On two ways to use determinantal point processes for Monte Carlo integration

Guillaume Gautier, R. Bardenet, Michal Valko

To cite this version:

Guillaume Gautier, R. Bardenet, Michal Valko. On two ways to use determinantal point processes for Monte Carlo integration. NEGDEPML 2019 - ICML Workshop on Negative Dependence in ML, Jun 2019, Long Beach, CA, United States. hal-02160382

HAL Id: hal-02160382
https://hal.archives-ouvertes.fr/hal-02160382

Submitted on 19 Jun 2019

On two ways to use determinantal point processes for Monte Carlo integration

Guillaume Gautier [1 2] **Rémi Bardenet** [1] **Michal Valko** [3 2]

Abstract

This paper focuses on Monte Carlo integration with determinantal point processes (DPPs) which enforce negative dependence between quadrature nodes. We survey the properties of two unbiased Monte Carlo estimators of the integral of interest: a direct one proposed by Bardenet & Hardy (2016) and a less obvious 60-year-old estimator by Ermakov & Zolotukhin (1960) that actually also relies on DPPs. We provide an efficient implementation to sample exactly a particular multidimensional DPP called *multivariate Jacobi ensemble*. This let us investigate the behavior of both estimators on toy problems in yet unexplored regimes.

1. Introduction

Numerical integration is a core task of many machine learning applications, including most Bayesian methods (Robert, 2007). Both deterministic and random algorithms have been proposed (Evans & Swartz, 2000). All methods require combining evaluations of the integrand at so-called *quadrature nodes* to minimize the approximation error.

We are motivated by a stream of work which makes use of prior knowledge on the smoothness of the integrand using kernels and RKHSs. Oates et al. (2017) and Liu & Lee (2017) made use of kernel-based control variates, splitting the computational budget into regressing the integrand and integrating the residual. Bach (2017) looked for the best way to sample i.i.d. nodes and combine the resulting evaluations. Bayesian quadrature (O'Hagan, 1991; Huszár & Duvenaud, 2012; Briol et al., 2015), herding (Chen et al., 2010; Bach et al., 2012), or the biased importance sampling estimate of Delyon & Portier (2016) all favor *dissimilar* quadrature nodes, where dissimilarity is measured by a kernel. Our work falls within this last cluster.

We build on the particular approach of Bardenet & Hardy (2016) for Monte Carlo integration based on projection *determinantal point processes* (DPPs, Hough et al., 2006; Kulesza & Taskar, 2012). DPPs are a repulsive distribution over configurations of points; repulsiveness is again parametrized by a kernel. In a sense, DPPs are the kernel machines of point processes.

Our contributions. First, we point out a mostly forgotten Monte Carlo estimator derived by Ermakov & Zolotukhin (1960) (EZ, 1960) that implicitly but crucially requires sampling from a DPP, more than a decade before Macchi (1975) even formalized DPPs! Second, we provide a simple proof of their result and survey the properties of the estimator with modern arguments. In particular, when the integrand is a linear combination of the eigenfunctions of the underlying DPP kernel, the corresponding Fourier-like coefficients can be estimated with zero variance. From one sample of the corresponding DPP, perfect reconstruction of the signal is granted by solving a linear system. Third, we propose an efficient Python implementation for sampling exactly a particular DPP called *multivariate Jacobi ensemble*. This implementation allows to numerically investigate the behavior of the two Monte Carlo estimators derived by Bardenet & Hardy (2016) and Ermakov & Zolotukhin (1960), in regimes yet unexplored for any of the two. Our point is not to compare DPP-based Monte Carlo estimators to the wide choice of numerical integration algorithms, but to get a fine understanding of their properties so as to fine-tune their design and guide theoretical developments.

2. Quadrature, DPPs, and the multivariate Jacobi ensembles

2.1. Standard quadrature

Let $\mu(\mathrm{d}x) = \omega(x)\,\mathrm{d}x$ be a positive Borel measure on $\mathbb{X} \subset \mathbb{R}^d$ with finite mass and density w w.r.t. the Lebesgue measure. This paper aims to compute integrals of the form $\int f(x)\mu(\mathrm{d}x)$ for some test function $f : \mathbb{X} \to \mathbb{R}$. A quadrature rule approximates such integrals as a weighted sum of evaluations of f at some points $\{x_1, \dots, x_N\} \subset \mathbb{X}$,

$$\int f(x)\mu(\mathrm{d}x) \approx \sum_{n=1}^{N} w_n f(x_n), \qquad (1)$$

with $w_n \triangleq w_n(x_1, \dots, x_N) \in \mathbb{R}$ do not need to sum to one.

[1] Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL, France [2] INRIA Lille - Nord Europe, France [3] DeepMind Paris. Correspondence to: Guillaume Gautier <g.gautier@inria.fr>.

Proceedings of the Workshop on Negative Dependence in Machine Learning at the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

Among the many quadrature designs mentioned in the introduction, we pay special attention to the textbook example of the (deterministic) Gauss-Jacobi rule. This scheme applies to $\mathbb{X} \triangleq [-1,1]$ with $\omega(x) \triangleq (1-x)^a(1+x)^b$ where $a, b > -1$. In this case, the nodes $\{x_1, \dots, x_N\}$ are taken to be the zeros of p_N, the orthonormal Jacobi polynomial of degree N, and the weights $w_n \triangleq 1/K(x_n, x_n)$ with $K(x,x) \triangleq \sum_{k=0}^{N-1} p_k(x)^2$. In particular, it allows to perfectly integrate polynomials up to degree $2N-1$ (Davis & Rabinowitz, 1984, Section 2.7). In a sense, the DPPs of Bardenet & Hardy (2016) are a random, multivariate generalization of Gauss-Jacobi quadrature, cf. Section 3.1.

Monte Carlo integration can be defined as random choices of nodes in (1). Importance sampling, corresponds to i.i.d. nodes, while Markov chain Monte Carlo corresponds to nodes drawn from Markov chain (Robert & Casella, 2004). Finally, quasi-Monte Carlo (Dick & Pillichshammer, 2010) apply to μ uniform over a compact subset of $\mathbb{R}^d$, and consists in constructing deterministic nodes that spread very uniformly, as measured by the so-called discrepancy.

2.2. Projection DPPs

DPPs can be understood as a parametric class of point processes, specified by a base measure μ and a kernel $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$. In this work, we take $\mathbb{X} = [-1,1]^d$ and assume K to be continuous, positive semi-definite. For the resulting process to be well defined, it is necessary and sufficient that the kernel K has eigenvalues in $[0,1]$ (Soshnikov, 2000, Theorem 3). When the eigenvalues further belong to $\{0,1\}$, we speak of (orthogonal) *projection* kernel and *projection* DPP. Projection DPPs have the practical feature of producing samples with fixed cardinality, μ almost surely, equal to the rank N of the kernel. More generally, they are the building blocks of DPPs. Indeed, under some general assumptions, all DPPs are mixtures of projection DPPs, (Hough et al., 2006). Hereafter, unless specifically stated, we consider projection DPPs.

One way to define a projection DPP with N points is to first take N orthonormal functions $\phi_0, \dots, \phi_{N-1}$ in $L^2(\mu)$, i.e., $\langle \phi_k, \phi_\ell \rangle_{L^2(\mu)} \triangleq \int \phi_k(x)\phi_\ell(x)\mu(\mathrm{d}x) = \delta_{k\ell}$. Then, consider the kernel K_N associated to the orthogonal projector onto $\mathcal{H}_N \triangleq \operatorname{span}\{\phi_k,\ 0 \leq k \leq N-1\}$, i.e.,

$$K_N(x,y) \triangleq \sum_{k=0}^{N-1} \phi_k(x)\phi_k(y). \tag{2}$$

The subset of $\{\mathbf{x}_1, \dots, \mathbf{x}_N\} \subset \mathbb{X}$ is said to be is drawn from the projection DPP with base measure μ and kernel K_N, denoted by $\{\mathbf{x}_1, \dots, \mathbf{x}_N\} \sim \operatorname{DPP}(\mu, K_N)$, when $(\mathbf{x}_1, \dots, \mathbf{x}_N)$ has joint distribution

$$\frac{1}{N!} \det[K_N(x_p, x_q)]_{p,q=1}^N \, \mu^{\otimes N}(\mathrm{d}x). \tag{3}$$

$\operatorname{DPP}(\mu, K_N)$ defines a probability measure over sets since (3) is invariant by permutation, the orthonormality of the ϕ_ks yields the normalization, see Appendix A.1 for details.

The repulsiveness of projection DPPs may be understood geometrically by rewriting (3) as

$$\prod_{n=1}^N \frac{1}{N-(n-1)} \left\| \Pi_{\mathcal{K}_{n-1}^\perp} K_N(x_n, \cdot) \right\|_{L^2(\mu)}^2 \mu(\mathrm{d}x_n), \tag{4}$$

where $\Pi_{\mathcal{K}_{n-1}^\perp}$ is the orthogonal projector onto the orthocomplement of $\operatorname{span}\{K_N(x_\ell, \cdot)\}_{\ell=1}^{n-1}$ in $\mathcal{H}_N$. Seeing (4) as a base×height formula, the joint distribution (3) is proportional to the squared volume of the parallelotope spanned by $K(x_1, \cdot), \dots, K(x_N, \cdot)$ in the *feature space* $\mathcal{H}_N$. Hence, the larger the volume, the more likely $x_1, \dots, x_N$ co-occur.

Moreover, using the same *normal* equations as in standard linear regression, the norms in (4) read

$$\begin{aligned} &\left\| \Pi_{\mathcal{H}_{n-1}^\perp} K_N(x_n, \cdot) \right\|_{L^2(\mu)}^2 \qquad (5) \\ &= \begin{cases} K_N(x_1, x_1), & \text{if } n = 1, \\ K_N(x_n, x_n) - \mathbf{K}_{n-1}(x_n)^\intercal \mathbf{K}_{n-1}^{-1} \mathbf{K}_{n-1}(x_n), & \text{else,} \end{cases} \end{aligned}$$

where $\mathbf{K}_{n-1}(\cdot) = (K_N(x_1, \cdot), \dots, K_N(x_{n-1}, \cdot))^\intercal$, and $\mathbf{K}_{n-1} = [K_N(x_p, x_q)]_{p,q=1}^{n-1}$.

The unnormalized conditionals densities (5) also shows up in Gaussian processes (GPs, Rasmussen & Williams, 2006) as the incremental posterior variances in a GP model with kernel K_N, giving yet another intuition for repulsiveness.

2.3. The multivariate Jacobi ensemble

We follow Bardenet & Hardy (2016) and consider eigenfunctions of the kernel in (2) to be the orthonormal polynomials w.r.t. μ. In dimension $d = 1$, the resulting projection DPP is 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. However, there is no natural order on multivariate monomials: an ordering $\mathfrak{b} : \mathbb{N}^d \to \mathbb{N}$ must be picked before we apply Gram-Schmidt to the monomials in $L^2(\mu)$. Bardenet & Hardy (2016, Section 2.1.3) consider multi-indices $k \triangleq (k^1, \dots, k^d) \in \mathbb{N}^d$ ordered by their maximum degree $\max_i k^i$, and for constant maximum degree, by the usual lexicographic order. We still denote the multivariate orthonormal polynomials by $(\phi_k)_{k \in \mathbb{N}^d}$.

By multivariate OPE we mean the projection DPP with base measure $\mu(\mathrm{d}x) \triangleq \omega(x)\,\mathrm{d}x$ and orthogonal projection kernel $K_N(x,y) \triangleq \sum_{\mathfrak{b}(k)=0}^{N-1} \phi_k(x)\phi_k(y)$. When the base measure is separable, i.e., $\omega(x) = \omega^1(x^1) \times \cdots \times \omega^d(x^d)$, multivariate orthonormal polynomials are products of univariate ones. and the kernel (2) reads

$$K_N(x,y) = \sum_{\mathfrak{b}(k)=0}^{N-1} \prod_{i=1}^{d} \phi^i_{k^i}(x^i)\phi^i_{k^i}(y^i), \tag{6}$$

with $(\phi^i_\ell)_{\ell\geq 0}$ the orthonormal polynomials w.r.t. $\omega^i(x^i)\,\mathrm{d}x^i$. For $\mathbb{X}=[-1,1]^d$ and $\omega^i(x^i) = (1-x^i)^{a_i}(1+x^i)^{b_i}$, with $a_i, b_i > -1$, the resulting DPP is called a *multivariate Jacobi ensemble*.

