From random matrices to Monte Carlo integration via Gaussian quadrature

R. Bardenet, Adrien Hardy

▶ **To cite this version:**

R. Bardenet, Adrien Hardy. From random matrices to Monte Carlo integration via Gaussian quadrature. IEEE Statistical Signal processing workshop, 2018, Freiburg, Germany. hal-01882393

HAL Id: hal-01882393
https://hal.science/hal-01882393v1

Submitted on 26 Sep 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

FROM RANDOM MATRICES TO MONTE CARLO INTEGRATION VIA GAUSSIAN QUADRATURE

Rémi Bardenet[1] *and Adrien Hardy*[2]

[1] Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL, 59651 Villeneuve d'Ascq, France
[2] Laboratoire Paul Painlevé, Univ. Lille, Cité Scientifique, 59655 Villeneuve d'Ascq, France

ABSTRACT

We introduced in [1] a new Monte Carlo estimator that relies on determinantal point processes (DPPs). We were initially motivated by peculiar properties of results from random matrix theory. This motivation is absent from the original paper [1], so we develop it here. Then, we give a non-technical overview of the contents of [1], insisting on points that may be of interest to the statistical signal processing audience.

***Index Terms*—** Monte Carlo, random matrices, DPPs

1. A PECULIAR CENTRAL LIMIT THEOREM

Much of random matrix theory deals with the behaviour of the eigenvalues of certain random matrices, as can be gathered from the reference book [2]. 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 [3].

Let $U_N \subset \mathcal{M}_N(\mathbb{C})$ 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

$$\mu_H(A) = \mu_H(\{ga; a \in A\}) \tag{1}$$

for any $g \in \mathrm{U}_N$ and any Borel set $A \subset \mathrm{U}_N$, and this measure is called *the Haar measure* on U_N, see for instance [2, 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 (θ_i). Johansson [3] used a result called the *strong Szegő* theorem to prove that for any $g : [0, 2\pi] \to \mathbb{R}$ such that $g \in L^1$ and its Fourier coefficients satisfy

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

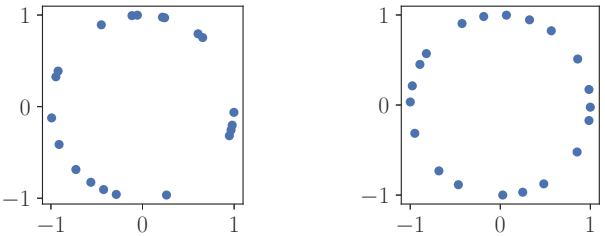
This project is funded by ANR grant BoB ANR-16-CE23-0003.

it holds

$$\sum_{i=1}^{N} g(\theta_i) - N\int_0^{2\pi} g(\theta)\frac{\mathrm{d}\theta}{2\pi} \to \mathcal{N}(0, \sigma^2), \tag{3}$$

where the convergence is in distribution when $N \to \infty$. The central limit theorem (CLT) in Equation (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) implies that the mean square error of I_N is asymptotically equivalent to $1/N$. This contrasts with traditional Monte Carlo errors of order $1/\sqrt{N}$.

Efficient cancellation happens in the variance of the left-hand side of (3). One way to get 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 1.



(a) Uniform (b) Haar

Fig. 1. (a) N i.i.d. uniform samples on the unit circle. (b) A sample of the eigenvalues of a Haar-distributed matrix in U_N.

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 (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).

1.1. The requirements for a *good* generalization

The starting point of our work [1] is an investigation whether results such as Johansson's can be generalized so as to become useful tools for numerical integration. In other words, we want to reverse-engineer (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 $d \geq 1$. Second, we would like to integrate against an arbitrary measure, whereas random matrix results such as (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 (2) as a measure of the decay of the Fourier coefficients of the integrand.

2. PROJECTION DPPS

The eigenvalues of a Haar-distributed unitary matrix are a projection determinantal point process (DPP). DPPs are a class of point processes that can encode repulsiveness, and we will show that they satisfy the requirements of Section 1.1.

The study of DPPs was pioneered by Macchi in 1975 [4], motivated by models for the spatial distribution of beams of fermions [5]. Since then, they have received a lot of attention in probability [6, 7, 8, 9], mostly for their applications in random matrix theory, and more recently in spatial statistics [10], machine learning [11], and signal processing [12, 13].

