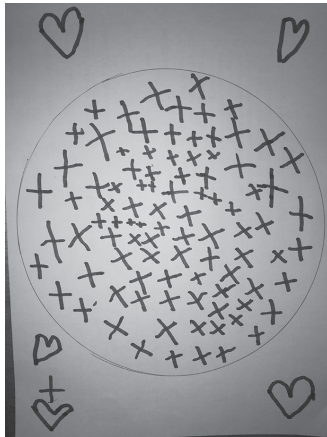


Determinantal Point Processes

Adrien Hardy (Université de Lille, France) and Mylène Maïda (Université de Lille, France)

1 Introduction

If you ask a five-year-old child to *randomly* draw points within a disc, they will probably produce a picture like this one:



which looks more like the picture on the left than the one of the right:

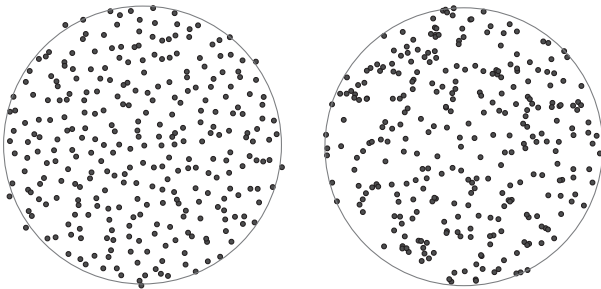


Figure 1. On the left, spectrum of a Ginibre matrix; on the right, sample of independent uniform points on the disc.

However, the figure on the right is a realisation of a point process¹ which naturally corresponds to the notion of randomness in mathematics: the well-known homogenous Poisson process, made of points that are chosen uniformly on the disc and *independently* of one to another. In particular, with the Poisson process, getting very close points is possible but the human brain tends to avoid such configurations for mysterious reasons.

The figure on the left-hand side shows the eigenvalues in the complex plane of a matrix with independent complex Gaussian entries², known as a *Ginibre matrix*. It seems that the intuition of randomness for a five-year-old child³ fits better with these kind of structured configurations rather than

1 In the whole article, the notion of point process refers to a spatial process, namely random configurations (i.e. locally finite subsets) of points in space; indeed, the word *process* is commonly used in this context, although there is no relation to time.
 2 A complex Gaussian is a complex random variable of which the real part and the imaginary part are two independent real Gaussian variables.

realisations of a Poisson process. In this article, we want to discuss that underlying structure, which is a particular case of a class of point process with some intriguing properties: the *determinantal point processes*, hereafter abbreviated DPPs.

The following is neither going to be a formal introduction nor an exhaustive survey of this wide topic, but rather a personal choice of some elegant DPPs collected in various mathematical domains⁴. For a more extensive and classical presentation, see, e.g., [5, 6, 10, 11].

As for lovers of practical applications, let us imagine that we type the query “jaguar” into a search engine. Among the top results, we don’t want to find 10 articles on sports cars, but also one on the animal with spotted fur, one on the film by Francis Veber and maybe one on an American football team. For a wide range of subjects, we need to introduce repulsion between similar items within the algorithms: if an item has been selected, the closely related items are less likely to be displayed. We are going to see, towards the end of this paper, that the DPP has already aroused curiosity among *machine learners* by providing implementable solutions to this kind of problem, and to many others.

My first DPP: The carries process.

Consider the following process, which might remind some of us of their primary school days: you have a column of digits which you add up one by one from top to bottom. At each step, you note the unit of the result on the right, note a dot if there is a number to carry, go on to the next line and so on.

$$\begin{array}{r}
 3 \quad 3 \\
 + 6 = 9 \\
 + 5 = 4 \bullet \\
 + 4 = 8 \\
 + 4 = 2 \bullet \\
 + 3 = 5 \\
 + 3 = 8 \\
 + 7 = 5 \bullet
 \end{array}$$

If the first column contains random, independent and identically distributed (iid) digits following the uniform distribution on $\llbracket 0, 9 \rrbracket := \{0, \dots, 9\}$, how will the distribution of the point process in relation to the carries be?

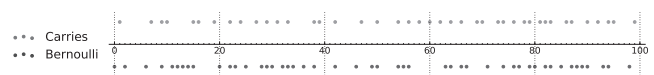


Figure 2. Above, the point process of carries; below, iid Bernoulli variables of the same parameter $p = 9/20$.

3 Study carried out on a sample of three participants, without quota sampling.
 4 Let us mention some noteworthy DPPs which are not quoted in our article: for a given graph, the process of the edges of a uniform spanning tree is a DPP; non-intersecting random walks on a bipartite graph form, at fixed time, a DPP; the zeros of the hyperbolic Gaussian analytic function form a DPP; etc.

As a column of iid uniform digits on $\llbracket 0, 9 \rrbracket$ leads, after successive additions, to a column of iid units that are also uniform on $\llbracket 0, 9 \rrbracket$, it is equivalent to consider the following descent point process, treated in a more general framework by Borodin, Diaconis and Fulman [4]: consider a column of $n+1$ digits S_0, S_1, \dots, S_n that are iid uniform on $\llbracket 0, 9 \rrbracket$ and, at each line, we note a dot on the right if the digit in this line is strictly inferior to the previous one. More precisely, if we put

$$X_i := \mathbf{1}_{\{S_i < S_{i-1}\}} \in \{0, 1\},$$

the descent point process is given by the random configurations $D_n := \{i \in \llbracket 1, n \rrbracket : X_i = 1\}$.

For instance, in the following example, we obtain $D_7 = \{2, 4, 7\}$:

i	S_i	X_i
0	3	
1	9	0
2	4	• 1
3	8	0
4	2	• 1
5	5	0
6	8	0
7	5	• 1

The simple computation

$$\begin{aligned} \mathbb{P}(\{i\} \subset D_n) &= \mathbb{P}(X_i = 1) = \mathbb{P}(S_i < S_{i-1}) \\ &= \frac{1}{10^2} \binom{10}{2} = \frac{9}{20} < \frac{1}{2} \end{aligned}$$

shows that, as one may expect, at every line the probability to have a descent/carry is a little bit less than $1/2$. Thus, the sequence X_1, \dots, X_n is a sequence of random Bernoulli variables⁵ of parameter $9/20$ but they are not independent. Intuitively, if you have a carry in a given line, the digit in that line tends to be small and there are less chances to have one in the next line. On Figure 2, we compare a realisation of the process D_{100} with a random configuration $B_{100} = \{i \in \llbracket 1, 100 \rrbracket : Y_i = 1\}$ where Y_1, \dots, Y_{100} are iid Bernoulli variables of the same parameter $9/20$.

