State-of-art minibatches via novel DPP kernels: discretization, wavelets, and rough objectives

Hoang-Son Tran[*]
Department of Mathematics
National University of Singapore
transonsp97@gmail.com

Pranav Gupta[†]
Division of Applied Mathematics
Brown University
pranav_gupta@brown.edu

Rémi Bardenet
Univ. Lille, CNRS, Centrale Lille
UMR 9189 – CRIStAL
remi.bardenet@cnrs.fr

Subhro Ghosh
Department of Mathematics
National University of Singapore
subhrowork@gmail.com

Abstract

Determinantal point processes (DPPs) have emerged as a kernelized alternative to vanilla independent sampling for generating efficient minibatches, coresets and other parsimonious representations of large-scale datasets. While theoretical foundations and promising empirical performance have been demonstrated, there are two challenges for current proposals for DPP-based coresets or minibatches. The first is the need for families of DPPs with certain key variance reduction properties, usually constructed in a continuous setting, of which there are few known examples. The second is the need for an ad-hoc construction of a discrete DPP defined on a given dataset, that inherits such variance reduction. In this work, we contribute to the programme of establishing DPPs as a subsampling toolbox for ML by advancing on these two fronts. First, we propose new DPPs on the Euclidean space based on wavelets, with provably better accuracy guarantees than the best known rates. Second, we introduce a general method to convert such continuous DPPs, which are more amenable to proving analytical statements, into discrete kernels, which are pertinent for subsampling tasks such as minibatch and coreset constructions. This conversion mechanism simultaneously preserves the desired variance decay and reveals a low-rank decomposition of the discrete kernel, which makes sampling the corresponding DPP computationally inexpensive. En route, we enlarge the class of ML tasks amenable to improvements via DPP-based minibatches and coresets to include objective functions with arbitrarily low regularity, and rate guarantees that explicitly adapt to this regularity.

1 Introduction

Determinantal point processes (DPPs; [19]) are random point configurations that are parametrized by kernel functions. Like kernel machines, DPPs are particularly tractable, both mathematically and computationally [13, 27]. In machine learning, since the seminal work [16], they have been used as diversity-promoting models. Closer to our purpose, DPPs have also been shown to provably improve on independent sampling when designing numerical quadratures [3, 7, 8], minibatches in stochastic gradient descent [4] or building coresets [28, 5, 14].

[*]Equal contribution.
[†]Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

To fix ideas, for a configuration of points $\mathcal{S} \subset \mathbb{R}^d$ of cardinality n and a class $\mathcal{F}$ of functions, we define the *linear statistic*

$$\Lambda_{\mathcal{S}}(f) = \sum_{Y \in \mathcal{S}} f(Y), \quad f \in \mathcal{F}. \tag{1}$$

The recipe for constructing a minibatch or coreset using DPPs typically requires two ingredients. The first is a probabilistic guarantee that, when $\mathcal{S} = \mathcal{S}_{\mathrm{c}}$ is a specific DPP whose points are allowed to vary in all of $\mathbb{R}^d$ (loosely termed a "continuous" DPP), a suitable set of unbiased linear statistics (1) has a standard error that decays faster than the classical rate $1/\sqrt{n}$ of i.i.d. sampling. Yet both coresets and minibatches are constrained to be subsets of an N-point dataset $\mathcal{X} \subset \mathbb{R}^d$, so the second ingredient is to design a second DPP $\mathcal{S}_{\mathrm{d}}$ that is *discrete*, i.e. supported on the subsets of $\mathcal{X}$, and inherits the variance decay of the original continuous DPP $\mathcal{S}_{\mathrm{c}}$.

Motivation and our contributions. Available designs for DPP minibatches and coresets with qualitatively improved theoretical performance over independent sampling [4, 28, 5, 14] rely on DPPs known as multivariate orthogonal polynomial ensembles [3]. The latter come with a standard error for linear statistics of order $n^{-(\frac{1}{2}+\frac{1}{2d})}$, holding when the test functions in $\mathcal{F}$ are essentially C^1. A key contribution of this paper is to introduce a novel class of DPP samplers, defined using *wavelets*, such that the standard error of $\Lambda(f)$ decays fast for quite general test functions, namely for Hölder continuous functions and functions belonging to fractional Sobolev spaces. Furthermore, the decay rates *adapt to* the *regularity of* f in an explicit manner, and in the C^1 setting provide an improved guarantee of a standard error in $n^{-(\frac{1}{2}+\frac{1}{d})}$, which matches the best known rates in that setting [18] without complex-analytic machinery. Our construction also *improves* on the best rates obtained with *weak smoothness requirements* for variance reduction with DPPs [8], and opens up applications to e.g. coresets for *non-smooth loss functions*, or quadrature in rough settings such as for financial data. An interesting by-product of our analysis is the identification of classical *stratified sampling* in the continuum as a special case of a wavelet-based DPP sampler, thereby explaining its superior empirical performance as a coreset sampling mechanism [5], and providing natural motivation to pursue determinantal samplers as a robust and structured subset generation mechanism.

The transfer of the favorable guarantees from the continuum background space to the discrete dataset has historically had the problem of being mathematically and computationally involved: the construction in [4], for instance, requires a cumbersome spectral truncation of an $N \times N$ matrix that is near-to but not low-rank. Alternatives include mimicking the continuum construction in the discrete setting, such as with the discrete OPE in [28], which requires customized arguments for their rates [14] and is very much tied to the structured scenario of orthogonal polynomials. A key contribution of our paper is a comprehensive pipeline for the *conversion of continuum kernels* with salutary variance decay properties to *canonical discrete kernels*, defined on the dataset, that inherit similar guarantees. The proposed procedure is both an algorithmic recipe that is *computationally cheaper* than previous alternatives, as well as a mathematical technique that provably ensures the *transfer of variance decay rates* from the continuum to the discrete. Using this pipeline as a turn-key mechanism potentially enables future investigations in this area to focus squarely on the more analytically tractable question of designing continuum DPP kernels with favourable variance decay. Our kernel transfer pipeline comes with an *in-built low-rank decomposition* of the discrete kernel, which has important computational benefits in the context of modern DPP sampling algorithms [16, 10, 6]. The pipeline is based on certain novel stochastic identities that could be of independent probabilistic interest. Finally, our pipeline is complementary to, but different from, the recent general results of [14] on variance transfer from continuous to discrete DPPs: while they implicitly assume the availability of a natural candidate discrete DPP (like a discrete OPE) that can be shown in a case-by-case analysis to approximate the underlying continuous kernel in a suitable sense, we rather assume that we can evaluate the kernel of the continuous DPP and use it to *define* a discrete DPP that automatically inherits its variance-decay properties.

The rest of the paper is organized as follows. In Section 2, we introduce some necessary background on DPPs. In Section 3, we present our novel discrete DPP construction; proofs are deferred to Appendix A. In Section 4, we present our new wavelet-based continuous DPPs; proofs are deferred to Appendix B. In Section 5, we give two examples of how to put our two contributions together and obtain new theoretical guarantees for coresets and numerical integration; proofs are deferred to Appendix C. In Section 7, we numerically validate our theoretical results; additional experiments can be found in Appendix D. We conclude with opening remarks in Section 8.

