelle, A. Beygelzimer, F. D'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, pages 11542–11554, Vancouver, Canada, 2019. Curran Associates, Inc. (Cited on page 41.)

A. Deshpande and L. Rademacher. Efficient volume sampling for row/column subset selection. In *Proceedings of the symposium on foundations of computer science*, 2010. (Cited on pages 2, 15 and 37.)

A. Deshpande and S. Vempala. 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, pages 292–303. Springer-Verlag, 2006. (Cited on pages 15, 35 and 36.)

A. Deshpande, L. Rademacher, S. Vempala, and G. Wang. Matrix approximation and projective clustering via volume sampling. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm*, SODA '06. Society for Industrial and Applied Mathematics, 2006. (Cited on pages 36 and 37.)

J. Dick and F. Pillichshammer. *Digital nets and sequences: discrepancy theory and quasi–Monte Carlo integration.* Cambridge University Press, 2010. (Cited on pages 2, 24 and 31.)

P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. *Mach. Learn.*, 56(1-3):9–33, June 2004. ISSN 0885-6125. doi: 10.1023/B:MACH.0000033113.59016.96. (Cited on page 35.)

P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Relative-error cur matrix decompositions, 2007. (Cited on pages 35 and 36.)

P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13 (Dec):3475–3506, 2012. (Cited on page 41.)

S. M. Ermakov and V. G. Zolotukhin. Polynomial Approximations and the Monte-Carlo Method. *Theory of Probability and Its Applications*, 5(4):428–431, jan 1960. ISSN 0040-585X. doi: 10.1137/1105046. (Cited on pages 20, 25 and 26.)

M. Fanuel and R. Bardenet. Nonparametric inference of continuous DPPs. In *Advances in Neural Information Processing Processing Systems (NeurIPS)*, 2021. (Cited on page 58.)

M. Fanuel and R. Bardenet. Sparsification of the regularized magnetic Laplacian with multi-type spanning forests. *arXiv preprint arXiv:2208.14797*, 2022. (Cited on page 58.)

P. Flandrin. *Time-frequency/time-scale analysis.* Academic press, 1998. (Cited on pages 43 and 44.)

P. Flandrin. Time–frequency filtering based on spectrogram zeros. *IEEE Signal Processing Letters*, 22(11):2137–2141, 2015. (Cited on pages 43, 47, 55 and 56.)

M. Gartrell, V.-E. Brunel, E. Dohmatob, and S. Krichene. Learning Nonsymmetric Determinantal Point Processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. (Cited on page 15.)

C. F. Gauss. *Methodus nova integralium valores per approximationem inveniendi.* Heinrich Dietrich, Göttingen, 1815. (Cited on page 23.)

G. Gautier. *On sampling determinantal point processes.* PhD thesis, Centrale Lille Institut, 2020. (Cited on pages 16 and 58.)

G. Gautier, R. Bardenet, and M. Valko. Zonotope hit-and-run for efficient sampling from projection DPPs. In *International Conference on Machine Learning (ICML)*, 2017. (Cited on pages 14 and 58.)

G. Gautier, R. Bardenet, G. Polito, and M. Valko. DPPy: Sampling determinantal point processes with Python. *Journal of Machine Learning Research; Open Source Software (JMLR MLOSS)*, 2019a. (Cited on page 58.)

G. Gautier, R. Bardenet, and M. Valko. On two ways to use determinantal point processes for Monte Carlo integration. In *Advances in Neural Information Processing Systems 32*, pages 7768—-7777. 2019b. (Cited on pages 15 and 20.)

G. Gautier, R. Bardenet, and M. Valko. On two ways to use determinantal point processes for Monte Carlo integration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019c. (Cited on pages 20, 25 and 26.)

G. Gautier, R. Bardenet, and M. Valko. Fast sampling of β-ensembles. *Statistics and Computing*, 2021. (Cited on page 58.)

S. Ghosh. Determinantal processes and completeness of random exponentials: the critical case. *Probability Theory and Related Fields*, 163(3):643–665, 2015. (Cited on page 56.)

S. Ghosh, M. Lin, and D. Sun. Signal analysis via the stochastic geometry of spectrogram level sets. *IEEE Transactions on Signal Processing*, 70:1104–1117, 2022. (Cited on page 55.)

R. J. Glauber. Quantum theory of coherence. In *Quantum optics: Proceedings of the 10th session of the Scottish universities summer school in Physics*, page 53, 1970. (Cited on page 15.)

I. Gohberg, S. Goldberg, and M. A. Kaashoek. *Classes of linear operators, Volume I.* Springer, 1990. (Cited on page 12.)