There is clearly a property of *negative association* or *repulsion* between the adjacent descents, a discrete and unidimensional analogue of what we observed in Figure 1 on the left. The following computation confirms that two adjacent descents are negatively correlated:

$$\begin{aligned} \mathbb{P}(\{i, i+1\} \subset D_n) &= \mathbb{P}(X_i = 1 \text{ and } X_{i+1} = 1) \\ &= \mathbb{P}(S_{i+1} < S_i < S_{i-1}) \\ &= \frac{1}{10^3} \binom{10}{3} \\ &= \frac{3}{25} < \left(\frac{9}{20}\right)^2 \\ &= \mathbb{P}(\{i\} \subset D_n) \mathbb{P}(\{i+1\} \subset D_n). \end{aligned}$$

⁵ That is to say, $\mathbb{P}(X_i = 1) = 1 - \mathbb{P}(X_i = 0) = 9/20$.

On the contrary, if $|i - j| > 1$, the independence implies

$$\begin{aligned} \mathbb{P}(\{i, j\} \subset D_n) &= \mathbb{P}(X_i = 1 \text{ and } X_j = 1) \\ &= \mathbb{P}(S_i < S_{i-1}) \mathbb{P}(S_j < S_{j-1}) \\ &= \left(\frac{9}{20}\right)^2; \end{aligned}$$

we say that the process is of range 1.

Now, in order to completely describe the law of the descent point process, it is enough to determine $\mathbb{P}(A \subset D_n)$ for any subset A of $\llbracket 1, n \rrbracket$. If A has cardinal k , it is called *k-point correlation function*, which we denote by $\rho_k(A)$. We have already determined the one-point correlation function $\rho_1(\{i\}) = 9/20$ for any i , as well as the two-point correlation function: for all $i \neq j$,

$$\rho_2(\{i, j\}) = \begin{cases} \left(\frac{9}{20}\right)^2 & \text{if } |i - j| > 1, \\ \frac{3}{25} & \text{if } |i - j| = 1. \end{cases}$$

When $k \geq 3$, if A is a sequence of k consecutive numbers of $\llbracket 1, n \rrbracket$, one has $\rho_k(A) = \frac{1}{10^{k+1}} \binom{10}{k+1}$. Otherwise, we can write $A = A_1 \cup A_2$ with A_1 and A_2 at distance at least 2 and with respective cardinals k_1 and k_2 , so that $\rho_k(A) = \rho_{k_1}(A_1) \rho_{k_2}(A_2)$, because the process is of range 1. Going a little further, we can verify that this collection of correlation functions is encoded by a function with two variables $K : \llbracket 1, n \rrbracket^2 \rightarrow \mathbb{R}$ in the sense that, for any $k \in \llbracket 1, n \rrbracket$ and $s_1, \dots, s_k \in \llbracket 1, n \rrbracket$ pairwise distinct, one has:

$$\rho_k(\{s_1, \dots, s_k\}) = \det[K(s_i, s_j)]_{1 \leq i, j \leq k}.$$

We say that the point process D_n is *determinantal with kernel* K . Borodin, Diaconis and Fulman [4] obtain a rather explicit formula for the kernel:

$$K(i, j) := \kappa(j - i) \text{ where } \sum_{m \in \mathbb{Z}} \kappa(m) z^m = \frac{1}{1 - (1 - z)^{10}}.$$

More generally, they show that any point process of range 1 on a segment of \mathbb{Z} is determinantal. They also study the descents of a random permutation: a permutation σ of $\llbracket 1, n \rrbracket$ has a descent at i if $\sigma(i - 1) > \sigma(i)$. If σ is chosen uniformly in the symmetric group, then the descent process is determinantal with kernel

$$K_{\text{per}}(i, j) := \kappa_{\text{per}}(j - i) \text{ where } \sum_{m \in \mathbb{Z}} \kappa_{\text{per}}(m) z^m = \frac{1}{1 - e^z}.$$

Below, we are going to come across some other interesting DPP related to random permutations.

DPP and random matrices

While the mathematical study of DPPs began in the 1970s with Odile Macchi's thesis, inspired by the formalism of fermions in quantum mechanics, their appearance in random matrix theory has widely popularised them. In the following, we present some results that show that the set of eigenvalues of certain matrix models form a DPP. For all the results that are mentioned in this paragraph, we recommend for instance the monograph [1].

A first example is given by the eigenvalues of unitary matrices sampled "uniformly". More precisely, for $n \geq 1$, we equip the unitary group $U_n(\mathbb{C}) := \{U \in M_n(\mathbb{C}) : UU^* = I_n\}$

with its unique probability measure ν_n invariant by left and right multiplication, namely its normalised Haar measure. We hereby obtain a random variable random V with values in $U_n(\mathbb{C})$ by specifying that $\mathbb{P}(V \in A) = \nu_n(A)$ for any Borel set $A \subset U_n(\mathbb{C})$. To obtain the joint probability distribution of the n (random) eigenvalues of V , we perform a change of variables mapping a unitary matrix to the set of its eigenvalues and eigenvectors,⁶ and then integrate on its eigenvectors. If we denote by $e^{i\theta_1}, \dots, e^{i\theta_n}$ the eigenvalues of the random matrix V , a classical Jacobian computation in Lie group theory (due to Weyl) shows that the phases $(\theta_1, \dots, \theta_n) \in [-\pi, \pi]^n$ follow the probability law:

$$d\mathbb{P}(\theta_1, \dots, \theta_n) = \frac{1}{n!} \prod_{j < k} |e^{i\theta_j} - e^{i\theta_k}|^2 \prod_{j=1}^n \frac{d\theta_j}{2\pi}. \quad (1.1)$$