3. Monte Carlo with projection DPPs

3.1. A natural estimator

Bardenet & Hardy (2016) used

$$\widehat{I}_N^{\mathrm{BH}}(f) \triangleq \sum_{n=1}^{N} \frac{f(\mathbf{x}_n)}{K_N(\mathbf{x}_n,\mathbf{x}_n)}, \quad f\in L^1(\mu), \tag{7}$$

as an unbiased estimator of $\int f(x)\mu(\mathrm{d}x)$, with variance

$$\frac{1}{2}\int\left(\frac{f(x)}{K_N(x,x)} - \frac{f(y)}{K_N(y,y)}\right)^2 |K_N(x,y)|^2 \mu(\mathrm{d}x)\mu(\mathrm{d}y).$$

This variance clearly captures a notion of smoothness of f w.r.t. the kernel but its interpretation is not obvious.

For $\mathbb{X}=[-1,1]^d$, the interest in multivariate Jacobi ensemble among DPPs comes from the fact that (7) can be understood as a (randomized) multivariate counterpart of the Gauss-Jacobi quadrature in Section 2.1. Besides, for f essentially $\mathcal{C}^1$, Bardenet & Hardy (2016, Theorem 2.7) also proved a CLT with faster-than-classical-Monte-Carlo decay,

$$\sqrt{N^{1+1/d}}\left(\widehat{I}_N^{\mathrm{BH}}(f) - \int f(x)\mu(\mathrm{d}x)\right) \xrightarrow[N\to\infty]{\text{law}} \mathcal{N}(0,\Omega^2_{f,\omega}), \tag{8}$$

with $\Omega^2_{f,\omega} \triangleq \frac{1}{2}\sum_{k\in\mathbb{N}^d}(k^1+\cdots+k^d)\mathcal{F}_{\frac{f\omega}{\omega_{\mathrm{eq}}}}(k)^2$, where $\mathcal{F}_g$ denotes the Fourier transform of g, and $\omega_{\mathrm{eq}}(x) \triangleq 1/\prod_{i=1}^d \pi\sqrt{1-(x^i)^2}$. In the fast CLT (8), the asymptotic variance is governed by the smoothness of f since $\Omega_{f,\omega}$ is a measure of the decay of its Fourier coefficients.

3.2. The Ermakov-Zolotukhin estimator

We first state the main result of Ermakov & Zolotukhin (1960), see also Evans & Swartz (2000, Section 6.4.3) and references therein. Using modern arguments and notation, we can provide a short and simple proof this results, cf. Appendix A.2. It is based on a generalization of the Cauchy-Binet formula established by Johansson (2006), see also Appendix A.1. We apply the result of Ermakov & Zolotukhin (1960) to build an unbiased estimator of $\int f(x)\mu(\mathrm{d}x)$ which comes with a practical variance.

Theorem 1. *Let* $\{\mathbf{x}_1,\dots,\mathbf{x}_N\}\sim \mathrm{DPP}(\mu,K_N)$ *as in* (3). *Then, the solution of*

$$\begin{pmatrix} \phi_0(\mathbf{x}_1) & \dots & \phi_{N-1}(\mathbf{x}_1)\\ \vdots & & \vdots\\ \phi_0(\mathbf{x}_N) & \dots & \phi_{N-1}(\mathbf{x}_N)\end{pmatrix} \begin{pmatrix} y^1\\ \vdots\\ y^N\end{pmatrix} = \begin{pmatrix} f(\mathbf{x}_1)\\ \vdots\\ f(\mathbf{x}_N)\end{pmatrix} \tag{9}$$

is unique, μ-almost surely, with coordinates satisfying

$$\mathbb{E}\big[y^k\big] = \langle f,\phi_{k-1}\rangle_{L^2(\mu)} \tag{10}$$

$$\mathbb{V}\mathrm{ar}\big[y^k\big] = \|f\|_{L^2(\mu)} - \sum_{k=0}^{N-1}\langle f,\phi_{k-1}\rangle_{L^2(\mu)} \tag{11}$$

where $\mathbf{\Phi}$ denotes the feature matrix in (9) *and $\mathbf{\Phi}_{\phi_{k-1},f}$ is defined as the matrix obtained by replacing the k-th column of $\mathbf{\Phi}$ by $(f(\mathbf{x}_1),\dots,f(\mathbf{x}_N))^{\mathsf{T}}$.*

Several remarks are in order. The latter theorem shows that solving the (random) linear system (9), provides unbiased estimates of the N Fourier-like coefficients $(\langle f,\phi_k\rangle)_{k=0}^{N-1}$. Remarkably, these estimates have the exact same variance (11) equal to the residual $\sum_{k=N}^{\infty}\langle f,\phi_k\rangle^2$. The faster the decay of the coefficients, the smaller the variance. When $f\in\mathcal{H}_N$, these estimators have zero variance: f can be reconstructed perfectly from only one sample of $\mathrm{DPP}(\mu,K_N)$.

In the setting of multivariate Jacobi ensemble described in Section 2.3, the first orthonormal polynomial ϕ_0 is constant. Hence, a direct application of Theorem 1 yields

$$\widehat{I}_N^{\mathrm{EZ}}(f) \triangleq \mu\big([-1,1]^d\big)^{\frac{1}{2}}\,\frac{\det \mathbf{\Phi}_{\phi_0,f}(\mathbf{x}_{1:N})}{\det \mathbf{\Phi}(\mathbf{x}_{1:N})} \tag{12}$$

as an unbiased estimator of $\int f(x)\mu(\mathrm{d}x)$, which can be viewed as a quadrature rule, cf. Appendix A.3. Unlike the variance of $\widehat{I}_N^{\mathrm{BH}}(f)$ in (3.1), the variance of $\widehat{I}_N^{\mathrm{EZ}}(f)$ clearly reflects the accuracy of the approximation of f by its projection onto $\mathcal{H}_N$ In particular, it allows to integrate and interpolate polynomials up to "degree" $\mathfrak{b}^{-1}(N-1)$, perfectly. Nonetheless, its limiting theoretical properties, like a CLT, look hard to establish. In particular, the dependence of each quadrature weight on all quadrature nodes makes the estimator a peculiar object that does not fit the assumptions of traditional CLTs for DPPs (Soshnikov, 2000).

3.3. Sampling

To perform Monte Carlo integration with DPPs, it is crucial to sample the points and evaluate the weights efficiently. Except for some specific instances, exact sampling from continuous projection DPPs requires the spectral decomposition of the kernel (2) before applying the chain rule (4) (Hough et al., 2006). The main challenge is to find good proposal distributions to efficiently sample the successive conditionals (Lavancier et al., 2012).

We focus on sampling the multivariate Jacobi ensemble, with parameters $a^i, b^i \in [-\frac{1}{2}, \frac{1}{2}]$. In dimension $d = 1$, it can be sampled at cost $\mathcal{O}(N^2)$ by computing the eigenvalues of a random tridiagonal matrix (Killip & Nenciu, 2004, Theorem 2, $\beta = 2$). For $d \geq 2$, we follow Bardenet & Hardy (2016) and use the same proposal distribution $\omega_{\text{eq}}(x)\,\mathrm{d}x$ and rejection bound to sample each conditional. The rejection constant is derived from a result of (Chow et al., 1994) on Jacobi polynomials. See Appendix A.4 for more details.

We remodeled the original implementation[1] of Bardenet & Hardy (2016) in a more Python*ic* way. Notably, when evaluating the kernel (6), we paid special attention to avoiding unnecessary evaluations of the univariate orthogonal Jacobi polynomials by propagation of three-term recurrence relations they satisfy. Comparatively, sampling $N = 100$ points in dimension $d = 1, 2$ was counted in minutes, now it takes milliseconds. In Appendix A.4, we display a 2D sample of size $N = 1000$, obtained in approximately 7 min compared to hours previously, on a modern laptop.

4. Empirical investigation

Appendix B collects the results of the following experiments as well as further experiments on non smooth functions.

4.1. The bump experiment

Bardenet & Hardy (2016, Section 3) illustrate the behavior of $\widehat{I}_N^{\text{BH}}$ and its CLT (8) on a unimodal, smooth *bump* function ($\varepsilon = 0.05$). The expected variance decay is of order $1/N^{1+1/d}$. We successfully reproduce their experiment in Figure 1 for larger N, and compare with the behavior of $\widehat{I}_N^{\text{EZ}}$. In short, $\widehat{I}_N^{\text{EZ}}$ dramatically outperforms $\widehat{I}_N^{\text{BH}}$ in $d \leq 2$, with surprisingly fast empirical convergence rates. When $d \geq 3$, performance decreases, and $\widehat{I}_N^{\text{BH}}$ shows both faster and more regular variance decay.

As to whether a CLT for $\widehat{I}_N^{\text{EZ}}$ could hold, we performed Kolmogorov-Smirnov tests for $N = 300$, see Appendix B.1. This yielded small p-values across dimensions, from 0.03 to 0.24. This is compared to the same p-values for $\widehat{I}_N^{\text{BH}}$, which range from 0.60 to 0.99. The lack of normality of the EZ estimator is partly due to a few outliers. Where these outliers come from is left for future work; ill-conditioning of the linear system (9) is an obvious candidate.

4.2. Integrating sums of eigenfunctions

To test the variance decay of $\widehat{I}_N^{\text{EZ}}(f)$ prescribed by Theorem 1, we consider functions of the form

$$f(x) = \sum_{\mathfrak{b}(k)=0}^{N_{\text{modes}}-1} \frac{1}{\mathfrak{b}(k)+1} \phi_k(x). \tag{13}$$

That is to say, the function f can be either fully ($N_{\text{modes}} \leq N$) or partially ($N_{\text{modes}} > N$) decomposed in the eigenbasis of the kernel. In both cases, we let N vary from 10 to 100 and the dimension d from 1 to 4.

In the first setting, we set $N_{\text{modes}} = 70$. Thus, N eventually reaches the number of functions used to build f in (13), after what $\widehat{I}_N^{\text{EZ}}$ is an exact estimator in any dimension, see Figure 1. The second setting has $N_{\text{modes}} = N + 1$, so that the number of points N is never enough for the variance (11) to be zero. The corresponding $1/N^2$ variance decay prescribed by Theorem 1 can be observed in Appendix B.2.

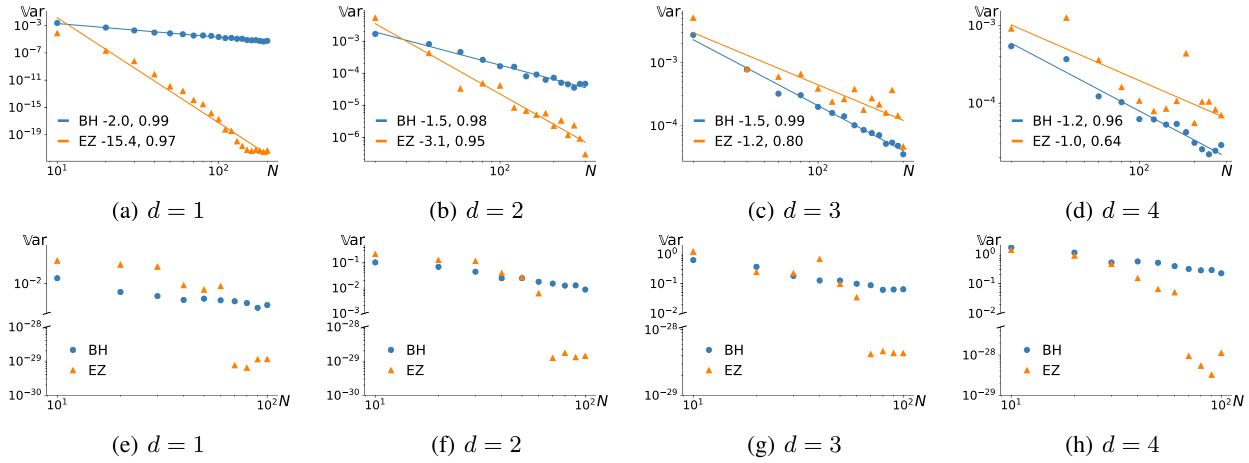
5. Conclusion

Ermakov & Zolotukhin (EZ, 1960) proposed a non-obvious unbiased Monte Carlo estimator using projection DPPs. It requires solving a linear system, which in turn involves evaluating both the N eigenfunctions of the corresponding kernel and the integrand at the N points of the DPP sample. This is yet another connection between DPPs and linear algebra. In fact, solving this linear system provides unbiased estimates of the Fourier-like coefficients of the integrand f with each of the N eigenfunctions of the DPP kernel. Remarkably, these estimators have identical variance measuring the accuracy of the approximation of f by its projection onto these eigenfunctions. With modern arguments, we have provided a much shorter proof of these properties than in the original work of (Ermakov & Zolotukhin, 1960). Beyond this, little is known on the EZ estimator. While coming with a less interpretable variance, the more direct estimator proposed by Bardenet & Hardy (BH, 2016) has an intrinsic connection with the classical Gauss quadrature and further enjoys stronger theoretical properties when using multivariate Jacobi ensemble.