G. Golub. Numerical methods for solving linear least squares problems. *Numerische Mathematik*, 7(3):206–216, 1965. ISSN 0945-3245. doi: 10.1007/BF01436075. (Cited on page 35.)

K. Gröchenig. *Foundations of time-frequency analysis.* Birkhäuser, 2001. (Cited on pages 44, 47, 53 and 56.)

L. Gross. *Abstract Wiener Spaces.* Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (1965/66), Vol. II. Univ. California Press., 1967. (Cited on pages 49 and 50.)

M. Gu and S. C. Eisenstat. Efficient algorithms for computing a strong rank-revealing QR factorization. *SIAM J. Sci. Comput.*, 17(4):848–869, July 1996. ISSN 1064-8275. doi: 10.1137/0917055. (Cited on page 35.)

A. Haimi, G. Koliander, and J. L. Romero. Zeros of Gaussian Weyl-Heisenberg functions and hyperuniformity of charge. *Journal of Statistical Physics*, 187(3):1–41, 2020. (Cited on page 54.)

A. Hardy. Attractive repulsion. HDR manuscript, Université de Lille, 2020. (Cited on page 16.)

D. Hawat, G. Gautier, R. Bardenet, and R. Lachièze-Rey. On estimating the structure factor of a point process, with applications to hyperuniformity. *arXiv preprint arXiv:2203.08749*, 2022. (Cited on page 58.)

N. Holighaus, G. Koliander, Z. Průša, and L. D. Abreu. Characterization of analytic wavelet transforms and a new phaseless reconstruction algorithm. *IEEE Transactions on Signal processing*, 67(15):3894–3908, 2019. (Cited on page 54.)

J. B. Hough, M. Krishnapur, Y. Peres, and B. Virág. Determinantal Processes and Independence. In *Probability Surveys*, 2006. doi: 10.1214/154957806000000078. (Cited on pages 9, 10, 12, 15 and 16.)

J. B. Hough, M. Krishnapur, Y. Peres, and B. Virág. *Zeros of Gaussian analytic functions and determinantal point processes*. American Mathematical Society, 2009. ISBN 9780821843734. (Cited on pages 17, 47 and 48.)

R. G. Impagliazzo. Hardness as randomness: a survey of universal derandomization. Lectures notes from the International Congress of Mathematicians, 2002. (Cited on page 2.)

H. Jaquard, M. Fanuel, P.-O. Amblard, R. Bardenet, S. Barthelmé, and N. Tremblay. Smoothing complex-valued signals on graphs with Monte-Carlo. *arXiv preprint arXiv:2208.14797*, 2022. (Cited on page 58.)

K. Johansson. On random matrices from the compact classical groups. *Annals of mathematics*, pages 519–545, 1997. (Cited on pages 20 and 21.)

K. Johansson. Random matrices and determinantal processes. *Les Houches Summer School Proceedings*, 2006. ISSN 09248099. doi: 10.1016/S0924-8099(06)80038-7. (Cited on page 16.)

R. H. Johnstone, R. Bardenet, D. J. Gavaghan, and G. R. Mirams. Hierarchical Bayesian inference for ion channel screening dose-response data. *Wellcome Open Research*, 2016a. (Cited on page 1.)

R. H. Johnstone, E. T. Y. Chang, R. Bardenet, T. P. De Boer, D. J. Gavaghan, P. Pathmanathan, R. H. Clayton, and G. R. Mirams. Uncertainty and variability in models of the cardiac action potential: Can we build trustworthy models? *Journal of molecular and cellular cardiology*, 96:49–62, 2016b. (Cited on page 1.)

A. Kassel and T. Lévy. Determinantal probability measures on Grassmannians. *arXiv preprint arXiv:1910.06312*, 2019. (Cited on page 41.)

W. König. Orthogonal polynomial ensembles in probability theory. *Probability Surveys*, 2004. ISSN 1549-5787. doi: 10.1214/154957805100000177. (Cited on page 22.)

A. Kulesza and B. Taskar. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012. (Cited on pages 13, 14, 15 and 16.)

G. Lambert. *Fluctuations of smooth linear statistics of determinantal point processes*. PhD thesis, KTH Stockholm, 2016. (Cited on page 10.)

F. Lavancier, J. Møller, and E. Rubak. Determinantal point process models and statistical inference. *Journal of the Royal Statistical Society, Series B*, 2014. (Cited on pages 12 and 13.)

F. Lavancier, J. Møller, and E. Rubak. Determinantal point process models and statistical inference: Extended version. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 2015. doi: 10.1111/rssb.12096. (Cited on page 16.)

R. Lyons. Determinantal probability measures. *Publications mathématiques de l'IHÉS*, 2002. ISSN 0073-8301. doi: 10.1007/s10240-003-0016-0. (Cited on page 16.)