2 Background

DPPs. Let E be a Polish space equipped with a reference measure μ. Typical examples of E include Euclidean spaces, manifolds, and discrete spaces such as graphs. A *point process* $\mathcal{S}$ on E is a random locally finite subset of E. We say that $\mathcal{S}$ is a DPP (w.r.t. μ) if there exists a measurable function $K : E^2 \to \mathbb{C}$ such that

$$\mathbb{E}\Big[\sum_{\neq} f(x_{i_1}, \ldots, x_{i_k})\Big] = \int_{E^k} f(x_1, \ldots, x_k) \det[K(x_i, x_j)]_{k\times k} \, \mathrm{d}\mu^{\otimes k}(x_1, \ldots, x_k), \tag{2}$$

where the sum in the LHS ranges over all pairwise distinct k-tuples of $\mathcal{S}$, for all bounded measurable $f : E^k \to \mathbb{R}$ and for all $k \in \mathbb{N}$. Such a function K is called a *kernel* of the DPP $\mathcal{S}$, and μ is the *reference measure* (see [20, 13]). We will denote $\mathcal{S} \sim \mathrm{DPP}(K, \mu)$.

An important class of DPPs is the *projection DPPs*, which are defined by projection kernels. To elaborate, let $n \geq 1$ and $\phi_1, \ldots, \phi_n$ be n (representatives of) orthonormal functions in $L^2(\mu)$, and define $K(x, y) := \sum_{k=1}^n \phi_k(x)\phi_k(y)$. Notice that, K, as an integral kernel, defines a projection operator on the space $L^2(\mu)$. The resultant DPP defined by (K, μ) is called a *projection DPP* of rank n. A useful probabilistic interpretation of projection DPPs is the following: if $\mathcal{S} \sim \mathrm{DPP}(K, \mu)$ where K is a projection kernel of rank n, then $\mathcal{S} = \{X_1, \ldots, X_n\}$ can be obtained by drawing $(X_1, \ldots, X_n)$ from the probability measure

$$p(x_1, \ldots, x_n) \, \mathrm{d}\mu(x_1) \cdots \mathrm{d}\mu(x_n) := \frac{1}{n!} \det\big(K(x_i, x_j)\big)_{1\leq i,j\leq n} \, \mathrm{d}\mu(x_1) \cdots \mathrm{d}\mu(x_n). \tag{3}$$

For more discussions on DPPs, we refer to the surveys [13, 16].

Discrete DPPs. When E has finite cardinality N and μ is taken to be counting measure on E, we speak of DPP(K, μ) as a *discrete DPP on* E. One can check that this is equivalent to the following more convenient definition. A random subset $\mathcal{S} \subset E$ is a discrete DPP if there is an $N \times N$-matrix $\mathbf{K}$ indexed by elements of E such that $\mathbb{P}(\mathcal{S} \supset T) = \det[\mathbf{K}_T]$ for all subset $T \subset E$. Here $\mathbf{K}_T$ denotes the principal submatrix of $\mathbf{K}$ whose rows and columns are indexed by the elements of T.

A popular class of discrete DPPs in ML is given by *L-ensembles*, which are random subsets $\mathcal{S} \subset E$ satisfying $\mathbb{P}(\mathcal{S} = T) \propto \det(\mathbf{L}_T)$ for all subsets $T \subset E$, where $\mathbf{L}$ is an $N \times N$ symmetric matrix indexed by E, with positive eigenvalues. It is known that every L-ensemble is a DPP with kernel $\mathbf{K} = \mathbf{L}(\mathbf{I} + \mathbf{L})^{-1}$. However, not every DPP is an L-ensemble. In particular, an L-ensemble has random cardinality, whereas a projection DPP has deterministic cardinality. To obtain a subset of fixed size when sampling with L-ensembles, a standard approach is to condition an L-ensemble $\mathcal{S}$ on the event $|\mathcal{S}| = m$, for some $m \in \mathbb{N}$ with $m \leq \mathrm{rank}(\mathbf{L})$. This gives rise to the so-called *m-DPPs*; see [16]. We also refer the interested reader to [29] for a unifying concept of *extended* L-ensembles.

Orthogonal Polynomial Ensembles (OPEs). When $E = \mathbb{R}^d$ and μ has a density w.r.t. Lebesgue, we speak of DPP(K, μ) as a *continuous* DPP. A classical example consists in letting $\phi_1, \ldots, \phi_n$ be the result of applying the Gram-Schmidt orthonormalization process in $L^2(\mu)$ to the n first multivariate monomials, after having chosen an order on multivariate indices. The resulting DPP is called *multivariate orthogonal polynomial ensemble* (OPE), and we denote its kernel as $K_{\mathrm{OPE},\mu}$. Under conditions on μ, when $\mathcal{S}$ is the multivariate OPE, for smooth enough $f : \mathbb{R}^d \to \mathbb{R}$, and the weight function $w : x \mapsto K_{\mathrm{OPE},\mu}(x, x)^{-1}$, [3] showed that the linear statistic $\Lambda_{\mathcal{S}}(fw)$ satisfies $\mathbb{E}\left[\Lambda_{\mathcal{S}}(fw)\right] = \int f \mathrm{d}\mu$ and $\sqrt{\mathrm{Var}\left[\Lambda_{\mathcal{S}}(fw)\right]} = O(1/n^{\frac{1}{2}+\frac{1}{2d}})$, improving over the classical Monte Carlo rate of $O(1/\sqrt{n})$. This was leveraged in [4] to construct DPP-based minibatch samplers, and in [5] to devise coreset sampling mechanisms.

3 A general pipeline from the continuum to the discrete

In line with Section 2, let E be a Polish space equipped with a reference μ, and $\mathcal{S}_{\mathrm{c}} \sim \mathrm{DPP}(K, \mu)$ with K the projection kernel onto the span of n orthonormal functions $\psi_1, \ldots, \psi_n \in L^2(\mu)$. Let $X_1, \ldots, X_N \sim \nu$ i.i.d., where $\nu(\mathrm{d}x) = \rho(x)\mu(\mathrm{d}x)$ is a probability measure with $\rho(x) > 0$ for μ-a.e x. We assume that ν is non-atomic so that the samples $X_1, \ldots, X_N$ are pairwise distinct, almost surely. We think of $\mathcal{X} = \{X_1, \ldots, X_N\}$ as a dataset. Conditionally on $\mathcal{X}$, in this section we build two discrete DPPs that inherit the variance of linear statistics of $\mathcal{S}_{\mathrm{c}}$.

First, define the matrix

$$\mathbf{\Psi} := \begin{pmatrix} \psi_1(X_1) & \dots & \psi_n(X_1) \\ \vdots & \ddots & \vdots \\ \psi_1(X_N) & \dots & \psi_n(X_N) \end{pmatrix} = [\psi_k(X_i)] \in \mathbb{C}^{N\times n}, \tag{4}$$

and set $m := \mathrm{rank}(\mathbf{\Psi})$. We remark that $\mathbf{\Psi}$ and m are measurable with respect to the data set $\mathcal{X}$.