Our experiments highlight the key features of both estimators when the underlying DPP is a multivariate Jacobi ensemble, and further demonstrate the known properties of the BH estimator in yet unexplored regimes. Although EZ shows a *surprisingly fast* empirical convergence rate for $d \leq 2$, its behavior is more erratic for $d \geq 3$. Ill-conditioning of the linear system is a potential source of outliers in the distribution of the estimator. Regularization may help but would introduce a stability/bias trade-off. More generally, EZ seems worth investigating for integration or even interpolation tasks where the function is known to be decomposable in the eigenbasis of the kernel, i.e., in a setting similar to the one of Bach (2017). Finally, the new implementation of an exact sampler for multivariate Jacobi ensemble unlocks more large-scale empirical investigations and asks for more theoretical work. The associated code is available in the DPPy toolbox of Gautier et al. (2018).

[1] https://github.com/rbardenet/dppmc

https://github.com/guilgautier/DPPy/notebooks



(a) $d = 1$ (b) $d = 2$ (c) $d = 3$ (d) $d = 4$

(e) $d = 1$ (f) $d = 2$ (g) $d = 3$ (h) $d = 4$

Figure 1. (a)-(d) cf. Section 4.1 the numbers in the legend are the slope and R^2 (e)-(h) cf. Section 4.2.

Acknowledgments

Many thanks to Mathieu Gerber of Univ. Bristol, UK for digging up this result of Ermakov & Zolotukhin (1960) from his human memory. We acknowledge funding by European CHIST-ERA project DELTA, the French Ministry of Higher Education and Research, the Nord-Pas-de-Calais Regional Council, Inria and Otto-von-Guericke-Universität Magdeburg associated-team north-European project Allocate, and French National Research Agency projects ExTraLearn (n.ANR-14-CE24-0010-01) and BoB (n.ANR-16-CE23-0003).

References

Bach, F. On the Equivalence between Kernel Quadrature Rules and Random Feature Expansions. *Journal of Machine Learning Research*, 18(21):1–38, 2017. arXiv:1502.06800.

Bach, F., Lacoste-Julien, S., and Obozinski, G. On the Equivalence between Herding and Conditional Gradient Algorithms. In *International Conference on Machine Learning (ICML)*, 2012. arXiv:1203.4523.

Bardenet, R. and Hardy, A. Monte Carlo with Determinantal Point Processes. *ArXiv e-prints*, 2016. arXiv:1605.00361.

Briol, F.-X., Oates, C. J., Girolami, M., and Osborne, M. A. Frank-Wolfe Bayesian Quadrature: Probabilistic Integration with Theoretical Guarantees. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1162–1170, jun 2015. arXiv:1506.02681.

Chen, Y., Welling, M., and Smola, A. Super-Samples from Kernel Herding. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 109–116. AUAI Press, mar 2010. arXiv:1203.3472. ISBN 9780974903965.

Chow, Y., Gatteschi, L., and Wong, R. A Bernstein-type inequality for the Jacobi polynomial. *Proceedings of the American Mathematical Society*, 121(3):703–703, 1994.

Davis, P. J. and Rabinowitz, P. Methods of numerical integration. Academic Press. 1984. ISBN 9780122063602.

Delyon, B. and Portier, F. Integral approximation by kernel smoothing. *Bernoulli*, 22(4):2177–2208, nov 2016. arXiv:1409.0733.

Derezínski, M. Fast determinantal point processes via distortion-free intermediate sampling. *ArXiv e-prints*, 2019. arXiv:1811.03717v2.

Dick, J. and Pillichshammer, F. Digital nets and sequences : discrepancy and quasi-Monte Carlo integration. Cambridge University Press. 2010. ISBN 9780521191593.

Ermakov, S. M. and Zolotukhin, V. G. Polynomial Approximations and the Monte-Carlo Method. *Theory of Probability & Its Applications*, 5(4):428–431, jan 1960.

Evans, M. and Swartz, T. Approximating integrals via Monte Carlo and deterministic methods. Oxford University Press. 2000. ISBN 9780198502784.

Gautier, G., Bardenet, R., and Valko, M. DPPy: Sampling Determinantal Point Processes with Python. *ArXiv e-prints*, 2018. arXiv:1809.07258.

Gautschi, W. How sharp is Bernstein's Inequality for Jacobi polynomials? *Electronic Transactions on Numerical Analysis*, 36:1–8, 2009.

Hough, J. B., Krishnapur, M., Peres, Y., and Virág, B. Determinantal Processes and Independence. In *Probability Surveys*, volume 3, pp. 206–229. The Institute of Mathematical Statistics and the Bernoulli Society, 2006. arXiv:math/0503110.

Huszár, F. and Duvenaud, D. Optimally-Weighted Herding is Bayesian Quadrature. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 377–386. AUAI Press, apr 2012. arXiv:1204.1664. ISBN 9780974903989.

Johansson, K. Random matrices and determinantal processes. *Les Houches Summer School Proceedings*, 83(C): 1–56, 2006.

Killip, R. and Nenciu, I. Matrix models for circular ensembles. *International Mathematics Research Notices*, 2004 (50):2665, 2004. arXiv:math/0410034.

König, W. Orthogonal polynomial ensembles in probability theory. *Probab. Surveys*, 2:385–447, 2004. arXiv:math/0403090.

Kulesza, A. and Taskar, B. Determinantal Point Processes for Machine Learning. *Foundations and Trends in Machine Learning*, 5(2-3):123–286, 2012. arXiv:1207.6083.

Launay, C., Galerne, B., and Desolneux, A. Exact Sampling of Determinantal Point Processes without Eigendecomposition. *ArXiv e-prints*, feb 2018. arXiv:1802.08429.

Lavancier, F., Møller, J., and Rubak, E. Determinantal point process models and statistical inference : Extended version. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 77(4):853–877, may 2012. arXiv:1205.4818.

Liu, Q. and Lee, J. D. Black-Box Importance Sampling. In *Internation Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017. arXiv:1610.05247.

Macchi, O. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7(01):83–122, mar 1975.

Oates, C. J., Girolami, M., and Chopin, N. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79 (3):695–718, jun 2017. arXiv:1410.2392.

O'Hagan, A. Bayes–Hermite quadrature. *Journal of Statistical Planning and Inference*, 29(3):245–260, nov 1991.

Poulson, J. High-performance sampling of generic Determinantal Point Processes. *ArXiv e-prints*, apr 2019. arXiv:1905.00165.

Rasmussen, C. E. and Williams, C. K. I. Gaussian processes for machine learning. MIT Press. 2006. ISBN 026218253X.

Robert, C. P. The Bayesian choice : from decision-theoretic foundations to computational implementation. Springer. 2007. ISBN 9780387952314.

Robert, C. P. and Casella, G. Monte Carlo statistical methods. Springer-Verlag New York. 2004. ISBN 9781441919397.

Soshnikov, A. Determinantal random point fields. *Russian Mathematical Surveys*, 55(5):923–975, feb 2000. arXiv:math/0002099.

A. Methodology

A.1. The generalized Cauchy-Binet formula: the modern argument

Johansson (2006, Section 2.2) developed a natural way to build DPPs associated to projection (potentially non hermitian) kernels. In this part, we focus on the generalization of the Cauchy-Binet formula (Johansson, 2006, Proposition 2.10). Its usefulness is twofold for our purpose. First, it serves to justify the fact that the normalization constant of the joint distribution (3) is one, i.e., it is indeed a probability distribution. Second, we use it as a modern and simple argument to prove the result of Ermakov & Zolotukhin (1960), cf. Theorem 1. An extended version of the proof is given in Appendix A.2.

Lemma A. *(Johansson, 2006, Proposition 2.10) Let $(\mathbb{X}, \mathcal{B}, \mu)$ be a measurable space and consider measurable functions $\phi_0, \dots, \phi_{N-1}$ and $\psi_0, \dots, \psi_{N-1}$, such that $\phi_k \psi_\ell \in L^1(\mu)$. Then,*

$$\det\Big(\langle \phi_k, \psi_\ell \rangle_{L^2(\mu)}\Big)_{k,\ell=1}^{N} = \frac{1}{N!} \int \det \mathbf{\Phi}(x_{1:N}) \det \mathbf{\Psi}(x_{1:N})\, \mu^{\otimes N}(\mathrm{d}x), \tag{A.1}$$

where

$$\mathbf{\Phi}(x_{1:N}) = \begin{pmatrix} \phi_0(x_1) & \dots & \phi_{N-1}(x_1) \\ \vdots & & \vdots \\ \phi_0(x_N) & \dots & \phi_{N-1}(x_N) \end{pmatrix} \quad \textit{and} \quad \mathbf{\Psi}(x_{1:N}) = \begin{pmatrix} \psi_0(x_1) & \dots & \psi_{N-1}(x_1) \\ \vdots & & \vdots \\ \psi_0(x_N) & \dots & \psi_{N-1}(x_N) \end{pmatrix}$$

A.2. Proof of Theorem 1

First, we recall the result of Ermakov & Zolotukhin (1960), cf. Theorem 1. Then, we provide a modern proof based on the generalization of the Cauchy-Binet formula, cf. Lemma A, where we exploit the orthonormality of the eigenfunctions of the kernel.

Theorem B. *Consider $f \in L^2(\mu)$ together with N orthonormal functions $\phi_0, \dots, \phi_{N-1} \in L^2(\mu)$:*

$$\langle \phi_k, \phi_\ell \rangle_{L^2(\mu)} \triangleq \int \phi_k(x)\phi_\ell(x)\mu(\mathrm{d}x) = \delta_{k\ell}, \quad \forall 0 \leq k, \ell \leq N-1. \tag{A.2}$$

Let $\{\mathbf{x}_1, \dots, \mathbf{x}_N\} \sim \mathrm{DPP}(\mu, K_N)$ with $K_N(x,y) = \sum_{k=0}^{N-1} \phi(x)\phi(y)$. That is to say $(\mathbf{x}_1, \dots, \mathbf{x}_N)$ has joint distribution

$$\frac{1}{N!} \det[K_N(x_p, x_q)]_{p,q=1}^{N}\, \mu^{\otimes N}(\mathrm{d}x). \tag{A.3}$$

Then, the solution of

$$\begin{pmatrix} \phi_0(\mathbf{x}_1) & \dots & \phi_{N-1}(\mathbf{x}_1) \\ \vdots & & \vdots \\ \phi_0(\mathbf{x}_N) & \dots & \phi_{N-1}(\mathbf{x}_N) \end{pmatrix} \begin{pmatrix} y^1 \\ \vdots \\ y^N \end{pmatrix} = \begin{pmatrix} f(\mathbf{x}_1) \\ \vdots \\ f(\mathbf{x}_N) \end{pmatrix} \tag{A.4}$$

is unique, μ-almost surely and the coordinates of the solution vector, namely

$$y^k = \frac{\det \mathbf{\Phi}_{\phi_{k-1},f}(\mathbf{x}_{1:N})}{\det \mathbf{\Phi}(\mathbf{x}_{1:N})}, \tag{A.5}$$

satisfy

$$\mathbb{E}\big[y^k\big] = \langle f, \phi_{k-1} \rangle_{L^2(\mu)}, \quad \textit{and} \quad \mathbb{V}\mathrm{ar}\big[y^k\big] = \|f\|_{L^2(\mu)}^2 - \sum_{\ell=0}^{N-1} \langle f, \phi_\ell \rangle_{L^2(\mu)}^2, \tag{A.6}$$

where $\mathbf{\Phi}(\mathbf{x}_{1:N})$ denotes the feature matrix in (A.4) and $\mathbf{\Phi}_{\phi_{k-1},f}(\mathbf{x}_{1:N})$ is defined as the matrix obtained by replacing the k-th column of $\mathbf{\Phi}(\mathbf{x}_{1:N})$ by $(f(\mathbf{x}_1), \dots, f(\mathbf{x}_N))^\mathsf{T}$.

Proof of Theorem B. First, the joint distribution (A.3) of $(\mathbf{x}_1, \dots, \mathbf{x}_N)$ is proportional to $(\det \mathbf{\Phi}(\mathbf{x}_{1:N}))^2 \mu^{\otimes N}(x)$. Thus, $\det \mathbf{\Phi}(x_{1:N}) \neq 0$, μ-almost surely. Hence, the matrix $\mathbf{\Phi}(\mathbf{x}_{1:N})$ defining the linear system (A.4) is invertible, μ-almost surely.

The expression of the coordinates (A.5) follows from Cramer's rule.