By noting that the interaction term between the eigenvalues is the square of a Vandermonde determinant, we obtain

$$d\mathbb{P}(\theta_1, \dots, \theta_n) = \frac{1}{n!} \det [K_{U_n(\mathbb{C})}(s_i, s_j)]_{1 \leq i, j \leq n} \prod_{j=1}^n \frac{d\theta_j}{2\pi} \quad (1.2)$$

where we introduced the kernel

$$K_{U_n(\mathbb{C})}(x, y) := \sum_{k=0}^{n-1} \varphi_k(x) \overline{\varphi_k(y)}, \quad \varphi_k(x) := \frac{e^{ikx}}{\sqrt{2\pi}}.$$

Observe that $K_{U_n(\mathbb{C})}(x, y)$ is the kernel of the projection operator onto the subspace of $L^2(-\pi, \pi)$ of the trigonometric polynomials of degree at most $n - 1$. Taking into account the continuous character of the eigenangles, we define the “infinitesimal” version of the k -point correlation function we introduced above for the carries process: for every $k \geq 1$ and $s_1, \dots, s_k \in [-\pi, \pi]$ pairwise distinct,

$$\rho_k(s_1, \dots, s_k) := \lim_{\varepsilon \rightarrow 0} \frac{1}{\varepsilon^k} \mathbb{P}(\forall j \in \llbracket 1, k \rrbracket, \{\theta_1, \dots, \theta_n\} \cap [s_j, s_j + \varepsilon] \neq \emptyset).$$

In other words, for any reasonable test function $\varphi : [-\pi, \pi]^k \rightarrow \mathbb{C}$, one has

$$\int_{[-\pi, \pi]^k} \varphi(s) \rho_k(s) ds = \int \sum_{i_1 \neq \dots \neq i_k} \varphi(\theta_{i_1}, \dots, \theta_{i_k}) d\mathbb{P}(\theta_1, \dots, \theta_n). \quad (1.3)$$

Thanks to the invariance by permutation of $d\mathbb{P}(\theta_1, \dots, \theta_n)$, we see that

$$\begin{aligned} \rho_k(s_1, \dots, s_k) &= \frac{n!}{(n-k)!} \int_{[-\pi, \pi]^{n-k}} d\mathbb{P}(s_1, \dots, s_k, s_{k+1}, \dots, s_n) \\ &= \frac{1}{(n-k)!} \int_{[-\pi, \pi]^{n-k}} \det [K_{U_n(\mathbb{C})}(s_i, s_j)]_{1 \leq i, j \leq n} \prod_{j=k+1}^n \frac{ds_j}{2\pi} \\ &= \det [K_{U_n(\mathbb{C})}(s_i, s_j)]_{1 \leq i, j \leq k}, \end{aligned} \quad (1.4)$$

⁶ This function is not a well-defined diffeomorphism. We fix it by reducing the source space to the matrices with simple eigenvalues (its complementary being of measure zero for ν_n), and the target space by ordering the eigenvalues and by taking a quotient to get rid of the freedom for the eigenvectors.

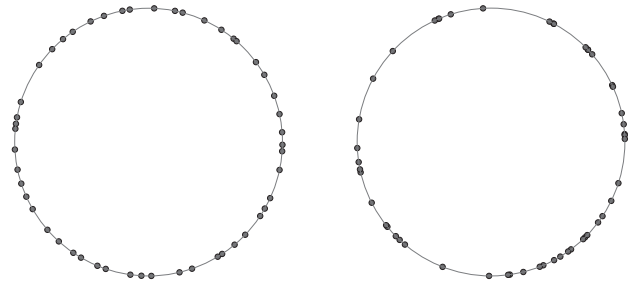


Figure 3. On the left, spectrum of a realisation of V ; on the right, independent samples of uniform points on the unit circle.

where the last identity is obtained by developing the determinant under the integral and by using that $K_{U_n(\mathbb{C})}(x, y)$ is the kernel of a projection operator:

$$\int_{-\pi}^{\pi} K_{U_n(\mathbb{C})}(s_i, s) K_{U_n(\mathbb{C})}(s, s_j) \frac{ds}{2\pi} = K_{U_n(\mathbb{C})}(s_i, s_j).$$

Thus, the point process of the eigenangles $\{\theta_1, \dots, \theta_n\}$ is determinantal. Let us visually compare a realisation of this process to iid uniform variables on the unit circle: as in the case of the descents, we observe a more regular distribution than in the independent case. The eigenangles are well-distributed and there is little variability; while in the uniform case, we have clusters of points which can be noticeably different from one realisation to another.

This repulsion is also visible in the interaction term of (1.1). It induces behaviours that are surprising for an unprepared probabilist. For example, the variance⁷ of a sum of n eigenangles is much smaller than the one of n iid angles. Indeed, if η_1, \dots, η_n are iid uniformly distributed on $[-\pi, \pi]$, one has:

$$\text{Var} \left[\sum_{j=1}^n e^{i\eta_j} \right] = \sum_{j=1}^n \text{Var} [e^{i\eta_j}] = n.$$

On the other hand, for the eigenangles we compute, using (1.4) with $k = 1, 2$:

$$\begin{aligned} \text{Var} \left[\sum_{j=1}^n e^{i\theta_j} \right] &= \int_{-\pi}^{\pi} |e^{ix}|^2 K_{U_n(\mathbb{C})}(x, x) dx \\ &\quad - \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} e^{i(x-y)} |K_{U_n(\mathbb{C})}(x, y)|^2 dx dy \\ &= n - \sum_{k, \ell=0}^{n-1} \frac{1}{(2\pi)^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} e^{i(x-y)(1+k-\ell)} dx dy \\ &= 1. \end{aligned}$$

For the sum of n random complex numbers of module 1 to have a unit variance, independently of n , many cancellations need to take place, and this forces the realisations of the eigenvalues not to differ too much from an equally-spaced configuration on the unit circle. This confirms the visual impression of Figure 3.