Remark 1. *We note that m is random in general, and clearly $m \leq n$ almost surely. However, we have $\mathbb{E}[\mathbf{\Psi}^*\mathbf{D}(\rho^{-1})\mathbf{\Psi}] = N\mathbf{I}_n$, where $\mathbf{D}(\rho^{-1}) = \mathrm{diag}(\rho(X_1)^{-1}, \dots, \rho(X_N)^{-1})$. This suggests that $\mathbf{\Psi}$ has maximal rank $m = n$ with high probability. In many examples of interest, such as graded polynomial feature functions ψ_i with ν having a density w.r.t. Lebesgue, we even have $m = n$ with probability* 1.

Now, conditionally on $\mathcal{X}$, define $\mathcal{S}_{\mathrm{d}}$ as the m-DPP on $\mathcal{X}$ corresponding to the L-ensemble defined by

$$\mathbf{L} = \frac{1}{N}\mathbf{D}(\rho^{-1/2})\,\mathbf{\Psi}\mathbf{\Psi}^*\,\mathbf{D}(\rho^{-1/2}), \tag{5}$$

see Section 2 for definitions. Alternately, if we do not have query access to ρ but only to a (positive) estimator $\hat{\rho}$ of ρ, we shall consider the m-DPP $\widehat{\mathcal{S}}_{\mathrm{d}}$ on $\mathcal{X}$ corresponding to the L-ensemble defined by

$$\widehat{\mathbf{L}} = \frac{1}{N}\mathbf{D}(\hat{\rho}^{-1/2})\,\mathbf{\Psi}\mathbf{\Psi}^*\,\mathbf{D}(\hat{\rho}^{-1/2}). \tag{6}$$

We prove the following in the appendix, see Lemma A.1.

Proposition 3.1. *Both $\mathcal{S}_{\mathrm{d}}$ and $\widehat{\mathcal{S}}_{\mathrm{d}}$ are actually (discrete) projection DPPs on $\mathcal{X}$, with respective kernel the projection onto the span of the eigenvectors of $\mathbf{L}$ and $\widehat{\mathbf{L}}$.*

In particular, there exist algorithms to exactly sample $\mathcal{S}_{\mathrm{d}}$ or $\widehat{\mathcal{S}}_{\mathrm{d}}$, at a cost dominated by the computation of the eigenvectors of $\mathbf{L}$ and $\widehat{\mathbf{L}}$, respectively; see [6] and references therein. Rewriting e.g. (5) as $\mathbf{B}\mathbf{B}^*$, one realizes that the cost of finding these eigenvectors by rather diagonalizing $\mathbf{B}^*\mathbf{B}$ is $O(Nn^2 + n^3)$. In minibatch and coreset constructions, where $n \ll N$, the sampling cost is thus essentially linear in N. Finally, note that unlike the discrete OPEs used in [28, 14], the definition of our DPP kernels requires the evaluation of the eigenfunctions of the kernel K of $\mathcal{S}_{\mathrm{c}}$ in (4). Note that the kernels used for subsampling tasks with DPPs are typically projection kernels that have accessible Mercer decompositions (i.e., eigenfunction expansions), so in practice it is not even necessary to compute these eigenfunctions separately. We now show that $\mathcal{S}_{\mathrm{d}}$ and $\widehat{\mathcal{S}_{\mathrm{d}}}$ inherit the variance reduction of linear statistics of $\mathcal{S}_{\mathrm{c}}$, without any argument specific to the choice of the kernel K. This leaves the choice of the kernel as a flexibility in the hands of the user.

Theorem 3.2 (Discretization Pipeline - I)**.** *Let $f : E \to \mathbb{R}$ be a bounded test function. Assume that*

$$\exists\, K_{\max} < \infty \text{ such that } \frac{K(x,x)}{\rho(x)} \leq K_{\max} \text{ for } \nu\text{-almost every } x. \tag{A1}$$

Recall the definition of linear statistics in (1). Then for any $\delta, \delta' \in (0,1)$, with probability at least $1 - \delta - \delta'$ in the data set $\mathcal{X}$, we have

$$\begin{aligned} \mathrm{Var}[\Lambda_{\mathcal{S}_{\mathrm{d}}}(f)|\mathcal{X}] &\leq 2\mathrm{Var}[\Lambda_{\mathcal{S}_{\mathrm{c}}}(f)] + 4\|f\|_\infty^2 K_{\max}^2 \mathscr{E}(n,N,\delta,\delta'), \quad \text{and} \\ \mathrm{Var}[\Lambda_{\mathcal{S}_{\mathrm{d}}}(f)|\mathcal{X}] &\geq \frac{N-1}{2N}\mathrm{Var}[\Lambda_{\mathcal{S}_{\mathrm{c}}}(f)] - 2\|f\|_\infty^2 K_{\max}^2 \mathscr{E}(n,N,\delta,\delta'), \end{aligned}$$

where

$$\mathscr{E}(n,N,\delta,\delta') := 4\sqrt{\frac{2\log(2/\delta')}{N}} + \frac{4}{9N^2}\log^2\left(\frac{n^2+1}{\delta}\right) + \frac{4}{N}\log\left(\frac{n^2+1}{\delta}\right). \tag{7}$$

Several comments are in order. First, in the cases we have in mind, $K_{\max} \asymp n$. This is true under mild assumptions for multivariate OPEs, see Section A.1, and for the wavelet-based DPPs we will introduce in Section 4. Second, our algorithm produces a minibatch $\mathcal{S}_{\mathrm{d}}$ of size $m \leq n$, rather than one that is necessarily of size exactly n. However, we note that if N is sufficiently large so that $K_{\max}^2 \mathscr{E}(n,N,\delta,\delta') < 1$ (which is the relevant regime in practice), then on the same event

appearing in the conclusion of Theorem 3.2, the minibatch has size exactly n. This follows from our proof technique; see Lemma A.4. Third, and most importantly, Theorem 3.2 establishes a *variance-transfer guarantee*: the conditional variance of the linear statistic associated with the subsample $\mathcal{S}_{\mathrm{d}}$ is bounded above and below by that of the underlying continuous DPP, up to *explicit* error terms. In the regime most relevant for applications, where $N \gg n$, these error terms are negligible, leaving $\mathrm{Var}[\Lambda_{\mathcal{S}_{\mathrm{c}}}(f)]$ as the dominant term. By choosing a continuous DPP kernel K that is known to enjoy variance reduction (i.e., $\mathrm{Var}[\Lambda_{\mathcal{S}_{\mathrm{c}}}(f)] = O(n^{1-\alpha})$ for some $\alpha > 0$), the resulting minibatch $\mathcal{S}_{\mathrm{d}}$ thus automatically inherits this variance reduction property.