2.1. Repulsiveness and volumes

In this paper, we only define a subfamily of DPPs called *projection DPPs*. Let μ be a positive Borel measure on $[-1, 1]^d$ with finite mass, and density ω w.r.t. to Lebesgue. Consider N orthonormal functions $\phi_0, \dots, \phi_{N-1}$ in $L^2(\mu)$ and let

$$K_N(x, y) \triangleq \sum_{k=0}^{N-1} \phi_k(x)\phi_k(y). \tag{4}$$

The set $X = \{\mathbf{x}_1, \dots, \mathbf{x}_N\} \subset [-1, 1]^d$ is said to be drawn from a DPP with base measure μ and kernel K_N, denoted by $X \sim \text{DPP}(\mu, K_N)$, if and only if the points $\mathbf{x}_1, \dots, \mathbf{x}_N$ have joint probability distribution

$$\frac{1}{N!} \det \left[K_N(x_i, x_\ell)\right]_{i,\ell=1}^{N} \prod_{i=1}^{N} \mu(\mathrm{d}x_i). \tag{5}$$

Note that (5) is invariant to permutations of the N labels, so that we have unambiguously defined a probability measure over sets. The fact that (5) is a probability density is a direct consequence of the orthonormality of the ϕ_is and developing the determinant in (5).

The potential to model repulsiveness for projection DPPs is best understood when rewriting (4) as

$$\prod_{i=1}^{N} \frac{1}{N - i + 1} \left\| P_{H_{i-1}} K_N(x_i, \cdot) \right\|^2_{L^2(\mu)} \mu(\mathrm{d}x_i). \tag{6}$$

where P_H is the orthogonal projection onto $H \subset L^2$,

$$H_0 = \text{Span}(\phi_0, \dots, \phi_{N-1}),$$

and H_{i-1} is the orthocomplement in H_0 of

$$\text{Span}\left(K_N(x_\ell, \cdot),\ 1 \leq \ell \leq i - 1\right).$$

Seeing (6) as a sequence of "base times height" formulas, the pdf of a projection DPP is proportional to the square of the volume of the parallelotope spanned by the N functions $K(x_i, \cdot)$ in H_0. The configurations X that have large pdf values are those that correspond to large volumes in H_0, which machine learners would call the *feature space*. These volumes determine which configurations X are repulsive.

Figure 2 further illustrates this concept. Let the base measure have density $\omega(x) = (1 - x)^4(1 + x)$, and let $N = 100$. Let (ϕ_k) be the orthonormal polynomials in $L^2(\mu)$, that is, $\deg(\phi_k) = k$ for all $k \geq 0$, the leading coefficient of ϕ_k is positive, and $\int \phi_k(x)\phi_\ell(x)\omega(x)\mathrm{d}x = \delta_{k\ell}$. We show two of the feature functions $f_k : x \mapsto \phi_k(x)\sqrt{\omega(x)}$ for $k = 2, 10$. For any choice of N points $x_1, \dots, x_N$, their pdf (6) is the volume in $\mathbb{R}^N$ of the parallelotope spanned by the N vectors $v(x_i) = (f_1(x_i), \dots, f_N(x_i)) \subset \mathbb{R}^N$. We know from the theory of orthogonal polynomials [14] that the norm of $v(x)$ is roughly constant when x is bounded away from -1 and 1, say $x \in [-0.9, 0.9]$. In that interval, pairs (x_1, x_2) that span large parallelograms thus correspond to vectors $v(x_1)$ and $v(x_2)$ with small inner product. Two equally-spaced pairs of abscissa are highlighted in Figure 2, one in plain red, and one in dashed red. We are now equipped to compare the chance of the red pair to be included in sample of the DPP with kernel (4), to that of the dashed red pair. For each abscissa x in these pairs, the intersections of the red and blue, and red and green lines give two coordinates of $v(x)$.

Orthogonal polynomials "oscillate" faster [14] in regions of high mass under the base measure $\mu(\mathrm{d}x) = \omega(x)\mathrm{d}x$. The plain red pair lies in the bulk of ω, and the fast oscillations of the f_ks make the inner product $\langle v(x_1), v(x_2)\rangle$ small compared to the same inner product evaluated at the dashed red pair. The plain red pair is thus more likely to co-occur in a DPP sample than the dashed red pair. Otherly put, repulsiveness is "weaker" in the bulk of ω. Note that regions close to the endpoints of $[-1, 1]$ are exceptions: they correspond both to large norms and small inner products, and we thus expect some samples of this particular DPP to cluster there.