A natural question is to study asymptotically the spacing between these eigenangles when $n \rightarrow \infty$, after an appropriate scaling. Indeed, as there are n eigenangles well spread on $[-\pi, \pi]$, the point process $\{\frac{n}{2\pi}\theta_1, \dots, \frac{n}{2\pi}\theta_n\}$ is a subset of

⁷ The variance of a random complex variable Z is defined by $\text{Var}(Z) := \mathbb{E}|Z|^2 - |\mathbb{E}(Z)|^2$.

$[-\frac{n}{2}, \frac{n}{2}]$, where the standard distance between two consecutive points is of order one. A change of variables yields a DPP with kernel:⁸

$$\begin{aligned} \widetilde{K}_{U_n(\mathbb{C})}(x, y) &:= e^{-i\frac{n-1}{2}(x-y)} \frac{2\pi}{n} K_{U_n(\mathbb{C})}\left(\frac{2\pi x}{n}, \frac{2\pi y}{n}\right) \\ &= \frac{\sin \pi(x-y)}{n \sin\left(\frac{\pi}{n}(x-y)\right)}. \end{aligned}$$

One obtains the uniform local convergence

$$\widetilde{K}_{U_n(\mathbb{C})}(x, y) \xrightarrow{n \rightarrow \infty} K_{\sin}(x, y) := \frac{\sin \pi(x-y)}{\pi(x-y)}$$

where, by convention, $K_{\sin}(x, x) := 1$. In probabilistic terms, when $n \rightarrow \infty$, the DPP of the normalised eigenangles $\frac{n}{2\pi}\theta_1, \dots, \frac{n}{2\pi}\theta_n$ of the random unitary matrices distributed according to the Haar measure on $U_n(\mathbb{C})$ converge, in the sense of the local uniform convergence of correlation functions, towards a limiting point process which is the DPP associated with the *sine kernel* $K_{\sin}(x, y)$: for all $k \geq 1$ and any reasonable function $\varphi : \mathbb{R}^k \rightarrow \mathbb{R}$,

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{E} \left[\sum_{i_1 \neq \dots \neq i_k} \varphi\left(\frac{n}{2\pi}\theta_{i_1}, \dots, \frac{n}{2\pi}\theta_{i_k}\right) \right] \\ = \int_{\mathbb{R}^k} \varphi(s) \det [K_{\sin}(s_i, s_j)]_{i,j=1}^k ds. \quad (1.5) \end{aligned}$$

Furthermore, K_{\sin} , seen as an operator acting on $L^2(\mathbb{R})$, is the projection on the space of functions whose Fourier transform has support in $[-\frac{1}{2}, \frac{1}{2}]$. This latter DPP generates almost surely infinite configurations on \mathbb{R} .

A paradigm in random matrix theory, already appearing in the pioneering works of Dyson and Wigner, is the idea that the local behaviour of the eigenvalues exhibits universality, unlike their global behaviour. To illustrate this phenomenon, let us first introduce another popular model of random matrices: the Gaussian Unitary Ensemble (GUE). This time we work on the space of Hermitian matrices $H_n(\mathbb{C}) := \{M \in \mathcal{M}_n(\mathbb{C}) : M^* = M\}$ equipped with a Gaussian measure. More precisely, the isomorphism $H_n(\mathbb{C}) \simeq \mathbb{R}^{n^2}$ induces a Lebesgue measure dM as well as an Euclidean norm $\|M\| = \text{Tr}(M^2)^{1/2}$ on $H_n(\mathbb{C})$, and we consider the Gaussian measure $g_n(M)$ with density proportional to $e^{-\frac{\alpha}{2}\|M\|^2}$. A change of variables and a computation of the Jacobian similar to the one for unitary matrices shows that the joint probability distribution of the eigenvalues $\lambda_1, \dots, \lambda_n \in \mathbb{R}$ of the random matrix of law $g_n(M)$ is given by:

$$d\mathbb{P}(\lambda_1, \dots, \lambda_n) := \frac{1}{Z_n} \prod_{j < k} |\lambda_j - \lambda_k|^2 \prod_{j=1}^n e^{-\frac{\alpha}{2}\lambda_j^2} d\lambda_j,$$

where $Z_n > 0$ is an explicit normalising constant⁹. After further manipulations one obtains that the set of eigenvalues $\{\lambda_1, \dots, \lambda_n\}$ is a DPP on \mathbb{R} with kernel:

$$K_{H_n(\mathbb{C})}(x, y) := \sum_{k=0}^{n-1} \Psi_k^n(x) \Psi_k^n(y)$$

8 We allowed ourselves to discretely add the term $e^{-i\frac{n-1}{2}(x-y)}$ because modifying a kernel $K(x, y)$ to $K(x, y)\frac{f(x)}{f(y)}$ where f does not vanish does not change the determinants associated with the correlation functions; thus, these two kernels generate the same DPP.

9 It is a ratio of products of Gamma functions, obtained as a limit case for Selberg's integral formula.

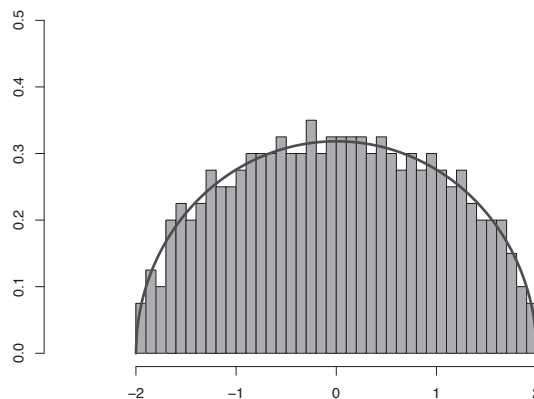


Figure 4. Histogramme des valeurs propres d'une réalisation du GUE de taille 300 et loi du demi-cercle

where, for all $n \geq 1$, $(\Psi_k^n)_{k \in \mathbb{N}}$ is an orthonormal basis of $L^2(\mathbb{R})$ given by renormalised Hermite functions.¹⁰ Here again, we are dealing with a projection kernel on a subspace of $L^2(\mathbb{R})$ of finite dimension, the subspace of functions which can be written $P(x)e^{-nx^2/4}$ where P is a polynomial of degree at the most $n - 1$. When $n \rightarrow \infty$, as illustrated in Figure 4, the eigenvalues concentrate on the compact $[-2, 2]$ with a density $\rho(x) := \frac{1}{2\pi} \sqrt{4 - x^2} \mathbf{1}_{[-2,2]}(x)$, known as the *semicircle law*.