In the regime where only an estimator $\hat{\rho}$ is available, we show that $\widehat{\mathcal{S}}_{\mathrm{d}}$ defined by (6) enjoys the same bounds as $\mathcal{S}_{\mathrm{d}}$, up to an explicit additional error coming from density estimation.

Theorem 3.3 (Discretization Pipeline - II)**.** *Let $f : E \to \mathbb{R}$ be a bounded test function. Assume both* (A1) *and that*

$$\mathbb{P}\Big(\max_{1 \le i \le N} \Big| \frac{\rho(X_i)}{\hat{\rho}(X_i)} - 1 \Big| > \varepsilon \Big) \le \beta, \quad \textit{for some } \varepsilon \in [0,1), \beta \in [0,1). \tag{A2}$$

Then, for any $\delta, \delta' \in (0,1)$, with probability at least $1 - \delta - \delta' - \beta$ in the data set $\mathcal{X}$,

$$\begin{aligned} \mathrm{Var}[\Lambda_{\widehat{\mathcal{S}}_{\mathrm{d}}}(f)|\mathcal{X}] &\le 2(1+\varepsilon)^2 \mathrm{Var}[\Lambda_{\mathcal{S}_{\mathrm{c}}}(f)] + 32\|f\|_\infty^2 K_{\max}^2 \mathscr{E}(n,N,\delta,\delta') + 8\|f\|_\infty^2 \varepsilon^2 n, \quad \textit{and} \\ \mathrm{Var}[\Lambda_{\widehat{\mathcal{S}}_{\mathrm{d}}}(f)|\mathcal{X}] &\ge \frac{(N-1)}{2N}(1-\varepsilon)^2 \mathrm{Var}[\Lambda_{\mathcal{S}_{\mathrm{c}}}(f)] - 16\|f\|_\infty^2 K_{\max}^2 \mathscr{E}(n,N,\delta,\delta') - 4\|f\|_\infty^2 \varepsilon^2 n. \end{aligned}$$

Compared to Theorem 3.3, $\widehat{\mathcal{S}}_{\mathrm{d}}$ only suffers an additional error term of order $O(\varepsilon^2 n)$. This term is not restrictive for our purposes. Indeed, to retain the variance reduction property, it is enough that $\varepsilon = O(n^{-\alpha})$ for some $\alpha > 0$. In practice, one may even obtain estimators for which ε decays with the full sample size N, which is much larger than the minibatch size n. For instance, in [15], if the density ρ is α-Hölder continuous on $\mathbb{R}^d$, then the KDE estimator $\hat{\rho}_{\mathrm{KDE}}$ satisfies $\|\hat{\rho}_{\mathrm{KDE}} - \rho\|_\infty = O(N^{-\alpha/(2\alpha+d)})$ (ignoring logarithmic factors) with high probability (c.f. Theorem 2 and Remark 8 in [15]). That being said, we point out that we actually only need in (A2) to control the relative error of the density estimate *on the dataset* $\{X_i\}_{i=1}^N$, with high probability. This should allow us to go beyond the usual L^∞ bounds on $\hat{\rho} - \rho$ and work in much weaker settings.

4 A new class of DPPs: wavelet bases and the variance reduction phenomenon

In this section, we introduce a new class of DPPs constructed from wavelets and show that they come with fast-decaying variance of linear statistics even for non-smooth test functions, making them ideal candidates as the continuous DPP in Theorems 3.2 and 3.3. To the best of our knowledge, this yields the first example of a class of DPPs in arbitrary dimension with this property that does not arise from orthogonal polynomial ensembles [3] or the Dirichlet ensemble [8].

From wavelets to a DPP kernel. Let $\phi : \mathbb{R} \to \mathbb{R}$ be a bounded, compactly supported, orthonormal scaling function, associated with a multiresolution analysis, normalized by $\int \phi(x)\mathrm{d}x = 1$. We refer to Appendix B for a self-contained introduction to multiresolution analysis. A typical example would be the Haar scaling function $\phi_{\mathrm{Haar}}(x) = \mathbf{1}_{[0,1)}(x)$. For each $j, k \in \mathbb{Z}$, we define

$$\phi_{j,k}(x) := 2^{-j/2}\phi(2^{-j}x - k) \quad , \quad \Phi_{j,\mathbf{k}}(\mathbf{x}) = \prod_{i=1}^d \phi_{j,k_i}(x_i),$$

where $\mathbf{k} = (k_1, \dots, k_d) \in \mathbb{Z}^d$ and $\mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$. Fix $j \in \mathbb{N}$, we define an index set

$$\mathcal{I} = \mathcal{I}(j) := \{\mathbf{k} = (k_1, \dots, k_d) \in \mathbb{Z}^d : \mathrm{supp}(\Phi_{-j,\mathbf{k}}) \subset [0,1]^d\}.$$

We define the wavelet DPP associated to the scaling function ϕ as the DPP $\mathcal{S}_{\mathrm{c}}$ with reference measure the Lebesgue measure on $\mathbb{R}^d$, and kernel

$$K(\mathbf{x}, \mathbf{y}) := \sum_{\mathbf{k} \in \mathcal{I}} \Phi_{-j,\mathbf{k}}(\mathbf{x})\Phi_{-j,\mathbf{k}}(\mathbf{y}), \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \tag{8}$$

We highlight that j is fixed, and that K is parametrized by the scaling function ϕ and the integer j. Moreover, by construction, the $(\Phi_{-j,\mathbf{k}})_{\mathbf{k}}$ are orthonormal in $L^2(\mathbb{R}^d, \mathrm{d}\mathbf{x})$, so that K is a projection kernel of rank $n := |\mathcal{I}(j)|$. We thus control the cardinality n of our DPP through the choice of j and ϕ. In particular, a simple exercise shows that if $\mathrm{supp}(\phi) \subset [a, b]$ then $n = (\lfloor 2^j - b \rfloor - \lceil -a \rceil + 1)^d \asymp 2^{dj}$. While the joint choice of j, a, b can help choose a particular cardinality, asymptotic results will let $j \to \infty$, and as such the number of points will be exponential in the dimension. This is a hint that wavelet DPPs should be used after some low-dimensional approximation, e.g. after running a PCA, as we demonstrate in the experimental Section 7. On the other hand, when d is the effective dimension of the data, an $\exp(d)$ points would be minimally required to fully "explore the space", so this sample size is rather reasonable. As we shall now see, the DPP with kernel (8) generalizes stratified sampling.

Proposition 4.1 (Connections to Stratified Sampling)**.** *When the scaling function is $\phi_{\mathrm{Haar}}(x) = \mathbf{1}_{[0,1)}(x)$, we have $n = 2^{dj}$. Furthermore, the supports of $\Phi_{-j,\mathbf{k}}$, $\mathbf{k} \in \mathcal{I}$, are disjoint, and $\mathcal{S}_{\mathrm{c}} \sim \mathrm{DPP}(K, \mathrm{d}\mathbf{x})$ has the same distribution as the set $\{Y_{\mathbf{k}} : \mathbf{k} \in \mathcal{I}\}$, where the $Y_{\mathbf{k}}$s are independent and $Y_{\mathbf{k}}$ is drawn uniformly on $\mathrm{supp}(\Phi_{-j,\mathbf{k}})$.*

Our continuous-to-discrete pipeline (see below) subsequently enables us to transfer this to a new kind of discretized analogue of stratified sampling that is tuned to the problem of subset sampling from a large dataset.