Then, we treat the case $k = 1$, the others follow the same lines. The proof relies on Lemma A where we exploit the orthonormality of the ϕ_ks (A.2). The expectation in (A.6) reads

$$\begin{aligned}
\mathbb{E}\left[\frac{\det \mathbf{\Phi}_{\phi_0,f}(\mathbf{x}_{1:N})}{\det \mathbf{\Phi}(\mathbf{x}_{1:N})}\right] &\overset{(A.3)}{=} \frac{1}{N!}\int \det \mathbf{\Phi}_{\phi_0,f}(x_{1:N}) \det \mathbf{\Phi}(x_{1:N})\, \mu^{\otimes N}(\mathrm{d}x)\\
&\overset{(A.1)}{=} \det \begin{pmatrix} \langle f, \phi_0\rangle^2_{L^2(\mu)} & \left(\langle f, \phi_\ell\rangle^2_{L^2(\mu)}\right)_{\ell=1}^{N-1} \\ \left(\langle f, \phi_0\rangle^2_{L^2(\mu)}\right)_{k=1}^{N-1} & \left(\langle \phi_k, \phi_\ell\rangle^2_{L^2(\mu)}\right)_{k,\ell=1}^{N-1} \end{pmatrix}\\
&\overset{(A.2)}{=} \det \begin{pmatrix} \langle f, \phi_0\rangle^2_{L^2(\mu)} & \left(\langle f, \phi_\ell\rangle^2_{L^2(\mu)}\right)_{\ell=1}^{N-1} \\ 0_{N-1,1} & I_{N-1} \end{pmatrix}\\
&= \langle f, \phi_0\rangle^2_{L^2(\mu)}.
\end{aligned} \tag{A.7}$$

Similarly, the second moment reads

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\det \mathbf{\Phi}_{\phi_0,f}(\mathbf{x}_{1:N})}{\det \mathbf{\Phi}(\mathbf{x}_{1:N})}\right)^2\right] &\overset{(A.3)}{=} \frac{1}{N!}\int \det \mathbf{\Phi}_{\phi_0,f}(x_{1:N}) \det \mathbf{\Phi}_{\phi_0,f}(x_{1:N})\, \mu^{\otimes N}(\mathrm{d}x)\\
&\overset{(A.1)}{=} \det \begin{pmatrix} \langle f, f\rangle^2_{L^2(\mu)} & \left(\langle f, \phi_\ell\rangle^2_{L^2(\mu)}\right)_{\ell=1}^{N-1} \\ \left(\langle f, \phi_k\rangle^2_{L^2(\mu)}\right)_{k=1}^{N-1} & \left(\langle \phi_k, \phi_\ell\rangle^2_{L^2(\mu)}\right)_{k,\ell=1}^{N-1} \end{pmatrix}\\
&\overset{(A.2)}{=} \det \begin{pmatrix} \|f\|^2_{L^2(\mu)} & \left(\langle f, \phi_\ell\rangle^2_{L^2(\mu)}\right)_{\ell=1}^{N-1} \\ \left(\langle f, \phi_k\rangle^2_{L^2(\mu)}\right)_{k=1}^{N-1} & I_{N-1} \end{pmatrix}\\
&= \|f\|^2_{L^2(\mu)} - \sum_{k=1}^{N-1} \langle f, \phi_k\rangle^2_{L^2(\mu)}.
\end{aligned} \tag{A.8}$$

Finally, the variance in (A.6) = (A.8) - (A.7)^2. □

A.3. EZ estimator as a quadrature rule

In this part, we consider Theorem B in the setting where one of the eigenfunctions of the kernel, say ϕ_0 is constant. In this case, we show that the EZ estimator defined to estimate $\int f(x)\mu(\mathrm{d}x)$ can be seen as a quadrature rule in the sense of (1), with weights ω_n that sum to $\mu\big([-1,1]^d\big)$. This is a non obvious fact, judging from the expression (12) of the estimator.

Proposition 1. *Consider ϕ_0 constant in Theorem B. Then, solving the corresponding linear system* (A.4) *allows to construct*

$$\widehat{I}_N^{EZ}(f) \overset{(12)}{=} \sqrt{\mu\Big([-1,1]^d\Big)}\ \frac{\det \mathbf{\Phi}_{\phi_0,f}(\mathbf{x}_{1:N})}{\det \mathbf{\Phi}(\mathbf{x}_{1:N})} \tag{A.9}$$

as an unbiased estimator of $\int f(x)\mu(\mathrm{d}x)$, with variance equal to the variance in (A.6)$\times\mu\Big([-1,1]^d\Big)$*. In particular it can be seen as a random quadrature rule* (1)*,*

$$\widehat{I}_N^{EZ}(f) = \sum_{n=1}^{N} \omega_n(\mathbf{x}_{1:N}) f(\mathbf{x}_n) \approx \int f(x)\mu(\mathrm{d}x) \tag{A.10}$$

such that $\sum_{n=1}^N \omega_n(\mathbf{x}_{1:N}) = \mu\Big([-1,1]^d\Big)$.

Proof. Take ϕ_0 constant. Since ϕ_0 has unit norm, cf. (A.2), it is straightforward to see that

$$\phi_0 = \frac{1}{\sqrt{\mu\Big([-1,1]^d\Big)}}$$

so that (A.9) can be written

$$\widehat{I}_N^{\mathrm{EZ}}(f) = \mu\Big([-1,1]^d\Big)\ \frac{\det \boldsymbol{\Phi}_{\phi_0,f}(\mathbf{x}_{1:N})}{\det \boldsymbol{\Phi}_{\phi_0,1}(\mathbf{x}_{1:N})}.$$

Expanding the numerator w.r.t. the first column yields

$$\widehat{I}_N^{\mathrm{EZ}}(f) = \sum_{n=1}^{N} f(\mathbf{x}_n) \underbrace{(-1)^{1+n} \det(\phi_k(x_p))_{k=1,p=1\neq n}^{N-1,N} \frac{\mu\Big([-1,1]^d\Big)}{\det \boldsymbol{\Phi}_{\phi_0,1}(\mathbf{x}_{1:N})}}_{\triangleq \omega_n(\mathbf{x}_{1:N})}.$$

In particular, there is a priori no reason for the weights to be nonnegative. Finally,

$$\sum_{n=1}^{N} \omega_n(\mathbf{x}_{1:N}) = \frac{\mu\Big([-1,1]^d\Big)}{\cancel{\det \boldsymbol{\Phi}_{\phi_0,1}(\mathbf{x}_{1:N})}} \underbrace{\sum_{n=1}^{N} (-1)^{1+n} \det(\phi_k(x_p))_{k=1,p=1\neq n}^{N-1,N}}_{=\cancel{\det \boldsymbol{\Phi}_{\phi_0,1}(\mathbf{x}_{1:N})}} = \mu\Big([-1,1]^d\Big).$$

This concludes the proof. □

A.4. Sampling multivariate Jacobi ensembles

In this part, we review briefly the main techniques for DPP sampling before we develop our method to generate samples from the multivariate Jacobi ensemble, as defined in Section 2.3. As an illustration, Figure A.1 displays a sample of a two dimensional Jacobi ensemble with $N = 1000$ points where the parameters a_1, b_1, a_2, b_2 were drawn i.i.d. uniformly on $[-1/2, 1/2]$.

In both finite and continuous cases, except for some specific instances, exact sampling from DPPs (with symmetric kernel) usually requires the spectral decomposition of the kernel before applying the chain rule (4), see, e.g., Hough et al. (2006); Kulesza & Taskar (2012). In the finite case, i.e., $\mathbb{X} = \{1, \dots, M\}$, sampling projection DPPs *does not require* the eigendecomposition of the kernel, and the chain rule costs $\mathcal{O}(MN^2)$, where N denotes the rank of kernel. Otherwise, there is a preprocessing cost of order $\mathcal{O}(M^3)$ which may become impractical for large M, just like other kernel methods. The same cubic cost applies to Cholesky-based samplers, see, e.g., Launay et al. (2018), or Poulson (2019) who can also treat non symmetric kernels. Note that, this cubic cost can be reduced when the kernel is given in a factored form (Kulesza & Taskar, 2012; Derezínski, 2019).

Unlike the discrete case, sampling from continuous DPPs, even projection ones remains challenging. The realizations of projection DPPs are usually generated by applying the chain rule (4), where the conditionals are sampled using rejection sampling. The main challenge is to find good proposal distributions to efficiently sample the successive conditionals (Lavancier et al., 2012). In this work, we take $\mathbb{X} = [-1,1]^d$ and focus on sampling the multivariate Jacobi ensemble, cf. 2.3, for a base measure with parameters $a^i, b^i \in [-\frac{1}{2}, \frac{1}{2}]$.

In dimension $d = 1$, to sample the univariate Jacobi ensemble, with base measure $\mu(\mathrm{d}x) = (1-x)^a (1+x)^b \,\mathrm{d}x$ where $a, b > -1$, we use the random tridiagonal matrix model of Killip & Nenciu (2004, Theorem 2). That is to say, computing the eigenvalues of a properly randomized tridiagonal matrix allows to get a sample of this continuous projection DPP at cost $\mathcal{O}(N^2)$!

For $d \geq 2$, we follow Bardenet & Hardy (2016, Section 3) who proposed to use the chain rule (4) to sample from the multivariate Jacobi ensemble with base measure $\mu(\mathrm{d}x) = \omega(x)\,\mathrm{d}x$, where

$$\omega(x) = \prod_{i=1}^{d} (1-x^i)^{a^i} (1-x^i)^{b^i}, \text{with } \left|a^i\right|, \left|b^i\right| \leq \frac{1}{2}. \tag{A.11}$$

To that end, we use the same proposal distribution and rejection bound to sample from each of the conditionals (4). The density (w.r.t. Lebesgue) of the proposal distribution writes

$$\omega_{\mathrm{eq}}(x) = \prod_{i=1}^{d} \frac{1}{\pi\sqrt{1-(x^i)^2}}. \tag{A.12}$$

The rejection constant is derived after by successive applications of the following result on Jacobi polynomials derived by Chow et al. (1994).

Proposition 2. *(Gautschi, 2009, Equation 1.3) Let $(\phi_k)_{k\geq 0}$ be the orthonormal polynomials w.r.t. the measure $(1-x)^a(1+x)^b \,\mathrm{d}x$ with $|a| \leq \frac{1}{2}, |b| \leq \frac{1}{2}$. Then, for any $x \in [-1,1]$ and $k \geq 0$,*

$$\pi(1-x)^{a+\frac{1}{2}}(1+x)^{b+\frac{1}{2}}\phi_k(x)^2 \leq \frac{2\,\Gamma(k+a+b+1)\,\Gamma(k+\max(a,b)+1)}{k!\,(k+\frac{a+b+1}{2})^{2\max(a,b)}\,\Gamma(k+\min(a,b)+1)}. \tag{A.13}$$

The domination of the acceptance ratio, i.e., the ratio of the n-th conditional density in (4) over the proposal density (A.12) is computed as follows