R. Lyons and Y. Peres. *Probabilities on trees and networks.* Cambridge University Press, 2016. (Cited on page 58.)

O. Macchi. The coincidence approach to stochastic point processes. 7:83–122, 03 1975. (Cited on pages 7, 12, 16, 20 and 25.)

O. Macchi. *Point processes and coincidences – Contributions to the theory, with applications to statistical optics and optical communication, augmented with a scholion by Suren Poghosyan and Hans Zessin.* Walter Warmuth Verlag, 2017. (Cited on pages 7 and 15.)

X. Mai and R. Bardenet. Un processus ponctuel déterminantal pour la sélection de variables supervisée. In *Actes de la conférence du GRETSI*, 2022. (Cited on page 41.)

U. Marteau-Ferey, F. R. Bach, and A. Rudi. Non-parametric models for non-negative functions. In *Advances in Neural Information Processing Systems 33*, 2020. (Cited on page 58.)

J. M. Miramont, R. Bardenet, P. Chainais, and F. Auger. A public benchmark for denoising and detection methods. In *Actes de la conférence du GRETSI*, 2022. (Cited on page 55.)

J. Møller and E. O'Reilly. Couplings for determinantal point processes and their reduced Palm distributions with a view to quantifying repulsiveness. *Journal of Applied Probability*, 58(2):469–483, 2021. (Cited on page 7.)

L. Mor-Yosef and H. Avron. Sketching for principal component regression. *SIAM Journal on Matrix Analysis and Applications*, 40(2):454–485, 2019. (Cited on page 40.)

E. Moulines and F. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 24, pages 451–459, 2011. (Cited on page 57.)

D. Papailiopoulos, A. Kyrillidis, and C. Boutsidis. Provable deterministic leverage score sampling. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 997–1006, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2956-9. doi: 10.1145/2623330.2623698. (Cited on page 38.)

B. Pascal and R. Bardenet. A covariant, discrete time-frequency representation tailored for zero-based signal detection. *IEEE Transactions on Signal Processing*, 2022. (Cited on pages 44, 45, 52, 53 and 54.)

A. Pinkus. *N-widths in Approximation Theory*, volume 7. Springer Science & Business Media, 2012. (Cited on page 30.)

A. Poinas and R. Bardenet. On proportional volume sampling for experimental design in general spaces. *To appear in Statistics and Computing*, 2022. (Cited on page 58.)

A. Rezaei and S. O. Gharan. A polynomial time MCMC method for sampling from continuous determinantal point processes. In *International Conference on Machine Learning*, pages 5438–5447, 2019. (Cited on page 31.)

B. J. Ridder, D. J. Leishman, M. Bridgland-Taylor, M. Samieegohar, X. Han, W. W. Wu, A. Randolph, P. Tran, J. Sheng, T. Danker, A. Lindqvist, D. Konrad, S. Hebeisen, L. Polonchuk, E. Gissinger, M. Renganathan, B. Koci, H. Wei, J. Fan, P. Levesque, J. Kwagh, J. Imredy, J. Zhai, M. Rogers, E. Humphries, R. Kirby, S. Stoelzle-Feix, N. Brinkwirth, M. Giustin, N. Becker, S. Friis, M. Rapedius, T. A. Goetze, T. Strassmaier, G. Okeyo, J. Kramer, Y. Kuryshev, C. Wu, H. Himmel, G. R. Mirams, D. G. Strauss, R. Bardenet, and Z. Li. A systematic strategy for estimating herg block potency and its implications in a new cardiac safety paradigm block potency and its implications in a new cardiac safety paradigm. *Toxicology and Applied Pharmacology*, 2020. (Cited on page 1.)

C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer, 2004. (Cited on page 24.)

Z. Rudnick and P. Sarnak. Zeros of principal L-functions and random matrix theory. *Duke Mathematical Journal*, 81(2):269–322, 1996. ISSN 0012-7094. doi: 10.1215/S0012-7094-96-08115-6. (Cited on page 10.)

B. Simon. *Trace Ideals and Their Applications*. American Mathematical Society, 2005. (Cited on page 27.)

M. Slawski. On principal components regression, random projections, and column subsampling. *Electronic Journal of Statistics*, 12(2):3673–3712, 2018. (Cited on page 39.)

A. Soshnikov. Determinantal random point fields. *Russian Mathematical Surveys*, 55: 923–975, Oct. 2000. doi: 10.1070/RM2000v055n05ABEH000321. (Cited on pages 12, 16 and 26.)

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387772413. (Cited on page 27.)

I. Steinwart and C. Scovel. Mercer's theorem on general domains: on the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35(3):363–417, 2012. (Cited on page 27.)