If we now zoom in at a factor n , around a point x_0 from $(-2, 2)$, it can be shown that the local behaviour of the eigenvalues around that point is again governed by the sine kernel K_{\sin} . More precisely, one has the uniform local convergence

$$\frac{1}{\rho(x_0)n} K_{H_n(\mathbb{C})}\left(x_0 + \frac{x}{\rho(x_0)n}, x_0 + \frac{y}{\rho(x_0)n}\right) \xrightarrow{n \rightarrow \infty} K_{\sin}(x, y).$$

Note that the limit does not depend on x_0 . The proof of this convergence is a little more delicate than in the case of unitary matrices. It requires a close analysis of the asymptotic behavior of the Hermite functions.

On the other hand, if we have a closer look at the points at the edge of the semicircle, after some rescaling, another kernel appears, known as the *Airy kernel*,

$$K_{\text{Airy}}(x, y) := \frac{\text{Ai}(x)\text{Ai}'(y) - \text{Ai}'(x)\text{Ai}(y)}{x - y}.$$

Here, $\text{Ai}(x)$ is the Airy function, which satisfies $\text{Ai}''(x) = x\text{Ai}(x)$ and which plays a role in the study of the optical properties of rainbows. For example, at point 2 (the analysis at -2 is the same),

$$\begin{aligned} K_{H_n(\mathbb{C})}^{\text{edge}}(x, y) &:= \frac{1}{n^{2/3}} K_{H_n(\mathbb{C})}\left(2 + \frac{x}{n^{2/3}}, 2 + \frac{y}{n^{2/3}}\right) \\ &\xrightarrow{n \rightarrow \infty} K_{\text{Airy}}(x, y). \end{aligned}$$

It is possible to show that this convergence takes place in the sense of the convergence of the trace-class operators on $L^2(s, \infty)$ for any $s > 0$. We can deduce the fluctuations of the largest eigenvalue around the right edge. Indeed, the distribution function of the largest particle of a DPP on \mathbb{R} can be expressed in terms of a Fredholm determinant, namely

$$\mathbb{P}\left(n^{2/3}\left(\max_{j=1}^n \lambda_j - 2\right) \leq s\right) = \det(I - K_{H_n(\mathbb{C})}^{\text{edge}})_{L^2(s, \infty)},$$

10 If $H_k(x) := (-1)^k e^{x^2/2} (\frac{d}{dx})^k (e^{-x^2/2})$ is the k -th Hermite polynomial, we use $\Psi_k^n(x) := c_k^n H_k(\sqrt{n}x) e^{-nx^2/4}$ with $c_k^n := \int_{\mathbb{R}} H_k(\sqrt{n}x)^2 e^{-nx^2/2} dx$.

which is continuous for this topology. Thus we obtain, for all $s \in \mathbb{R}$,

$$\mathbb{P}\left(n^{2/3} \left(\max_{j=1}^n \lambda_j - 2\right) \leq s\right) \xrightarrow{n \rightarrow \infty} F_2(s) := \det(I - K_{\text{Airy}})_{L^2(s, \infty)}.$$

The probability law associated with the distribution function F_2 is known as the *Tracy–Widom law*. The latter has established an explicit formula for $F_2(s)$ in terms of the Hastings–McLeod solution of the Painlevé II equation. It is also the law of the largest particle of the DPP associated with the *Airy kernel*. We are going to meet this law again a little bit later, in a different context than random matrices.

After a series of works that cannot possibly all be quoted here, it could be observed that the sine and Airy kernels are ubiquitous in the description of the local behaviour of the eigenvalues in a huge number of random matrix models, linked with DPP or not. We speak of *universality phenomenon* in random matrices. It is surprising to observe that the two universal DPPs also appear outside the context of random matrices; in the following, we are going to present three particularly striking examples.

Orthogonal polynomials and the sine kernel

At first we would like to present the elegant approach of Lubinsky [9] for the DPPs related with a family of orthogonal polynomials. We consider a positive Borel measure μ supported in $[-1, 1]$ such that $L^2(\mu)$ contains all polynomials. We can then define an orthonormal family of polynomials $(p_k)_{k \in \mathbb{N}}$ in $L^2(\mu)$ with positive dominant coefficients γ_k . We denote by $w(x)$ the Radon–Nikodym derivative of μ with respect to the Lebesgue measure, such that $\mu(dx) = w(x)dx + \mu_s$ with μ_s being singular with respect to the Lebesgue measure. We next introduce the Christoffel–Darboux kernel, well-known in approximation theory:

$$K_n^{(\mu)}(x, y) = \sqrt{w(x)w(y)} \sum_{k=0}^{n-1} p_k(x)p_k(y).$$

The DPP associated with this kernel almost surely generates configurations of n points in $[-1, 1]$. We suppose that μ is a *regular* measure on $[-1, 1]$, in the sense that $\gamma_k^{1/k}$ converges towards 2 when $k \rightarrow \infty$. This is the case for, e.g., the classic families like the Legendre, Tchebychev and, more generally, Jacobi polynomials. It is also the case once $w(x) > 0$ almost everywhere on $[-1, 1]$. Lubinsky proves the following result: for any x_0 of $(-1, 1)$ such that μ is absolutely continuous on a neighbourhood of x_0 , and whose density w is continuous and strictly positive at x_0 , one has the uniform local convergence:

$$\frac{1}{n\rho(x_0)} K_n^{(\mu)}\left(x_0 + \frac{x}{n\rho(x_0)}, x_0 + \frac{y}{n\rho(x_0)}\right) \xrightarrow{n \rightarrow \infty} K_{\text{sin}}(x, y)$$

where $\rho(x) := 1/(\pi(\sqrt{1-x^2}))$. Note that, under the stronger assumption that $\mu(dx) = w(x)dx$, with a positive and continuous w on $[-1, 1]$, one has a global convergence of the DPP points associated with $K_n^{(\mu)}$ towards the arcsine law with density $\rho(x)$, as illustrated by Figure 5. The proof of this result is surprisingly short, as compared to what is usually done in this field of mathematics. It uses elementary analysis in a very clever way to offer a robust method for the comparison between the kernels associated with different measures.