2.2. DPPs are kernel machines

DPPs are akin to support-vector machines (SVM, [15]) or Gaussian processes [16], in that the statistical properties of

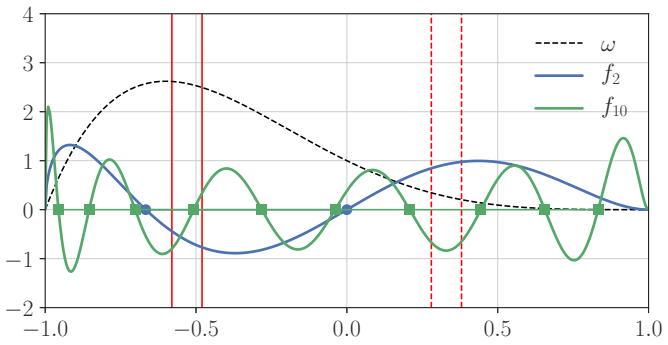


Fig. 2. A so-called *Jacobi* base measure, and two orthonormal eigenfunctions $\phi_2\sqrt{\omega}$ and $\phi_{10}\sqrt{\omega}$.

the point configurations are given by geometric quantities in high-dimensional feature space. In the same manner that the predictors in SVM regression/classification depend on the eigendecomposition of the SVM kernel, the repulsiveness in a DPP is encoded by its eigendecomposition, which for a projection DPP takes the simple form (4).

We further note that using the same *normal* equations as in standard linear regression, it comes

$$\left\|P_{H_{i-1}}K_N(x_i,\cdot)\right\|^2_{L^2(\mu)} = \begin{cases} K_N(x_1,x_1) & \text{if } i=1,\\ K_N(x_i,x_i)-\mathbf{k}_{i-1}(x_i)^T\mathbf{K}_{i-1}^{-1}\mathbf{k}_{i-1}(x_i) & \text{else,}\end{cases} \tag{7}$$

where $\mathbf{k}_{i-1}(\cdot)=(K_N(x_1,\cdot),\dots,K_N(x_{i-1},\cdot))^T$, and

$$\mathbf{K}_{i-1}=\Big[K_N(x_k,x_\ell)\Big]_{1\le k,\ell\le i-1}.$$

Equation (7) will be familiar to users of Gaussian processes (GPs; [16]): the unnormalized conditional densities (7) are the incremental posterior variances in a GP model with the same kernel, giving yet another intuition for repulsiveness. Furthermore, in the absence of a connection to the eigenvalues of a random matrix, the chain rule (6) and the explicit form (7) are commonly used to sample projection DPPs, using rejection sampling [17] to sample each conditional in turn.

2.3. Multivariate orthogonal polynomial ensembles

When $d=1$, letting ϕ_k in (4) be the orthonormal polynomials w.r.t. μ results in a DPP called an *orthogonal polynomial ensemble* (OPE, [18]). 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 need to pick an order $\mathfrak{b}:\mathbb{N}\to\mathbb{N}^d$ before we apply Gram-Schmidt to the first N monomials. In [1] we pick 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 $K_N(x,y)=\sum\Phi_{\mathbf{k}}(x)\Phi_{\mathbf{k}}(y)$, where the sum runs over multi-indices $\mathbf{k}$ such that $0\le\mathfrak{b}(\mathbf{k})\le N-1$. Note that for a separable base measure, i.e. if

$$\omega(x)=\omega^1(x^1)\dots\omega^d(x^d), \tag{8}$$

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

$$K_N(x,y)=\sum_{\mathfrak{b}(\mathbf{k})=\mathbf{0}}^{N-1}\prod_{j=1}^{d}\phi^j_{k_j}(x^j)\phi^j_{k_j}(y^j), \tag{9}$$

where $(\phi^j_k)_{k\ge0}$ are the orthonormal polynomials w.r.t. ω^j.