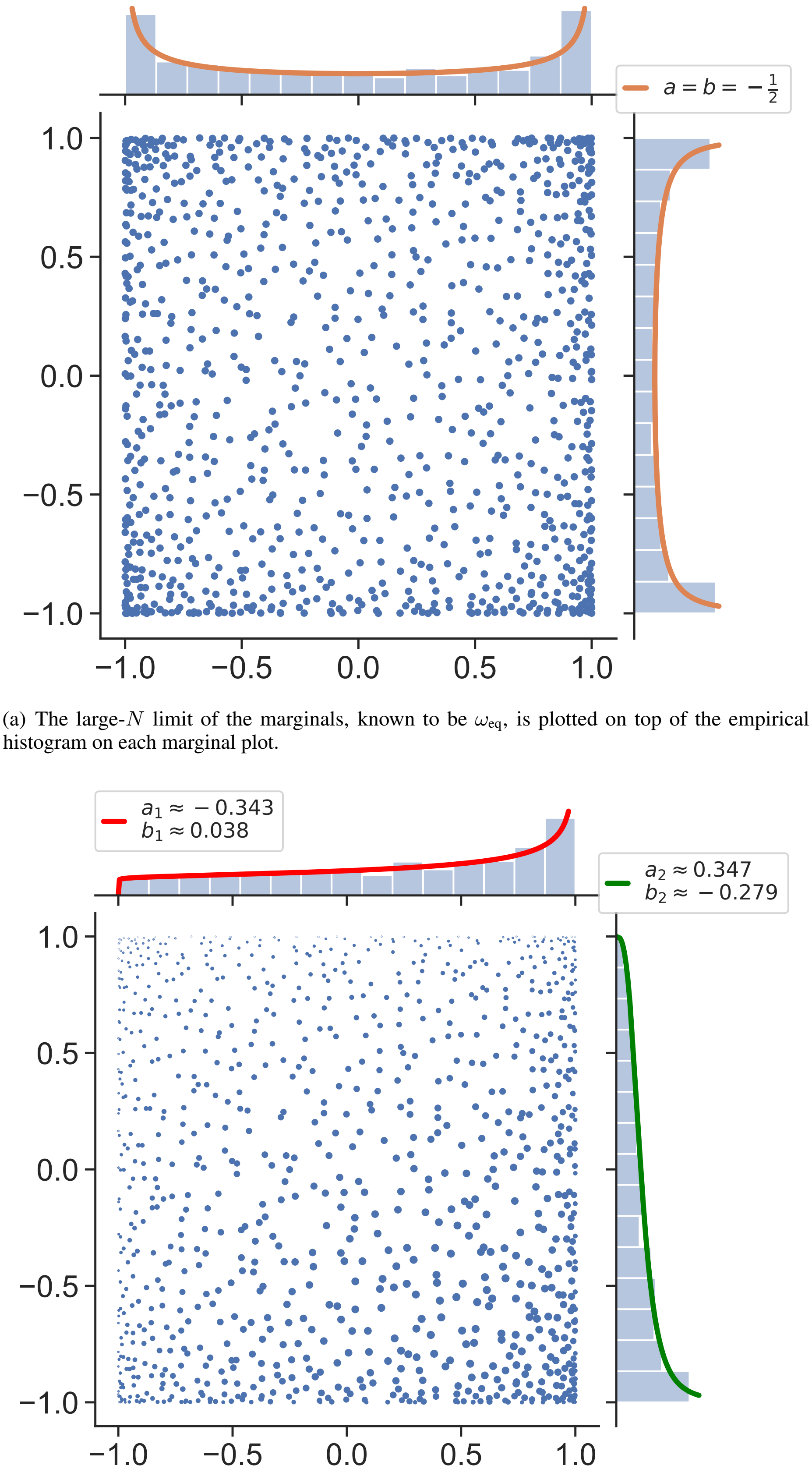
$$\begin{aligned} &\frac{K_N(x,x) - \mathbf{K}_{n-1}(x)^{\mathsf{T}}\mathbf{K}_{n-1}^{-1}\mathbf{K}_{n-1}(x)}{N-(n-1)}\omega(x) \times \frac{1}{\omega_{\mathrm{eq}}(x)} \\ &\leq \frac{1}{N-(n-1)}\frac{K(x,x)\omega(x)}{w_{\mathrm{eq}}(x)} \underset{(\mathrm{A.12})}{\overset{(6)}{=}} \frac{1}{N-(n-1)}\sum_{\mathfrak{b}(k)=0}^{N-1}\prod_{i=1}^{d}\pi(1-x^i)^{a^i+\frac{1}{2}}(1+x^i)^{b^i+\frac{1}{2}}\phi_{k^i}^i(x^i)^2. \end{aligned} \tag{A.14}$$

Finally, each of the terms that appear in (A.14) can be bounded using the following recipe:

1. For $k^i > 0$, we use the bound (A.13)
2. For $k^i = 0$, the domination of the left hand side (LHS) of (A.13) is not tight enough ($= 2$), so we proceed as follows. In this case, ϕ_0 is constant equal to $\left(\int (1-x)^a(1+x)^b \,\mathrm{d}x\right)^{-1/2}$ and since $|a|, |b| \in [-1/2, 1/2]$ we upper bound $(1-x)^{a+1/2}(1+x)^{b+1/2}$ by the evaluation at its mode.

Point 2. is crucial to tighten the rejection constant. Indeed, because of the choice of the ordering $\mathfrak{b}$ (cf. Section 2.3), the number of times that ϕ_0^i appears in (A.14) increases with the dimension. Hence, the tighter the bound on the LHS of (A.13) for $k = 0$ the best the rejection constant.

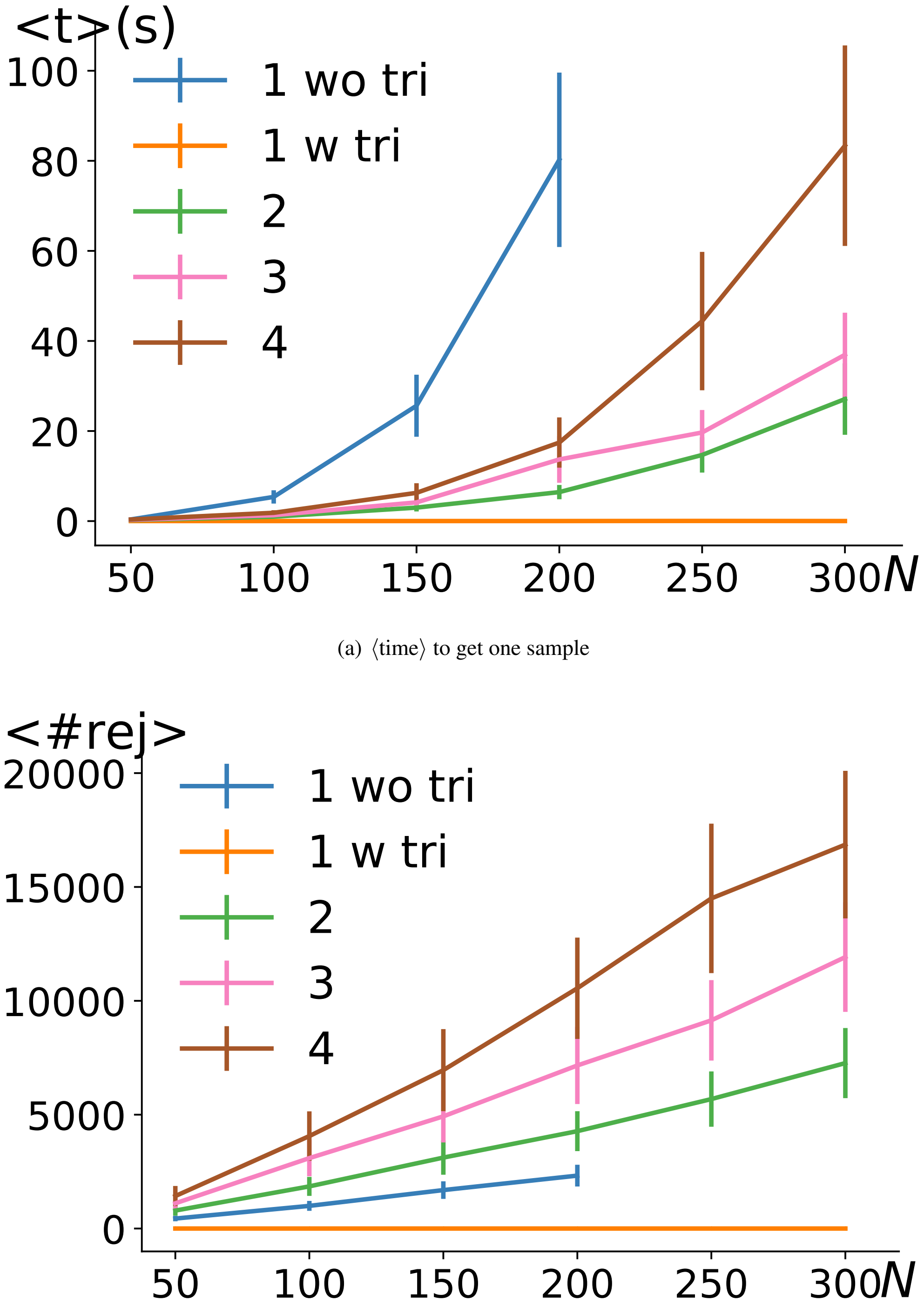
In Figure A.2 we illustrate the following observations. We note that computing the acceptance ratio requires to propagate these recurrence relations up to order $\sqrt[d]{N}$. Thus, for a given N, the larger the dimension, the smaller the depth of the recurrence. This could hint that, evaluating the kernel (6) becomes cheaper as d increases. However, the rejection rate also increases, so that in practice, it is not cheaper to sample in larger dimensions because the number of rejections dominates. In the particular case of dimension $d = 1$, samples are generated using the fast and rejection-free tridiagonal matrix model of Killip & Nenciu (2004, Theorem 2). This grants huge time savings compared to the acceptance-rejection method. Without it, sampling N points in dimension $d = 1$ would take more time than in larger dimension, although the associated rejection constants are smaller, as it can be seen in Figure 2(a) and Figure 2(b).



(a) The large-N limit of the marginals, known to be ω_{eq}, is plotted on top of the empirical histogram on each marginal plot.

(b) Same sample as in Figure 1(a) but the disk centered at $\mathbf{x}_n$ has an area proportional to the weight $1/K_N(\mathbf{x}_n, \mathbf{x}_n)$ as in $\widehat{I}_N^{\text{BH}}$ (7).

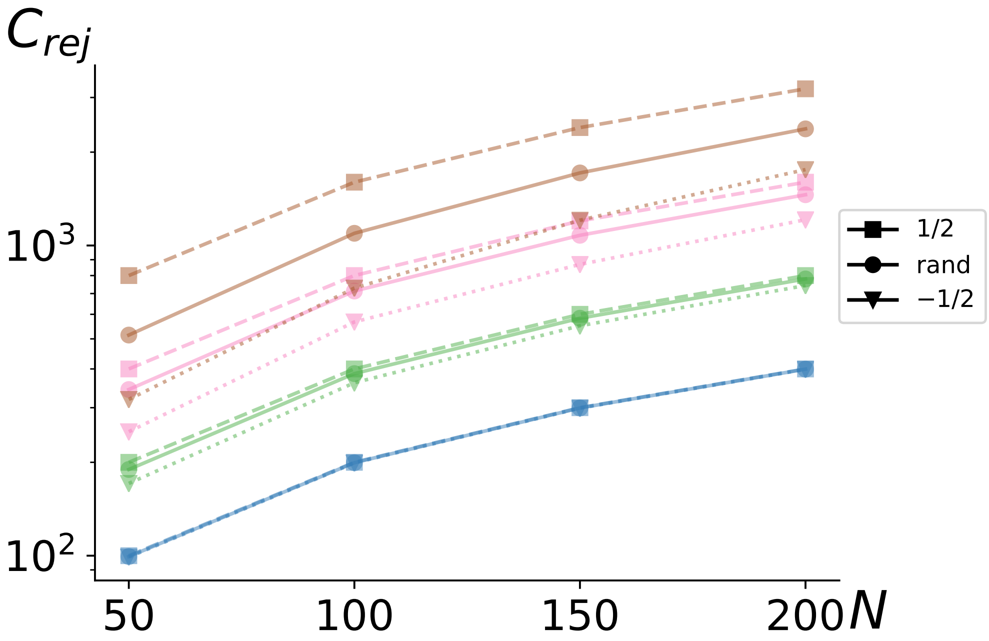
Figure A.1. A sample of a 2D Jacobi ensemble with $N = 1000$ points and parameters a_1, b_1, a_2, b_2 drawn i.i.d. uniformly on $[-1/2, 1/2]$.



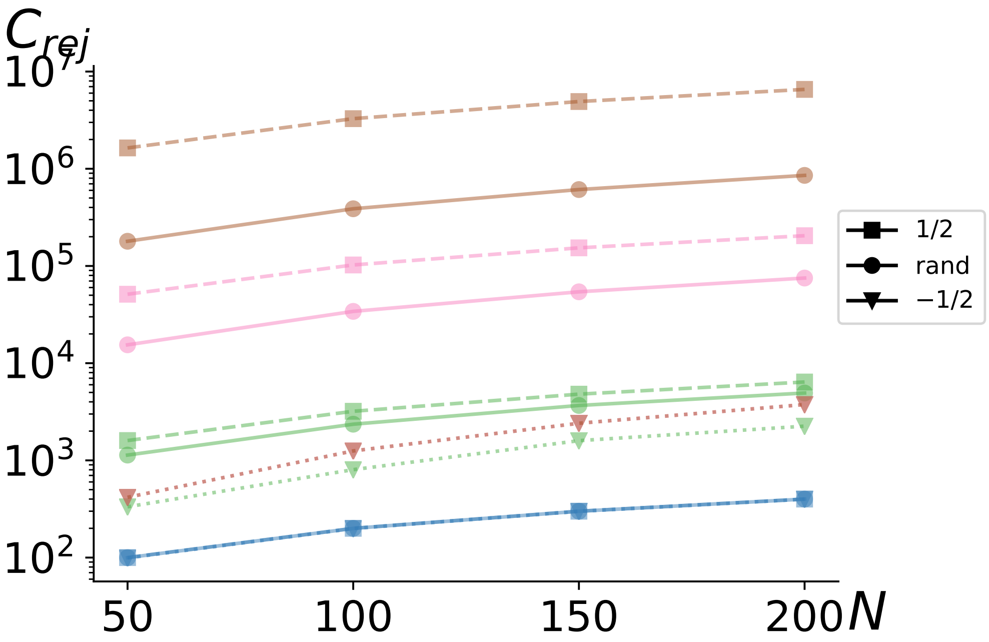
(a) ⟨time⟩ to get one sample

(b) ⟨#rejections⟩

Figure A.2. The colors and numbers correspond to the dimension. (a)-(b) all parameters equal $-1/2$. For $d = 1$, the tridiagonal model (tri) of Killip & Nenciu offers tremendous savings, without it is cheaper to get a sample in larger dimension. The number of rejections grows as $N2^d$.