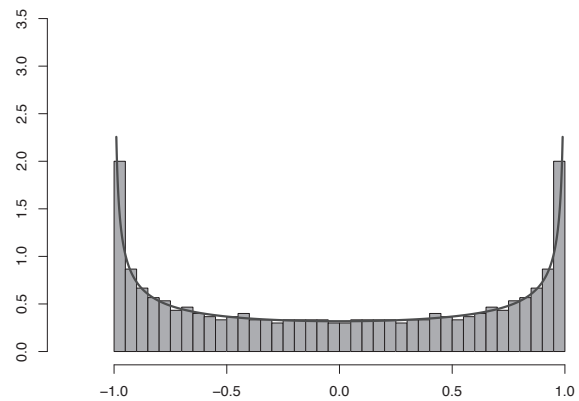


Figure 5. Histogram of a DPP of size 300 associated to the Christoffel–Darboux kernel and the arcsine law

Riemann zeta function and the sine kernel

Here we are going to discuss a somewhat unexpected occurrence of the sine process in analytic number theory. For a complex number s with real part $\Re(s) > 1$, the Riemann zeta function is defined by

$$\zeta(s) := \sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_{p \text{ prime}} \frac{1}{1 - p^{-s}}.$$

This function admits a meromorphic extension to \mathbb{C} , which possesses “trivial” zeros at the negative even integers: $-2, -4, -6$, etc. The other zeros are localized in the region $0 \leq \Re(s) \leq 1$. The celebrated Riemann conjecture claims that those zeros all have a real part equal to $1/2$. Under Riemann hypothesis, one can thus write those zeros into the form $1/2 \pm it_j$ with $0 < t_1 < t_2 < \dots$ due to the symmetry $\zeta(\bar{s}) = \overline{\zeta(s)}$. Thus, a classical result is that, as $n \rightarrow \infty$,

$$N_n := \#\{j \geq 1 : t_j \leq n\} \sim \frac{n}{2\pi} \log \frac{n}{2\pi},$$

which leads us to set

$$w_j := \frac{t_j}{2\pi} \log \frac{t_j}{2\pi},$$

so that $\#\{j \geq 1 : w_j \leq n\} \sim n$ when $n \rightarrow \infty$; the typical distance between the consecutive w_j should be of order one. In order to better understand the distances between the renormalized zeros, at least asymptotically, we consider the analogue of correlation functions for point processes (1.3):

$$R_k^{(n)}(\varphi) := \frac{1}{n} \sum_{1 \leq i_1 \neq \dots \neq i_k \leq n} \varphi(w_{i_1}, \dots, w_{i_k}),$$

for $k \geq 1$ and smooth functions $\varphi : \mathbb{R}^k \rightarrow \mathbb{R}$. Rudnick and Sarnak obtain, under rather strong assumptions on φ , that

$$\lim_{n \rightarrow \infty} R_k^{(n)}(\varphi) = \int_{s_1 + \dots + s_k = 0} \varphi(s) \det \left[K_{\text{sin}}(s_i, s_j) \right]_{i,j=1}^k ds,$$

which is a result similar to (1.5). Thus, asymptotically, the distances between the renormalized zeros behave like a typical realisation of the sine process. The restriction of the integration domain to the hyperplane $s_1 + \dots + s_k = 0$ is due to the fact that $R_k^{(n)}(\varphi)$ may depend only on the distances between the w_j 's. In the case of pairwise correlations, one has, for instance, for $\varphi(x, y) = f(x - y)$,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{1 \leq j \neq \ell \leq n} f(w_j - w_\ell) = \int_{-\infty}^{\infty} f(y) \left(1 - \left(\frac{\sin \pi y}{\pi y}\right)^2\right) dy.$$

This particular case was proven by Montgomery in the 1970s, for the functions $f : \mathbb{R} \rightarrow \mathbb{R}$ whose Fourier transform is C^∞ and with support in $[-1, 1]$. The Montgomery conjecture, which is still open, states that the result must be true without any support restriction. It is said that, in 1972, Montgomery, who was then still a student but already author of the above result, met the physicist Dyson at an afternoon tea in Princeton. Dyson, an expert in random matrices, immediately recognized in Montgomery's result the famous sine kernel.

Random permutations and Airy kernel

We end with a surprising occurrence of the Tracy–Widom law in the study of random permutations. If $\sigma \in \mathfrak{S}_n$ is a permutation of $\llbracket 1, n \rrbracket$, we say that $\sigma(i_1) < \sigma(i_2) < \dots < \sigma(i_k)$ with $i_1 < i_2 < \dots < i_k$ is an increasing subsequence of σ . We denote by $\ell(\sigma)$ the longest length of an increasing subsequence. We thus have $1 \leq \ell(\sigma) \leq n$. For example, if

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 5 & 3 & 2 & 6 & 1 & 7 & 9 & 4 & 8 \end{pmatrix},$$

then $\ell(\sigma) = 4$ and it is reached at the increasing subsequence 5, 6, 7, 8 (but also at 2, 6, 7, 9). If now σ_n is a random permutation sampled uniformly on the symmetric group \mathfrak{S}_n , the study of the asymptotic behavior of $\ell(\sigma_n)$ for large n is known as the *Ulam problem*. After the works of Hammersley, Vershik, Kerov, Logan and Schepp, one obtains that

$$\mathbb{E}[\ell(\sigma_n)] \sim 2\sqrt{n}, \quad n \rightarrow \infty.$$

An important breakthrough has been made by Baik, Deift and Johansson [2] who obtained the fluctuations of $\ell(\sigma_n)$ around its average value: for any $s \in \mathbb{R}$,

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(n^{-1/6} (\ell(\sigma_n) - 2\sqrt{n}) \leq s \right) = F_2(s).$$