Variance reduction. Consider bounded, compactly supported test functions,

$$f : \mathbb{R}^d \to \mathbb{R}, f(\mathbf{x}) = 0 \text{ on } \mathbb{R}^d \setminus (0,1)^d \text{ and } \|f\|_\infty := \sup_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) < \infty. \tag{$*$}$$

To study the impact of smoothness, we further consider two classes of functions: the space $C^{0,s}$ of Hölder continuous functions of order s, and the Sobolev space H^s, both with parameter $s \in (0, 1]$. For completeness, we recall the definitions of these spaces and their norms in Appendix B.2. We denote by $\mathcal{F}_s$ the space of functions f satisfying ($*$) and belonging to either $C^{0,s}$ or H^s. Furthermore, when clear from context what space f belongs to, we denote by $|f|_s$: either (a) for Hölder classes, the Hölder seminorm $|f|_{C^{0,s}}$ ($s \in (0, 1]$); or (b) for Sobolev classes, the Gagliardo seminorm $|f|_{H^s}$ (for $s \in (0, 1)$) and $\|\nabla f\|_{L^2}$ (for $s = 1$).

Theorem 4.2 (Variance reduction for wavelet kernels)**.** *Let $s \in (0, 1]$. Consider the DPP $\mathcal{S}_{\mathrm{c}}$ on $(\mathbb{R}^d, \mathrm{d}\mathbf{x})$ defined by the kernel (8), then there exists a constant C (depending only on ϕ and d) such that,*

$$\forall f \in \mathcal{F}_s, \quad \mathrm{Var}[\Lambda_{\mathcal{S}}(f)] \leq C\, |f|_s^2\, n^{1-2s/d}.$$

It is remarkable that when $s = 1$, we have a variance of the order $n^{1-2/d}$, compared to $n^{1-1/d}$ for multivariate OPEs. We thus match the rate obtained in [18], without the strong assumption that we are working on a compact complex manifold. For the Sobolev space H^s, our construction improves on the rate $n^{1-\min(2s,1)/d}$ obtained in [8] when $s > 1/2$, and matches it when $s \leq 1/2$. In the special case $\phi = \phi_{\mathrm{Haar}}$, we note that Theorem 4.2 provides explicit *variance reduction guarantees for stratified sampling* for objective functions of *low regularity* ($s < 1$), which is potentially new and of independent interest.

5 Applications: quadrature and coreset construction

Linear statistics (1) abound in ML, so that a variety of results can be derived from the results of Sections 3 and 4. As examples, we show first how to derive faster-than-Monte-Carlo quadrature from Theorem 4, and then how to combine this fast quadrature with the continuous-to-discrete conversion in Section 3 to obtain more-accurate-than-independent coresets.

Quadrature using wavelet DPPs. Let $\omega : \mathbb{R}^d \to \mathbb{R}$ be a C^1 density with[3] $\mathrm{supp}(\omega) \Subset (0, 1)^d$. Consider approximating $\int f(\mathbf{x})\omega(\mathbf{x})\mathrm{d}\mathbf{x}$, where $f \in \mathcal{F}$ belongs to a prescribed class of integrable functions. Let K be the wavelet DPP kernel (8) for some j and ϕ, and $\mathcal{S}_{\mathrm{c}} = \{Y_1, , \ldots, Y_n\}$ be the

[3]Hereafter, $A \Subset B$ means that the closure of A is a compact subset of B.

corresponding DPP. For $f : \mathbb{R}^d \to \mathbb{R}$, it is natural to define

$$\mathcal{L}_{\mathcal{S}_c}(f) := \sum_{i=1}^{n} \frac{f(Y_i)\omega(Y_i)}{K(Y_i, Y_i)}, \tag{9}$$

noting that this is a linear statistic (1) of the DPP $\mathcal{S}_c$. In particular, one can deduce from (3) that (9) is an unbiased estimator of $\int f\omega \mathrm{d}x$.

In the special case $\phi = \phi_{\mathrm{Haar}}$, the kernel diagonal $x \mapsto K(x, x)$ is constant, and we prove in Proposition C.1 that the mean squared error of (9) scales as $O(n^{-1-2s/d})$ for f in either $C^{0,s}$ or H^s. In view of Proposition 4.1, this was somehow expected: indeed, we know since the early work of [11] that the MSE of the stratified sampling estimator is $O(n^{-1-2/d})$ for $f : [0, 1]^d \to \mathbb{R}$ continuously differentiable. As noted in [8], this rate outperforms the OPE rate $O(n^{-1-1/d})$ for the same kind of estimator in [3], though the choice of the cardinality of the Haar-wavelet-based quadrature is limited to the set $\{2^d, 2^{2d}, \dots\}$, making it impractical in large dimensions.

For a general scaling function, the choice of the cardinality is more flexible, but the diagonal $x \mapsto K(x, x)$ is not necessarily constant, and we have found it more mathematically tractable to modify the linear statistic (9) as follows. With the notation of Section 4, fix a design $\{\mathbf{x_k} : \mathbf{k} \in \mathcal{I}\}$ such that, for each $\mathbf{k} \in \mathcal{I}$, $\mathbf{x_k}$ is any fixed point in $\mathrm{supp}(\Phi_{-j,\mathbf{k}})$. We consider the estimator

$$\widetilde{\mathcal{L}}_{\mathcal{S}_c}(f) := \mathcal{L}_{\mathcal{S}_c}(f) - \frac{1}{2^{dj/2}} \sum_{i=1}^{n} \sum_{\mathbf{k}\in\mathcal{I}} \frac{\Phi_{-j,\mathbf{k}}(Y_i)}{K(Y_i, Y_i)} f(\mathbf{x_k})\omega(\mathbf{x_k}) + \frac{1}{2^{dj}} \sum_{\mathbf{k}\in\mathcal{I}} f(\mathbf{x_k})\omega(\mathbf{x_k}). \tag{10}$$

This new linear statistic is reminiscent of control variates estimators in Monte Carlo integration [25, Section 4.4.2]. At the price of a moderate increase of n to $2n$ in the number of required evaluations of f when changing from (9) to (10), our main quadrature result states that the MSE of the latter scales like that of the former, without any restriction on ϕ.