(a) Larger Figure **??**, $d = 1, 2, 3, 4$



(b) $d = 1, 5, 10, 15$

Figure A.3. Rejection bounds. Given the proposal distribution $\omega_{\text{eq}}(x)\,\mathrm{d}x$, it is not surprising to see that the rejection procedure is the more efficient when the base measure $\mu(\mathrm{d}x) = \omega_{\text{eq}}(x)\,\mathrm{d}x$, i.e., coefficients $= -\frac{1}{2}$, than for larger coefficients. The larger the coefficients the greater the gap.

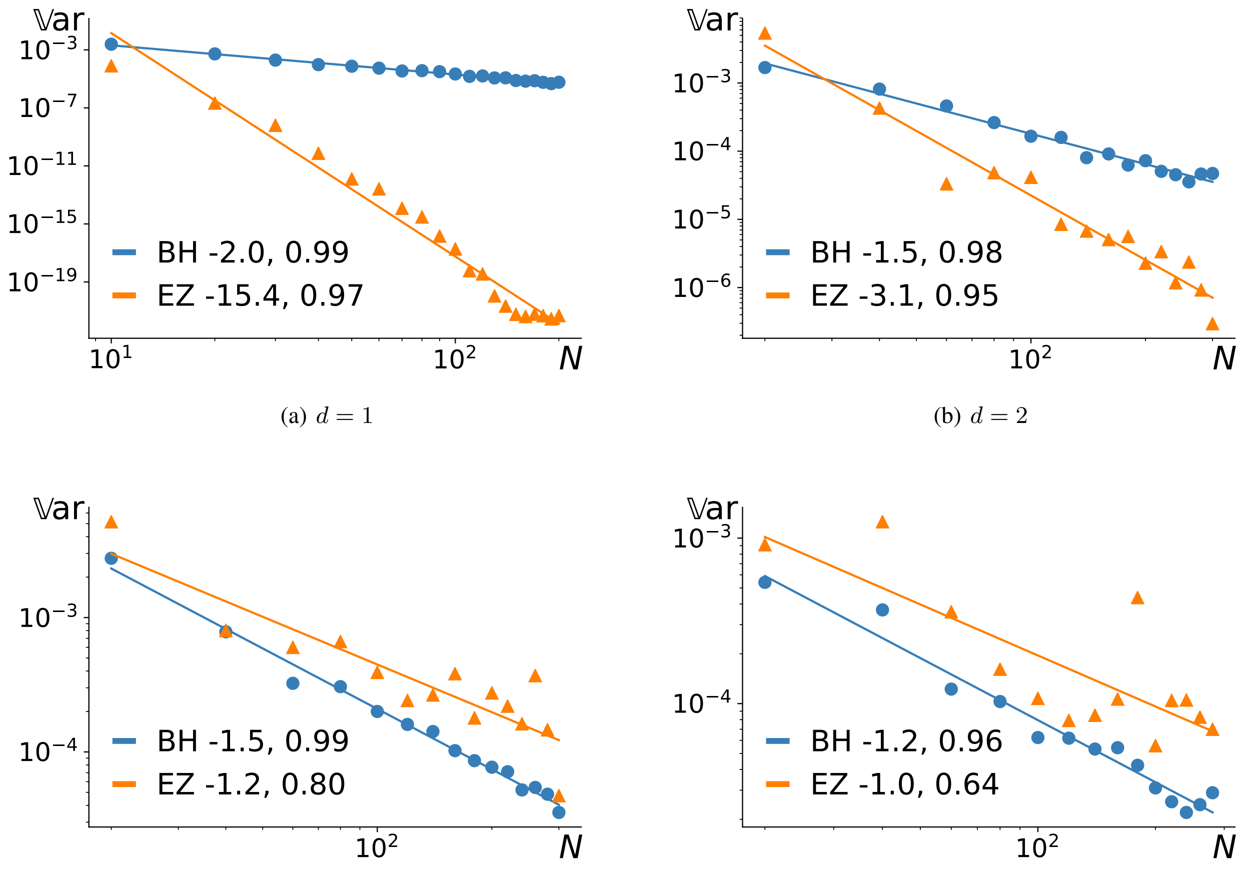
B. Experiments

B.1. Reproducing the bump example

In Section 4.1, we reproduce the experiment of Bardenet & Hardy (2016, Section 3) where they illustrate the behavior of $\widehat{I}_N^{\mathrm{BH}}$ on a unimodal, smooth bump function:

$$f(x) = \prod_{i=1}^{d} \exp\left(-\frac{1}{1-\varepsilon-(x^i)^2}\right) \mathbb{1}_{[-1+\varepsilon,1-\varepsilon]}(x^i). \tag{B.1}$$

For each value of N, we sample 100 times from the same multivariate Jacobi ensembles with i.i.d. uniform parameters on $[-1/2, 1/2]$, compute the resulting 100 values of each estimator, and plot the two resulting sample variances. In addition, in Figure B.2 we test the potential hope for a CLT for $\widehat{I}_N^{\mathrm{EZ}}$ and compare with $\widehat{I}_N^{\mathrm{BH}}$ for which the CLT (8) holds, in the regime $N = 300$.



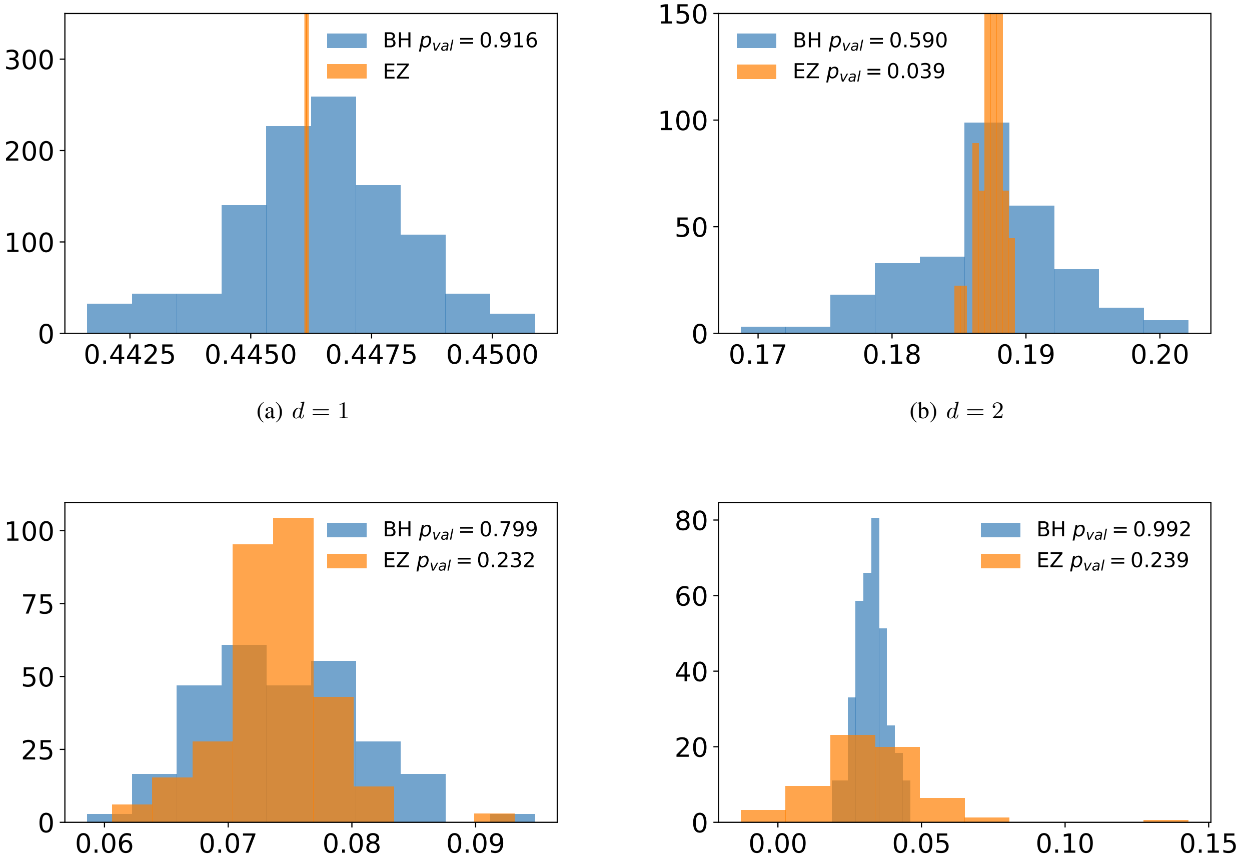
(a) $d = 1$

(b) $d = 2$

(c) $d = 3$

(d) $d = 4$

Figure B.1. Reproducing the bump function ($\varepsilon = 0.05$) experiment of Bardenet & Hardy (2016), cf. Section 4.1. The expected variance decay of order $1/N^{1+1/d}$ is observed for BH. For $d = 1$, EZ has almost no variance for $N \geq 100$: the bump function is extremely well approximated by a polynomials of degree $N \geq 100$.



(a) $d = 1$

(b) $d = 2$

(c) $d = 3$

(d) $d = 4$

Figure B.2. Histogram of 100 independent estimates $\widehat{I}_N^{\mathrm{BH}}$ and $\widehat{I}_N^{\mathrm{EZ}}$ of the integral of the bump function ($\varepsilon = 0.05$) with $N = 300$ and associated p-value of Kolmogorov-Smirnov test, cf. Section 4.1. The fluctuations of BH confirm to be Gaussian (cf. CLT (8)). (a) the bump function is extremely well approximated by a polynomial of degree 300 hence $\widehat{I}_N^{\mathrm{EZ}}$ has almost no variance. (b)-(c)-(d) few outliers seem to break the potential Gaussianity of $\widehat{I}_N^{\mathrm{EZ}}(f)$. (d) $\widehat{I}_N^{\mathrm{EZ}}(f)$ does not preserves the sign of the integrand.

B.2. Integrating sums of eigenfunctions

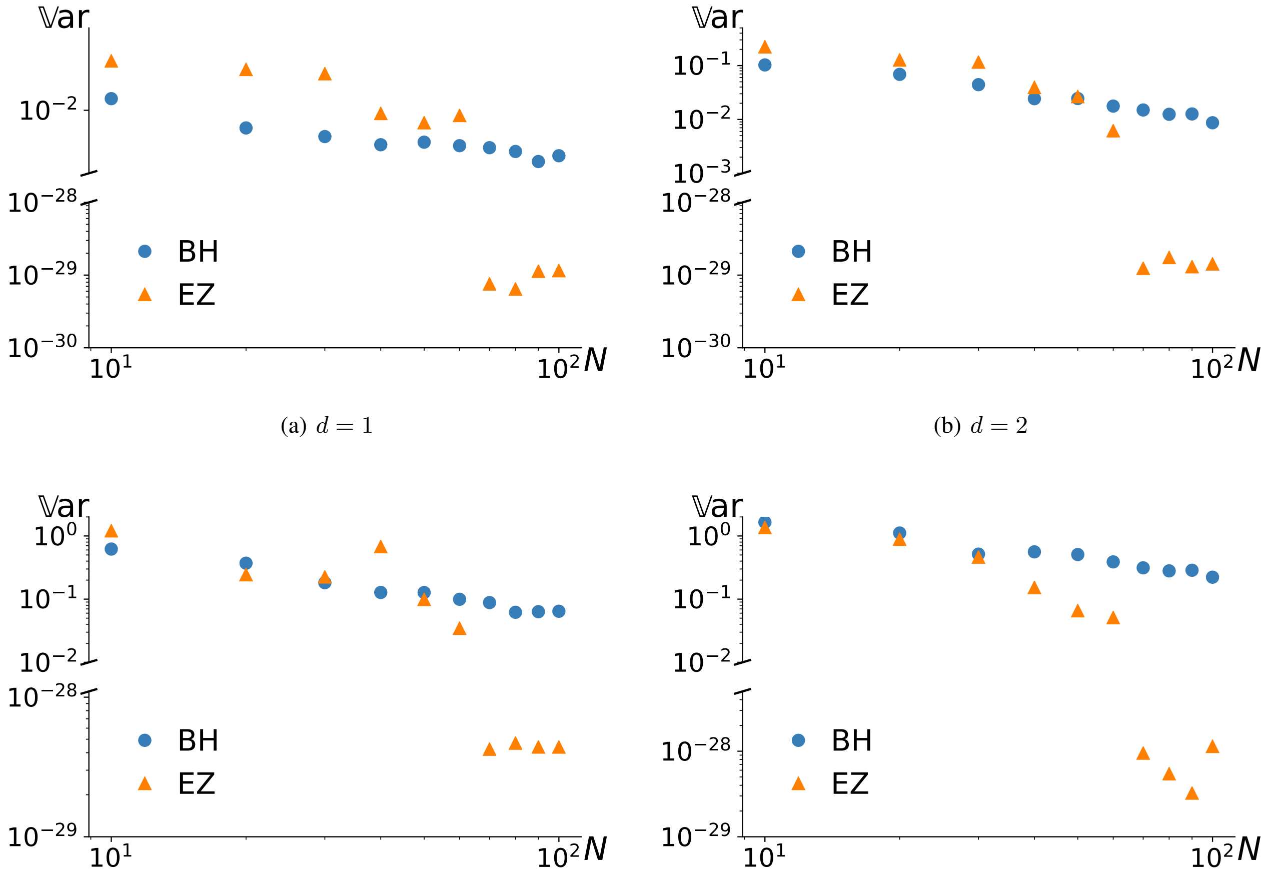
In the next series of experiments, we are mainly interested in testing the variance decay of $\widehat{I}_N^{\mathrm{EZ}}(f)$ prescribed by Theorem 1. To that end, we consider functions of the form given by (13), i.e.,