In other words, the random variable $n^{-1/6}(\ell(\sigma_n) - 2\sqrt{n})$ converges (in the weak * topology) towards the Tracy–Widom distribution when $n \rightarrow \infty$. To understand how the problem is related to DPPs, we won't present the original proof by Baik, Deift and Johansson but rather the one by Borodin, Okounkov and Olshanski [3]. Their analysis is based on the Robinson–Schensted (RS) correspondence, well-known in representation theory of the symmetric group, which, to a permutation $\sigma \in \mathfrak{S}_n$ associates a pair of Young tableaux of the same shape. More precisely, a partition $\lambda = (\lambda_1, \dots, \lambda_\ell)$ of an integer n , that is, a decreasing sequence of positive integers with sum n , is encoded by a *Ferrers diagram* of n boxes with λ_i boxes on the i -th line. For example, the partition (4, 4, 3, 1) of $n = 12$ is encoded by the diagram:



Given a Ferrers diagram λ with n boxes, a *Young tableau* of shape λ is a filling of the n boxes with the integers from 1 to n in a strictly increasing way along the lines and columns. The RS correspondence associates to a permutation σ two Young tableaux of the same shape. We successively place the integers $\sigma(1), \sigma(2), \dots$ in the first tableau, starting in the first line and with the following rule: every new element $\sigma(j)$ is inserted into the first line. If $\sigma(j)$ is the largest integer of the line, we place it into a new box on the right. Otherwise, it

takes the place of the smallest integer that is larger than itself. The latter is thus put down to the next line where the same rules apply. The second tableau keeps track of the order in which the boxes have been created. Thus for

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 5 & 3 & 2 & 6 & 1 & 7 & 9 & 4 & 8 \end{pmatrix},$$

the RS correspondence follows the successive steps:

5	3	2	26	16	167	1679	1479	1478
	5	3	3	2	2	2	26	269
		5	5	3	3	3	3	3
				5	5	5	5	5
<hr/>								
1	1	1	14	14	146	1467	1467	1467
		2	2	2	2	2	28	289
			3	3	3	3	3	3
				5	5	5	5	5

Two key properties are used here. First, $\ell(\sigma) = \lambda_1$, the length of the first line of the tableaux obtained by the RS correspondence. Then, when we apply the RS correspondence to the uniform permutation σ_n , the probability that the shape of the resulting tableaux is the partition λ of n equals $\text{PL}_n(\lambda) := \frac{1}{n!} \dim(\lambda)^2$. Here $\dim(\lambda)$ is the number of Young tableaux with shape λ ; it is also the dimension of the irreducible representation of \mathfrak{S}_n indexed by the partition λ . This probability distribution PL_n on the Ferrers diagrams with n boxes is called the *Plancherel measure*. Consider now the *poissonization* \mathbb{P}^θ of parameter $\theta > 0$ of the Plancherel measures, which generates Ferrers diagrams of random size: we sample an integer N with Poisson law of parameter θ and then sample a Ferrers diagram of N boxes following the Plancherel measure PL_N . Thus, if one denotes by $|\lambda| := \lambda_1 + \dots + \lambda_\ell$ the number of boxes of a Ferrers diagram λ , the probability under \mathbb{P}^θ to obtain a diagram λ is:

$$\mathbb{P}^\theta(\lambda) := e^{-\theta} \theta^{|\lambda|} \left(\frac{\dim(\lambda)}{|\lambda|!} \right)^2.$$

This poissonisation is motivated by the following fact: if $\lambda = (\lambda_1, \lambda_2, \dots)$ has \mathbb{P}^θ as its distribution, then the random configuration $x_i := \lambda_i - i$ is a DPP on \mathbb{Z} . Its kernel $K^\theta(x, y)$ can be written as a double integral on complex contours; this rewriting is well-suited for asymptotic analysis (saddle-point method). From there, Borodin, Okounkov and Olshanski have proven that

$$\lim_{\theta \rightarrow \infty} \theta^{1/6} K^\theta(2\sqrt{\theta} + x\theta^{1/6}, 2\sqrt{\theta} + y\theta^{1/6}) = K_{\text{Airy}}(x, y).$$

The convergence of $n^{-1/6}(\ell(\sigma_n) - 2\sqrt{n}) = n^{-1/6}(\lambda_1 - 2\sqrt{n})$ towards the Tracy–Widom law cited above is a consequence of the latter, because, under \mathbb{P}^θ , the number of boxes of a tableau concentrates around θ when $\theta \rightarrow \infty$ (depoissonisation procedure). In other words, for large integer θ , the distributions \mathbb{P}^θ and PL_θ generate diagrams which are similar in a sense that can be quantified.

DPP and machine learning

To conclude our little journey through the vast land of DPP, we would like to go back to the jaguar example from the introduction and outline some aspects of the use of DPP in the

context of machine learning. The main message is that when it comes to modelling situations which involve repulsion, diversity and negative correlations between objects, DPP can provide efficient models that are easy to handle and easy to sample. To go deeper into the subject, we recommend for example [7].

In most of these applications, one tries to select a (random) subset of objects in a database, that is, in a discrete set \mathcal{E} of cardinal n , possibly very large; typically images or texts. We then consider a class of DPP which generate configurations $\Xi \subset \mathcal{E}$ of random cardinal. In that context, we attribute to the i -th element of \mathcal{E} a vector $B_i \in \mathbb{R}^d$, where the dimension d is fixed by the user. For instance, B_i encodes the pixels of the i -th image and d will depend on the chosen resolution. We then consider the semi-definite positive matrix $L := B^t B$ where B is the $d \times n$ matrix with columns B_i . We next take $K := L(I + L)^{-1}$ as a kernel of a DPP $\Xi \subset \mathcal{E}$. Thus, for all $A \subset \mathcal{E}$,