Theorem 5.1 (Fast quadrature for Hölder functions)**.** *Let $\mathcal{S}_c$ be the DPP on $(\mathbb{R}^d, \mathrm{d}\mathbf{x})$ defined by the kernel K in (8). Let $s \in (0, 1]$. Then for $f \in C^{0,s}$ satisfying (∗),*

$$\mathbb{E}\Big[\Big|\widetilde{\mathcal{L}}_{\mathcal{S}}(f) - \int f(\mathbf{x})\omega(\mathbf{x})\mathrm{d}\mathbf{x})\Big|^2\Big] = O(n^{-1-2s/d}).$$

Discretized Haar wavelet DPPs as coresets. Let $\mathcal{X}$ be a data set of N points, and let $\mathcal{F}$ be a family of test functions on $\mathcal{X}$. In ML, we think of $\mathcal{F}$ as indexed by predictors, and evaluating f at x gives the loss incurred by the predictor indexed by f at point x. For each $f \in \mathcal{F}$, define $L(f) := \sum_{x\in\mathcal{X}} f(x)$. When N is huge, it is natural to approximate $L(f)$ by a weighted sum of the form $L_{\mathcal{C}}(f) := \sum_{x\in\mathcal{C}} w(x)f(x)$, where $\mathcal{C} \subset \mathcal{X}$ is a subset of much smaller cardinality and $w : \mathcal{C} \to \mathbb{R}$ is a suitable weight function. Note that $L_{\mathcal{C}}(f) = \Lambda_{\mathcal{C}}(fw)$ is a linear statistic of the configuration $\mathcal{C}$, i.e. is of the form (1). Let $\varepsilon > 0$, we say that $\mathcal{C}$ is an ε-coreset for $\mathcal{X}$ (with respect to $\mathcal{F}$) if the relative error of $L_{\mathcal{C}}(f)$ is less than ε, uniformly in $\mathcal{F}$. For more on coresets, we refer to [2].

To derive a coreset guarantee from our results, we first assume that the data set is $\mathcal{X} = \{X_1, \dots, X_N\}$ consisting of i.i.d. samples from a density ρ such that $\rho \in C^1([0, 1]^d)$ and $\rho(\mathbf{x}) \geq \rho_{\min} > 0$ on $[0, 1]^d$. For brevity, we assume the test functions $f \in \mathcal{F}$ belong to $C^{0,s}$, the case of H^s being analogous. For simplicity, we also restrict to the kernel K in (8) with the *Haar* scaling function ϕ_{Haar}, which allows us to manipulate the simpler linear statistic (9) instead of (10).

Following Section 3, given the data set $\mathcal{X}$, we use the Haar wavelets to define (4) and (5), resulting in a DPP $\mathcal{S}_d$ with kernel $\mathbf{K}$ projecting onto the eigenvectors of (5). Note that, while Proposition 4.1 states that the Haar-wavelet DPP *is* stratified sampling on dyadic hypercubes, the discrete DPP $\mathcal{S}_d$ is a more sophisticated object that provides a *novel mechanism dedicated to the problem subset sampling from large datasets*. Finally, consider the estimator for $L(f)$ defined by

$$L_{\mathcal{S}_d}(f) := \sum_{x\in\mathcal{S}_d} \frac{f(x)}{\mathbb{P}(x \in \mathcal{S}_d|\mathcal{X})},$$

where $\mathbb{P}(x \in \mathcal{S}_d|\mathcal{X})$ is the diagonal element of the matrix $\mathbf{K}$ indexed by x.

Theorem 5.2 (Fast MSE for coresets)**.** *Let $f \in C^{0,s}$ satisfy (∗) and $\|f\|_\infty \leq 1, |f|_{C^{0,s}} \leq 1$. Then, given the data set $\mathcal{X}$, $L_{\mathcal{S}_d}(f)$ is an unbiased estimator of $L(f)$. Moreover, there exists a constant*

$C > 0$ depending only on d and ρ such that: for all $\delta, \delta' \in (0,1)$, for all N large enough such that $\widetilde{\mathscr{E}}(n, N, \delta, \delta') \leq n^{-1-2s/d}$, with probability at least $1 - \delta - \delta'$ in the data set $\mathcal{X}$, we have

$$\mathbb{E}\Big[\Big|\frac{L_{\mathcal{S}_\mathrm{d}}(f) - L(f)}{N}\Big|^2 \Big| \mathcal{X}\Big] \leq C n^{-1-2s/d},$$

where

$$\widetilde{\mathscr{E}}(n, N, \delta, \delta') = 4\sqrt{\frac{2\log(2/\delta')}{N}} + \frac{n+4}{9N^2}\log^2\Big(\frac{n^2+1}{\delta}\Big) + \frac{n+4}{N}\log\Big(\frac{n^2+1}{\delta}\Big). \tag{11}$$

Theorem 5.2 shows that, under mild regularity assumptions on f, the statistic $L_{\mathcal{S}_\mathrm{d}}(f)$ is an unbiased estimator of $L(f)$ with fast MSE decay. On the other hand, [5] develops a general machinery for discrete DPPs that converts such MSE decay estimates into coreset guarantees, i.e. to a guarantee of low mean square relative error *uniformly* in $\mathcal{F}$. This is the celebrated *Probably Approximately Correct* (PAC)-style guarantee that is well-suited to underpin practical ML procedures. In Appendix C.4, we apply this machinery to the bound in Theorem 5.2 and give an explicit coreset guarantee.

6 Implementation in High Dimensions via Hilbert's Space Filling Curve

In our tensor product construction of the wavelet DPP in section 4, the rank of the wavelet DPP kernel in (8) scales as $n \asymp 2^{dj}$, which restricts the choice of sample sizes in high dimensions. We address this limitation using Hilbert's space-filling curve; see, for example, [12]. Let $H_d : [0,1] \to [0,1]^d$ denote this curve, which is measure-preserving, injective almost everywhere, and lies in the space $C^{0,1/d}([0,1])$. In particular, for $f \in C^{0,s}([0,1]^d)$, $s \in (0,1]$, we have $f \circ H_d \in C^{0,s/d}([0,1])$, with $|f \circ H_d|_{C^{0,s/d}} \leq C_d^s |f|_{C^{0,s}}$. Consequently, applying Theorem 4.2 in dimension one to the linear statistic $\Lambda_{\mathcal{S}_c}(f \circ H_d)$, we obtain

$$\mathrm{Var}[\Lambda_{\mathcal{S}_c}(f \circ H_d)] \leq C |f|^2_{C^{0,s}} n^{1-2s/d}.$$

Thus, the variance is of the same order, while the cardinality of the wavelet-DPP can now be chosen as $n \asymp 2^j$, regardless of the dimension. We note that since H_d is measure-preserving, integrating $f \circ H_d$ on $[0,1]$ is equivalent to integrating f on $[0,1]^d$. Similarly, the guarantees of Theorem 5.1 transfer to this construction.

The discretization pipeline of Section 3 also applies without modification. Given a dataset $\mathcal{X} = \{X_1, \ldots, X_N\} \subset [0,1]^d$, we apply the construction (4)-(6) to $\{h_d(X_1), \ldots, h_d(X_N)\} \subset [0,1]$, where h_d denotes an almost-everywhere inverse of H_d. After sampling the discrete DPP in $d = 1$, we compose the sampled points with H_d, to select the corresponding points in the original dataset.