$$f(x) = \sum_{\mathfrak{b}(k)=0}^{N_{\mathrm{modes}}-1} \frac{1}{\mathfrak{b}(k)+1} \phi_k(x), \tag{B.2}$$

whose integral w.r.t. μ is to be estimated based on realizations of the multivariate Jacobi ensemble with kernel $K_N(x,y) = \sum_{\mathfrak{b}(k)=0}^{N-1} \phi_k(x)\phi_k(y)$ where $N \neq N_{\mathrm{modes}}$ a priori. This means that the function f can be either fully ($N_{\mathrm{modes}} \leq N$) or partially ($N_{\mathrm{modes}} > N$) decomposed in the eigenbasis of the kernel. In both cases, we let the number of points N used to build the two estimators vary from 10 to 100 in dimensions $d = 1$ to 4.

In the first setting, we set $N_{\mathrm{modes}} = 70$. Thus, N eventually reaches the number of functions used to build f in (13), after what $\widehat{I}_N^{\mathrm{EZ}}$ is an exact estimator. For each dimension d, Figure B.3 indeed shows the drop in the variance of $\widehat{I}_N^{\mathrm{EZ}}$ once the number of points of the DPP hits the threshold $N = N_{\mathrm{modes}}$. This is in perfect agreement with Theorem 1: once $f \in \mathcal{H}_{N_{\mathrm{modes}}} \subseteq \mathcal{H}_N$, the variance (11) is zero.

The second setting has $N_{\mathrm{modes}} = N + 1$, so that the number of points N is never enough for the variance (11) to be zero. As N increases the contribution of the extra mode $\phi_{\mathfrak{b}^{-1}(N)}$ in (13) decreases as $\frac{1}{N}$. Hence, from Theorem 1 we expect a variance decay of order $\frac{1}{N^2}$, which we observe in practice, cf. Figure B.4.



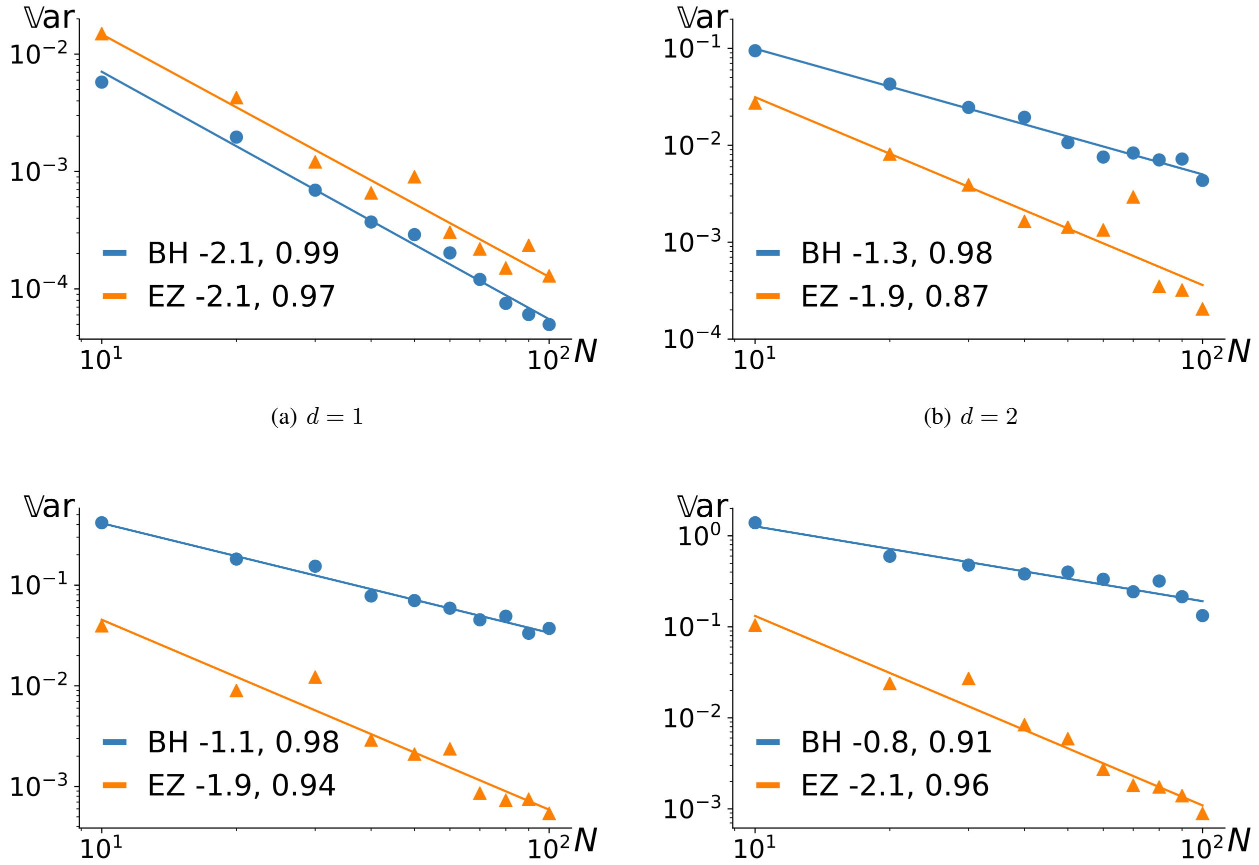
(a) $d = 1$

(b) $d = 2$

(c) $d = 3$

(d) $d = 4$

Figure B.3. Comparison of $\widehat{I}_N^{\mathrm{BH}}$ and $\widehat{I}_N^{\mathrm{EZ}}$ integrating a finite sum of 70 eigenfunctions of the DPP kernel as in (13), cf. Section 4.2.



(a) $d = 1$

(b) $d = 2$

(c) $d = 3$

(d) $d = 4$

Figure B.4. Comparison of $\widehat{I}_N^{\mathrm{BH}}$ and $\widehat{I}_N^{\mathrm{EZ}}$ for a linear combination of $N+1$ eigenfunctions of the DPP kernel as in (13), cf. Section 4.2.

B.3. Further experiments

In Appendices B.3.1-B.3.4 we test the robustness of both BH and EZ estimators, when applied to functions presenting discontinuities or which do not belong to the span of the eigenfunctions of the kernel. Although the conditions of the CLT (8) associated to $\widehat{I}^{\mathrm{BH}}$ are violated, the corresponding variance decay is smooth but not as fast. For $\widehat{I}^{\mathrm{EZ}}$ the performance deteriorate with the dimension. Indeed, the cross terms arising from the Taylor expansion of the different functions introduce monomials, associated to large coefficients, that do not belong to $\mathcal{H}_N$. Sampling more points would reduce the variance (11). But more importantly, for EZ to excel, this suggests to adapt the kernel to the basis where the integrand is known to be sparse or to have fast-decaying coefficients.

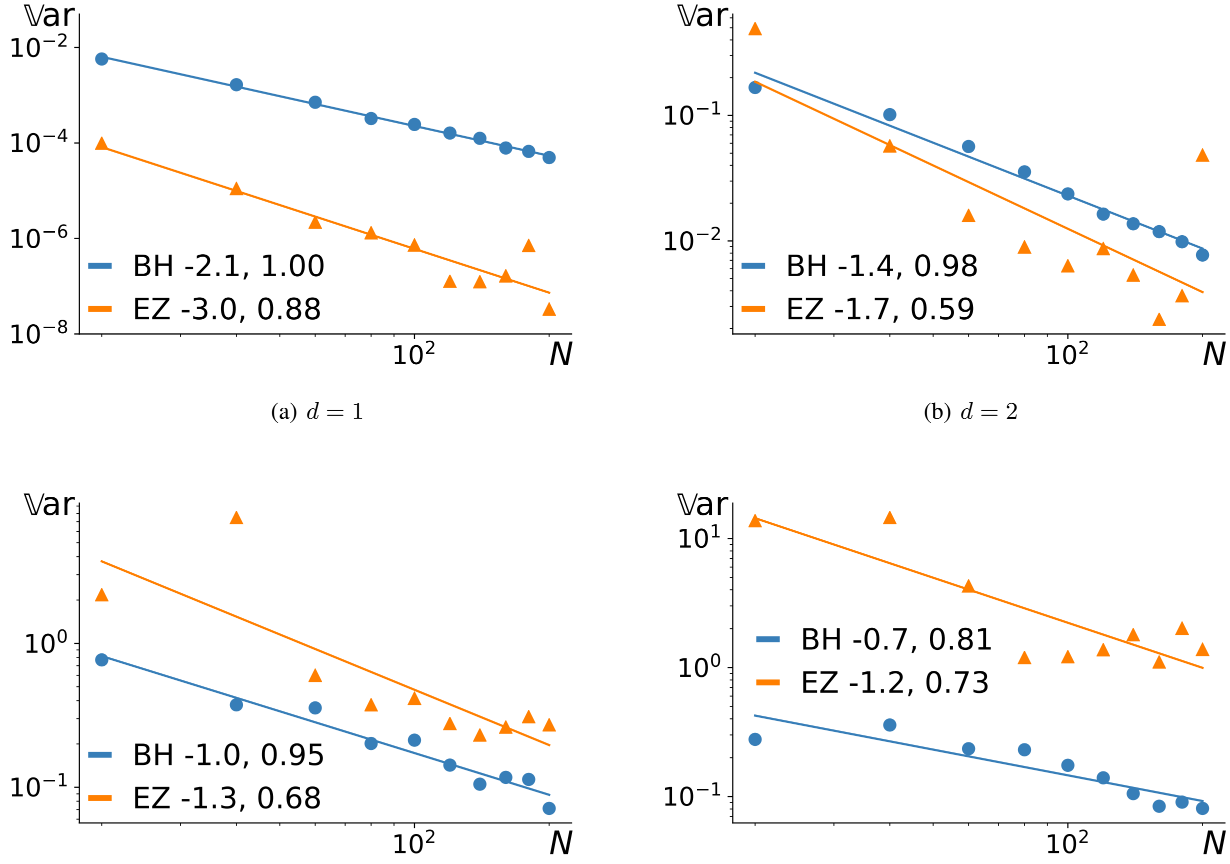
B.3.1. Integrating absolute value

We consider estimating the integral

$$\int_{[-1,1]^d} \prod_{i=1}^{d} |x^i|(1-x^i)^{a^i}(1-x^i)^{b^i}\,\mathrm{d}x^i \tag{B.3}$$

where $a^1, b^1 = -\frac{1}{2}$ and a^i, b^i i.i.d. uniformly in $[-\frac{1}{2}, \frac{1}{2}]$, using BH (7) and EZ (12) estimators.

Results are given in Figure B.5.



(a) $d = 1$

(b) $d = 2$

(c) $d = 3$

(d) $d = 4$

Figure B.5. Comparison of $\widehat{I}_N^{\mathrm{BH}}$ and $\widehat{I}_N^{\mathrm{EZ}}$ for absolute value, cf. Section B.3.