$$\mathbb{P}(A \subset \Xi) = \det(K_A) := \det[K_{ij}]_{i,j \in A},$$

with the convention $\det(K_\emptyset) := 1$. In terms of the matrix L , one has

$$\mathbb{P}(\Xi = A) = \frac{\det(L_A)}{\det(I + L)}.$$

We see that $\mathbb{P}(\Xi = A)$ equals, up to a normalisation constant, the square volume of the polytope generated by the columns B_i of B with $i \in A$. Thus, if $B_i = q_i \varphi_i$ with $q_i \in \mathbb{R}^+$ and $\varphi_i \in \mathbb{R}^d$ with $\|\varphi_i\| = 1$, then q_i is interpreted as a measure of the importance of the i -th object of \mathcal{E} , while $S_{ij} := \varphi_i^t \varphi_j \in [-1, 1]$ represents a measure of the similarity between the i -th and the j -th object. More precisely, $\mathbb{P}(\Xi = A)$ is proportional to $\prod_{i \in A} q_i^2 \det(S_A)$. In practice, every object of \mathcal{E} needs to be labelled with its attribute (q_i, φ_i) , i.e., in the same way that PageRank assigns a level of importance to every web page, and then we compute the associated matrices L and K . A realisation of the DPP with kernel K will then provide a subset of \mathcal{E} presenting diversity in the sense of the matrix S . For example, if \mathcal{E} is the set of images associated with the key word “jaguar” on the internet, and we have previously labelled each of its elements with the attributes (q_i, φ_i) , then a realisation of the DPP Ξ will provide subset of images exhibiting diversity.

It is also possible to restrict oneself to a parametrised family of kernels K_θ and estimate θ by means of the usual statistical methods from the training data. By this method, Kulesza and Taskar [7] develop, as an example, the extractive text summarisation where you have numerous texts on the same subject (i.e., newspaper articles on a time line) and you want to extract a summary, this is to say, a small number of sentences which contain as much information as possible.

The interest of these models is mainly due to the fact that they are easy to implement in an exact manner: one can reply to most of the questions on inference in polynomial time, essentially by multiplying, inverting, diagonalising or computing the determinants of matrices of size n , at a cost of $O(n^3)$ elementary operations. In particular, we have exact algorithms at our disposal for sampling a DPP at this cost.

The use of DPP in machine learning, as well as in spatial statistics ([8]) or in numerical integration, is still in its infancy but seems to spark the interest of more and more users of applied mathematics.

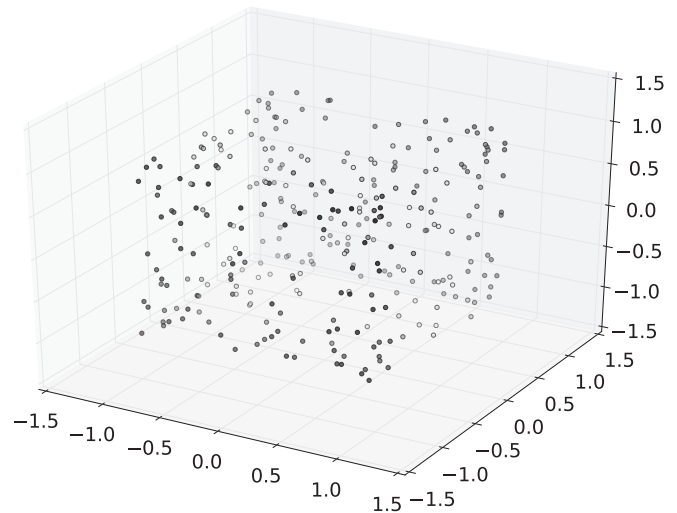


Figure 6. Sampling of a DPP associated to orthogonal polynomials in several variables on the cube

Acknowledgements

Thanks to Raphael, Rosalie and Elvire for their innocent simulations, furthermore to Rémi Bardenet and Guillaume Gauthier. Special thanks to Sophie Grivaux for the attentive checking of the text’s first french version and for her patience. The *EMS Newsletter* thanks *La Gazette des Mathématiciens* for authorisation to republish this text.

Bibliography

- [1] G. W. Anderson, A. Guionnet, and O. Zeitouni. *An Introduction to Random Matrices*, volume 118 of Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, 2010.
- [2] J. Baik, P. Deift, and K. Johansson. On the distribution of the length of the longest increasing subsequence of random permutations. *J. Amer. Math. Soc.*, 12(4):1119–1178, 1999. DOI:10.1090/S0894-0347-99-00307-0.
- [3] A. Borodin, A. Okounkov, and G. Olshanski. Asymptotics of Plancherel measures for symmetric groups. *J. Amer. Math. Soc.*, 13(3):481–515, 2000. DOI:10.1090/S0894-0347-00-00337-4.
- [4] A. Borodin, P. Diaconis, and J. Fulman. On adding a list of numbers (and other one-dependent determinantal processes). *Bull. Amer. Math. Soc. (N.S.)*, 47(4):639–670, 2010. DOI:10.1090/S0273-0979-2010-01306-9.
- [5] J. B. Hough, M. Krishnapur, Y. Peres, and B. Virág. Determinantal processes and independence. *Probab. Surv.*, 3:206–229, 2006. DOI:10.1214/154957806000000078
- [6] K. Johansson. Random matrices and determinantal processes. In *Mathematical statistical physics*, pages 1–55. Elsevier B. V., Amsterdam, 2006. DOI:10.1016/S0924-8099(06)80038-7
- [7] A. Kulesza and B. Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 16, 2012.
- [8] F. Lavancier, J. Møller, and E. Rubak. Determinantal point process models and statistical inference. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 77(4): 853–877, 2015. DOI:10.1111/rssb.12096.

-
- [9] D. S. Lubinsky. A new approach to universality limits involving orthogonal polynomials. *Ann. of Math. (2)*, 170(2):915–939, 2009. DOI:10.4007/annals.2009.170.915.
- [10] R. Lyons. *Determinantal Probability Measures*. Publications Mathématiques de l’Institut des Hautes Études Scientifiques, 2003.
- [11] A. Soshnikov. Determinantal random point fields. *Uspekhi Mat. Nauk*, 55(5(335)):107–160, 2000. DOI:10.1070/rm2000v055n05ABEH000321



Adrien Hardy [adrien.hardy@univ-lille.fr] is an associate professor at the University of Lille. His research interest lies principally in understanding large random interacting particle systems and their applications.



Mylène Maïda [mylene.maida@univ-lille.fr] has been a professor at the University of Lille since 2013. Her research interests are in probability theory and statistical physics.