7 Experiments

We first extend an experiment on coresets for k-means from [28, 5]. Then we demonstrate DPP mini-batches in SGD applied to a non-smooth loss. In Appendix D, we provide quadrature experiments.[4]

7.1 A controlled experiment on k-means

We compare five methods. Given an input dataset $\mathcal{X}$ and an integer m, each method returns a random coreset of size m. The baseline is i.i.d. uniform sampling, and we implement four DPP samplers. The first DPP is the OPE built on the empirical measure of the dataset $\mathcal{X}$, denoted VDM-DPP as in the original reference [28]. The other three DPP samplers are built using our continuous-to-discrete pipeline of Section 3, namely Equation (6). We take $\hat{\rho}$ to be the kernel density estimator (KDE) built on $\mathcal{X}$ using the Epanechnikov kernel, with Scott's bandwidth selection method. The three methods differ by the underlying continuous DPP Kernel they use: OPE uses the multivariate OPE kernel of [3], HAAR uses the wavelet DPP kernel (8) with the Haar scaling function, and DB2 the wavelet DPP kernel with the popular[5] *periodized Daubechies-2* scaling function [9, 21]. Note that for this DB2 sampler, we use the adjusted linear statistic $\tilde{\mathcal{L}}_{\mathcal{S}_d}$ in (10) to estimate the full empirical loss.

[4] Our code will be made publicly available upon publication.

[5] For wavelets, we use the Python package PyWavelets [17], MIT Licence.

Our performance metric is $Q_{\mathcal{S}}(0.9)$, where $Q_{\mathcal{S}}$ is the quantile function of the relative error $\sup |L_{\mathcal{S}_d}(f) - L(f)|/|L(f)$. While evaluating the supremum of the relative error, we uniformly sample without replacement k elements of $\mathcal{X}$ 150 times, and we take the maximum value of the relative error among these values. Moreover, for each method and each coreset size m, the quantile is estimated by an empirical quantile over 150 independent coresets.

Results. We first conduct our experiment using a trimodal synthetic dataset of 1024 points supported on $[-1, 1]^2$. The dataset is drawn from a balanced Gaussian mixture on $\mathbb{R}^2$, with centers at $(\pm\sqrt{2}/2, \pm\sqrt{2}/2)$ and the origin. The second dataset is the MNIST dataset,[6] and reduced to $d = 2$ using PCA. The experiment is performed with $k = 3$ for the trimodal dataset, and with $k = 10$ for the MNIST dataset. Results are shown in Figure 7.1. For both datasets, the DPP samplers beat uniform sampling, as expected. In the case of MNIST, both HAAR and DB2 improve over the OPE samplers, with VDM-DPP slightly outperforming OPE. In the case of the trimodal dataset, the performance of HAAR regresses to that of the OPE samplers, while the improvement offered by DB2 persists. Overall, the experiment supports the recommendation to use wavelet-based samplers.

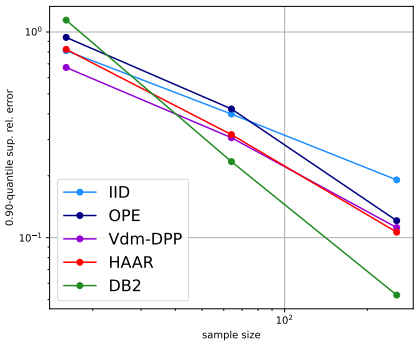

Figure 1: $Q_{\mathcal{S}}(0.9)$ vs. m: Trimodal Dataset

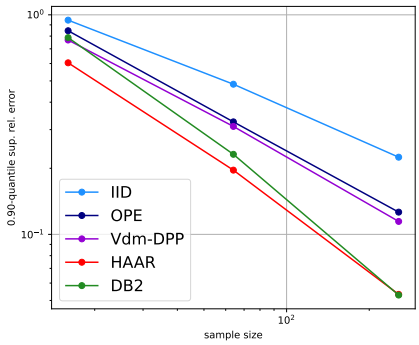

Figure 2: $Q_{\mathcal{S}}(0.9)$ vs. m: MNIST Dataset

7.2 Classification of MNIST with a non-smooth loss

To demonstrate the benefits of coupling wavelet-based DPPs and our discretization mechanism in minibatch sampling for non-smooth losses, we consider the *Pegasos* algorithm [26]. *Pegasos* is sub-gradient descent on the hinge loss, i.e. for basic SVMs. The full data sub-gradient at θ_t is given by

$$\nabla_t = \lambda\theta_t - \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{y_i\langle\theta_t, x_i\rangle < 1\} y_i x_i,$$

We compare minibatches $\mathcal{S}_d$ produced by four samplers: uniform sampling (IID), which we use as a baseline, and three DPP-based samplers built using (6) again, with $\hat{\rho}$ given by a Gaussian KDE on $\mathcal{X}$ with Scott's bandwidth. The DPP samplers only vary by the kernel of the underlying continuous DPP. The DPP built from the multivariate OPE kerenel is labeled as OPE, the wavelet-based kernel with the Haar scaling function yields HAAR, and DB2 corresponds to the periodized Daubechies-2 scaling function. With the DB2 sampler, we again use the adjusted linear statistic (10).

Results. We use the (balanced) dataset $\mathcal{X}$ of $N = 10000$, of 4s and 9s in MNIST, which we reduce to $d = 5$ and $d = 10$ using PCA. We use a train-test split of 70:30. For each sampler, we form two minibatches of size $m = 16$, one from each class. This gives a total minibatch size of $m = 32$. For the wavelet-based samplers HAAR, and DB2, we use the Hilbert's space filling curve approach to sample points and compute the relevant linear statistic. Pegasos is run for $T = 50$ iterations with regularization $\lambda = 0.1$ and step size $\eta_t = 1/(\lambda t)$. Our performance metrics are the test classification error, the full-batch sub-gradient norm $\|\nabla f(\theta_t)\|_2$, and the parameter-error $\|\theta_t - \theta^\star\|_2$. The reference solution $\theta^\star$ is obtained using Scikit-learn's linear SVM solver `sklearn.svm.SVC` [24].

[6] made available under the terms of the Creative Commons Attribution-Share Alike 3.0 license.

The performance metrics are averaged over 100 independent trials for each sampler, and the error bars are for ± 2 standard errors.