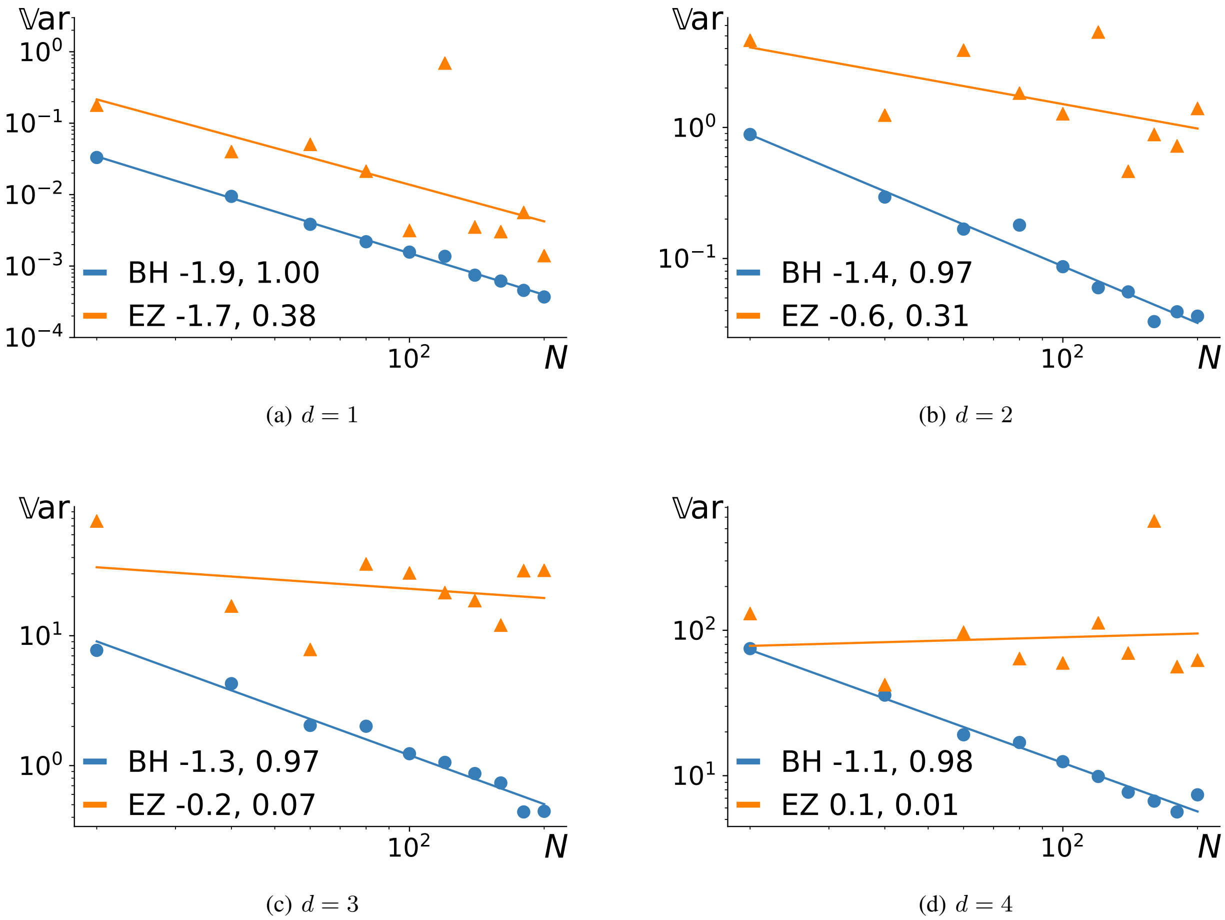
B.3.2. Integrating Heaviside

Let $H(x) = \begin{cases} 1, & \text{if } x > 0 \\ -1, & \text{otherwise} \end{cases}$. We consider estimating the integral

$$\int_{[-1,1]^d} \prod_{i=1}^{d} H(x^i)(1-x^i)^{a^i}(1-x^i)^{b^i} \,\mathrm{d}x^i \tag{B.4}$$

where $a^1, b^1 = -\frac{1}{2}$ and a^i, b^i i.i.d. uniformly in $[-\frac{1}{2}, \frac{1}{2}]$, using BH (7) and EZ (12) estimators.

Results are given in Figure B.6.



(a) $d = 1$

(b) $d = 2$

(c) $d = 3$

(d) $d = 4$

Figure B.6. Comparison of $\widehat{I}_N^{\mathrm{BH}}$ and $\widehat{I}_N^{\mathrm{EZ}}$ for Heaviside function, cf. Section **??**.

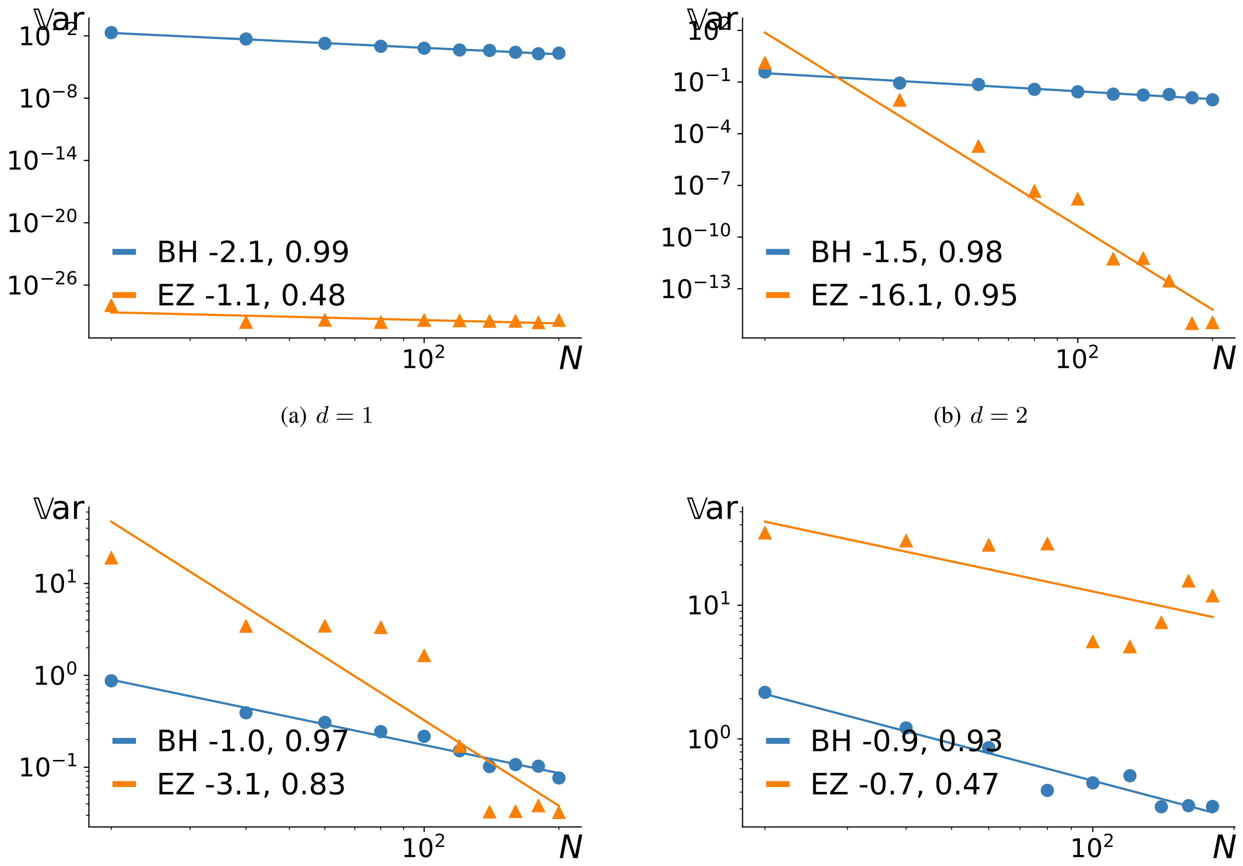
B.3.3. Integrating cosine

We consider estimating the integral

$$\int_{[-1,1]^d} \prod_{i=1}^{d} \cos(\pi x^i)(1-x^i)^{a^i}(1-x^i)^{b^i} \, \mathrm{d}x^i \tag{B.5}$$

where $a^1, b^1 = -\frac{1}{2}$ and a^i, b^i i.i.d. uniformly in $[-\frac{1}{2}, \frac{1}{2}]$, using BH (7) and EZ (12) estimators.

Results are given in Figure B.7



(a) $d = 1$

(b) $d = 2$

(c) $d = 3$

(d) $d = 4$

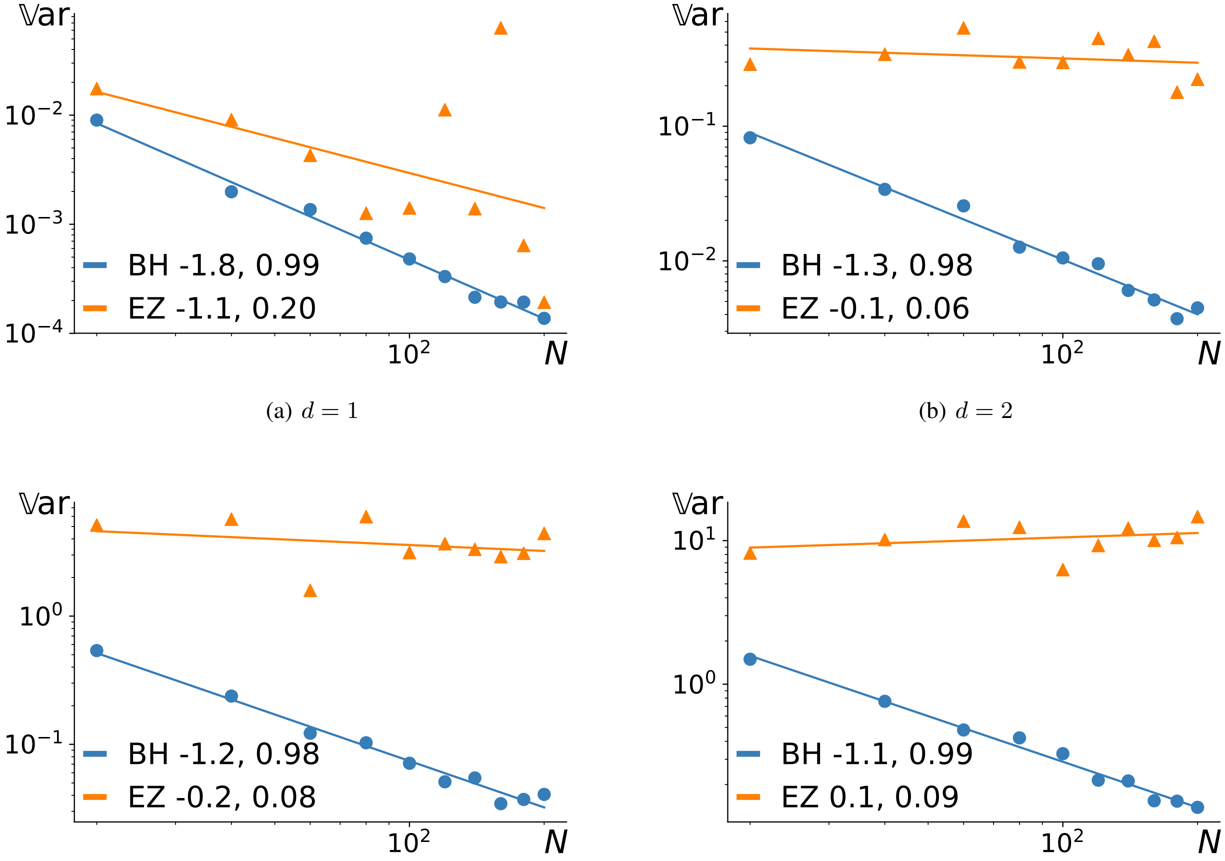
Figure B.7. Comparison of $\widehat{I}_N^{\mathrm{BH}}$ and $\widehat{I}_N^{\mathrm{EZ}}$ for cosine, cf. Section **??**.

B.3.4. Integrating a mixture of smooth and non smooth functions

Let $f(x) = H(x)(\cos(\pi x) + \cos(2\pi x) + \sin(5\pi x))$. We consider estimating the integral

$$\int_{[-1,1]^d} \prod_{i=1}^{d} f(x^i)(1-x^i)^{a^i}(1-x^i)^{b^i}\,\mathrm{d}x^i \tag{B.6}$$

where $a^1, b^1 = -\frac{1}{2}$ and a^i, b^i i.i.d. uniformly in $[-\frac{1}{2}, \frac{1}{2}]$, using BH (7) and EZ (12) estimators.



(a) $d = 1$

(b) $d = 2$

(c) $d = 3$

(d) $d = 4$

Figure B.8. Comparison of $\widehat{I}_N^{\mathrm{BH}}$ and $\widehat{I}_N^{\mathrm{EZ}}$, cf. Section **??**.