3. MONTE CARLO WITH DPPS

3.1. A stochastic Gaussian quadrature

Let μ be separable as in (8), and $X\sim\mathrm{DPP}(\mu,K_N)$ be a multivariate OPE given by (9). If we integrate (5) $N-1$ times, we obtain that any point in our DPP has marginal pdf $K_N(x,x)\omega(x)$. For $g\in L^1(\mathrm{d}\mu)$, we deduce that

$$\hat{I}_N\triangleq\sum_{i=1}^{N}\frac{g(\mathbf{x}_i)}{K_N(\mathbf{x}_i,\mathbf{x}_i)} \tag{10}$$

is an unbiased estimator of $\int g(x)\omega(x)\mathrm{d}x$. This estimator is familiar to users of Gaussian quadrature. Gauss indeed showed [19] that if you restrict to $d=1$, and replace the DPP samples in (10) by the zeros of ϕ_N, then $\hat{I}_N$ is deterministic and has zero error when g is a polynomial of degree less than $2N-1$. This is (or was) unexpected as ϕ_N has degree N and thus only N zeros. These zeros are real and distinct in non-pathological cases [14], but still, there are only N of them and they allow integrating polynomials of degree twice that. Remembering the intuition about the repulsiveness in Section 2.1, we see that OPEs are a good candidate for the natural stochastic version of Gaussian quadrature.

To further visualize similarity of OPEs with Gaussian quadrature, we plot in Figure 3 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 $\mathbf{x}_i$ is depicted

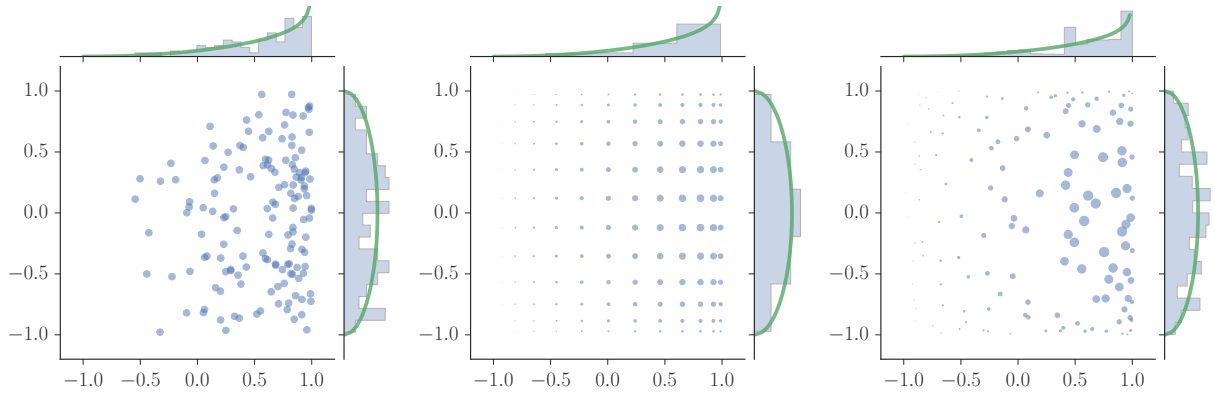


Fig. 3. Left: i.i.d. sample, middle: product Gaussian quadrature, right: bivariate OPE sample

with a circle, the area of which is proportional to the node's weight $1/K_N(\mathbf{x}_i, \mathbf{x}_i)$ in (10). 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. Our generalization of Johansson's CLT

Theorem 1 ([1]). *Let $\mu(\mathrm{d}x) = \omega(x)\mathrm{d}x$ with ω separable, $\mathscr{C}^1$, positive on the open set $(-1,1)^d$, and satisfying a technical regularity assumption. If $\mathbf{x}_1, \ldots, \mathbf{x}_N$ stands for the associated multivariate OPE, then for every g $\mathcal{C}^1$ vanishing outside $[-1+\epsilon, 1-\epsilon]^d$ for some $\epsilon > 0$,*

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

where

$$\Omega^2_{g,\omega} = \frac{1}{2}\sum_{k_1,\ldots,k_d=0}^{\infty}(k_1+\cdots+k_d)\widehat{\left(\frac{g\omega}{\omega_{eq}^{\otimes d}}\right)}(k_1,\ldots,k_d)^2, \tag{11}$$

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

If we take Johansson's result (3), and project the eigenvalues orthogonally from the unit circle onto the x-axis, we obtain a special case of our Theorem 1 where $d = 1$ and $\mu = \mu_{eq}$. But for some additional assumptions on f, we have thus generalized (3). Looking back at our requirements in Section 1.1, we have extended (3) to an arbitrary dimension d, and an arbitrary choice of target measure. We have preserved the fast convergence of (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 only assumes that the integrand is $\mathcal{C}^1$, while typical faster-than-Monte-Carlo algorithms like quasi-Monte Carlo [20] 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 us weaker assumptions, and gives a fast CLT with interpretable asymptotic variance.

3.3. An importance sampling version

One downside of Theorem 1 is that in practice, one rarely knows 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 tasks where access to μ is limited to pointwise evaluation up to a multiplicative constant [17]. In [1, Theorem 2.9], we show that if $\mathbf{x}_1, \ldots, \mathbf{x}_N$ is the multivariate OPE associated to a pdf q that satisfies the assumption of Theorem 1, then $\tilde{I}_N \triangleq \sum_{i=1}^N \frac{g(\mathbf{x}_i)}{K_N(\mathbf{x}_i,\mathbf{x}_i)}\frac{\omega(\mathbf{x}_i)}{q(\mathbf{x}_i)}$ satisfies the same CLT as in Theorem 1, with the same asymptotic variance. In other words, you do not incur any cost, asymptotically, for having replaced ω by another base 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.

4. CONCLUSION

We have taken a different view at the results in [1], highlighting the motivation that stemmed from random matrix theory. The results take us to a stochastic version of Gaussian quadrature, and ask questions of harmonic analysis: if g is known to be sparse in some basis of L^2, say a given wavelet basis, can we prove a CLT like Theorem 1 for a DPP that projects onto the span of N of these wavelets? In that case, would the asymptotic variance measure the decay of the coefficients of g in that wavelet basis? The tools we used in [1] rely heavily on the basis being polynomial, so new ideas are needed...

5. REFERENCES

[1] R. Bardenet and A. Hardy, "Monte Carlo with determinantal point processes," *arXiv preprint arXiv:1605.00361*, 2016.

[2] G. W. Anderson, A. Guionnet, and O. Zeitouni, *An introduction to random matrices*, vol. 118, Cambridge university press, 2010.

[3] K. Johansson, "On random matrices from the compact classical groups," *Ann. Math.*, vol. 145, no. 3, pp. 519–545, 1997.

[4] O. Macchi, "The coincidence approach to stochastic point processes," *Advances in Applied Probability*, vol. 7, pp. 83–122, 1975.

[5] C. Bénard, "Fluctuations of beams of quantum particles," *Physical Review A*, vol. 2, no. 5, pp. 2140, 1970.

[6] A. Soshnikov, "Gaussian limit for determinantal random point fields," *Ann. Probab.*, vol. 30, no. 1, pp. 171–187, 2002.

[7] R. Lyons, "Determinantal probability measures," *Publ. Math. Inst. Hautes Etudes Sci.*, vol. 98, pp. 167–212, 2003.

[8] J. B. Hough, M. Krishnapur, Y. Peres, and B. Virág, "Determinantal processes and independence," *Probab. Surv.*, vol. 3, pp. 206–229, 2006.

[9] K. Johansson, *Random matrices and determinantal processes, Mathematical Statistical Physics*, Elsevier B.V. Amsterdam, 2006.

[10] F. Lavancier, J. Møller, and E. Rubak, "Determinantal point process models and statistical inference," *Journal of Royal Statistical Society: Series B (Statistical Methodology)*, vol. 77, pp. 853–877, 2015.

[11] A. Kulesza and B. Taskar, "Determinantal point processes for machine learning," *Foundations and Trends in Machine Learning*, 2012.

[12] R. Bardenet, J. Flamant, and P. Chainais, "On the zeros of the spectrogram of white noise," arXiv preprint arXiv:1708.00082, 2017.

[13] N. Tremblay, P.-O. Amblard, and S. Barthelmé, "Graph sampling with determinantal processes," in *Proceedings of EUSIPCO*, 2017.

[14] V. Totik, "Orthogonal polynomials," *Surveys in Approximation Theory*, vol. 1, pp. 70–125, 2005.

[15] N. Cristianini and J. Shawe-Taylor, *Kernel methods for pattern recognition*, Cambridge University Press, 2004.

[16] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.

[17] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer-Verlag, New York, 2004.

[18] W. Köning, "Orthogonal polynomial ensembles in probability theory," *Probab. Surv.*, vol. 2, no. 385–447, 2005.

[19] C. F. Gauss, *Methodus nova integralium valores per approximationem inveniendi*, Heinrich Dietrich, Göttingen, 1815.

[20] J. Dick and F. Pilichshammer, *Digital nets and sequences. Discrepancy theory and quasi-Monte Carlo Integration*, Cambridge University Press, 2010.