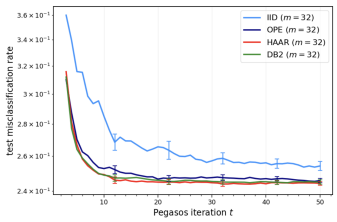

Figure 3: $d = 5$:Test Error

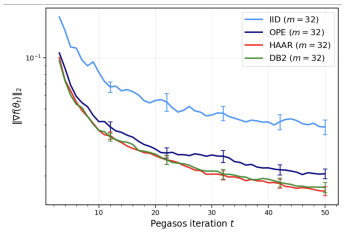

Figure 4: $d = 5 : \|\nabla f(\theta_t)\|_2$

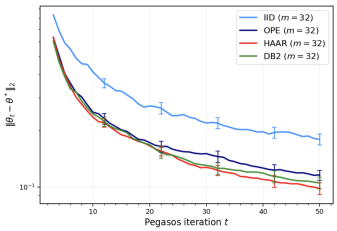

Figure 5: $d = 5 : \|\theta_t - \theta^*\|_2$

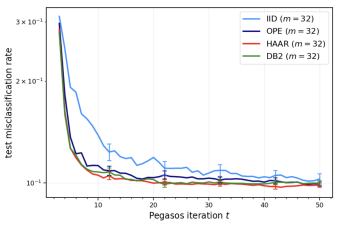

Figure 6: $d = 10$:Test Error

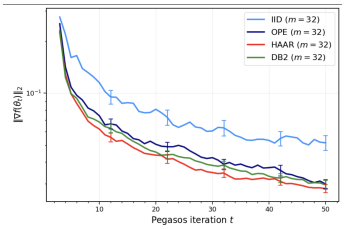

Figure 7: $d = 10 : \|\nabla f(\theta_t)\|_2$

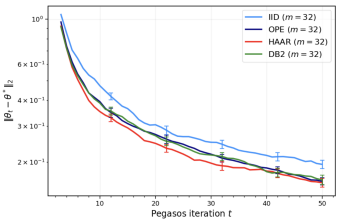

Figure 8: $d = 10 : \|\theta_t - \theta^*\|_2$

The DPP samplers offer considerable improvement over the iid baseline on all metrics. Moreover, both wavelet DPPs also outperform OPE.

8 Discussion

This work contributes to the programme of developing DPPs as a subsampling toolbox for ML in multiple ways. First, we propose new DPPs on the Euclidean space based on wavelets, with provably better accuracy guarantees than the best known rates. In this vein, we are able to address classes of highly irregular functions, with a view to ML applications with non-smooth objective functions (e.g. coresets with non-differentiable loss functions) and rough data classes (such as financial data). Second, we introduce a general pipeline to convert such continuous DPPs, which are more amenable to proving analytical estimates, into discrete kernels, which are pertinent for subsampling tasks such as minibatch and coreset constructions. This very general conversion pipeline, acting on an input continuous kernel in a completely turn-key fashion, simultaneously preserves the desired variance decay and automatically reveals a low-rank decomposition of the discrete kernel, which makes sampling the corresponding DPP computationally inexpensive. This work also raises natural questions for further investigation – principal among them being how much sharper one can push the variance reduction properties of determinantal samplers through more efficient kernel designs. An associated question is that of improving the dependence of the rates on the effective dimension. Our results on non-smooth objective functions raise natural questions on optimal dependencies of variance reduction rates on their regularity. Finally, and more specific to the wavelet-based DPP, it would also be practically relevant to decouple the cardinality of the DPP from the dimension.

Acknowledgments and Disclosure of Funding

HST was supported by the Singapore MOE grants A-8003576-00-00 and A-8002488-00-00. RB was supported by ERC grant *Blackjack* ERC-2019-STG-851866 and ANR grant *Baccarat* ANR-20-CHIA-0002. SG was supported in part by the NUS Dean's Chair Associate Professorship E-146-00-0037-01 and the Singapore MOE grants A-8002014-00-00 and A-8003802-00-00.

References

[1] G. Acosta and J.P. Borthagaray. A fractional Laplace equation: regularity of solutions and finite element approximations. *SIAM Journal on Numerical Analysis*, 55(2):472–495, 2017.

[2] O. Bachem, M. Lucic, and A. Krause. Practical coreset constructions for machine learning. *arXiv preprint arXiv:1703.06476*, 2017.

[3] R. Bardenet and A. Hardy. Monte Carlo with determinantal point processes. *The Annals of Applied Probability*, 30(1):368 – 417, 2020. doi: 10.1214/19-AAP1504. URL https://doi.org/10.1214/19-AAP1504.

[4] R. Bardenet, S. Ghosh, and M. Lin. Determinantal point processes based on orthogonal polynomials for sampling minibatches in sgd. *Advances in Neural Information Processing Systems*, 34:16226–16237, 2021.

[5] R. Bardenet, S. Ghosh, H. Simon-Onfroy, and H.S. Tran. Small coresets via negative dependence: Dpps, linear statistics, and concentration. *Advances in Neural Information Processing Systems*, 37:84329–84349, 2024.

[6] Simon Barthelmé, Nicolas Tremblay, and Pierre-Olivier Amblard. A faster sampler for discrete determinantal point processes. In *International Conference on Artificial Intelligence and Statistics*, pages 5582–5592. PMLR, 2023.

[7] A. Belhadji, R. Bardenet, and P. Chainais. Kernel quadrature with determinantal point processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[8] J.-F. Coeurjolly, A. Mazoyer, and P.-O. Amblard. Monte Carlo integration of non-differentiable functions on $[0,1]^\iota$, $\iota = 1, \ldots, d$, using a single determinantal point pattern defined on $[0,1]^d$. *arXiv preprint arXiv:2003.10323*, 2020.

[9] I. Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, 1992. doi: 10.1137/1.9781611970104. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611970104.

[10] J. Gillenwater, A. Kulesza, Z. Mariet, and S. Vassilvtiskii. A tree-based method for fast repeated sampling of determinantal point processes. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2260–2268. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/gillenwater19a.html.

[11] S. Haber. A modified monte-carlo quadrature. *Mathematics of Computation*, 20(95):361–368, 1966.

[12] Zhijian He and Art B. Owen. Extensible grids: uniform sampling on a space-filling curve. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(4):917–931, 2016.

[13] J. B. Hough, M. Krishnapur, Y. Peres, and B. Virág. Determinantal processes and independence. *Probability surveys*, 2006.

[14] H. Jaquard and N. Keriven. Statistical consistency of discrete-to-continuous limits of determinantal point processes, 2026. URL https://arxiv.org/abs/2603.01670.

[15] H. Jiang. Uniform convergence rates for kernel density estimation. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1694–1703. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/jiang17b.html.

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

[17] G. R. Lee, R. Gommers, F. Waselewski, K. Wohlfahrt, and A. O'Leary. Pywavelets: A python package for wavelet analysis. *Journal of Open Source Software*, 4(36):1237, 2019. doi: 10.21105/joss.01237. URL https://doi.org/10.21105/joss.01237.

[18] T. Lemoine and R. Bardenet. Monte Carlo methods on compact complex manifolds using Bergman kernels. *arXiv preprint arXiv:2405.09203*, 2024.

[19] O. Macchi. *Processus ponctuels et coincidences – Contributions à l'étude théorique des processus ponctuels, avec applications à l'optique statistique et aux communications optiques*. PhD thesis, Université Paris-Sud, 1972.

[20] O. Macchi. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7:83–122, 1975.

[21] S. Mallat. *A wavelet tour of signal processing*. Elsevier, third edition, 2008.

[22] C. McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1): 148–188, 1989.

[23] L. E. Payne and H. F. Weinberger. An optimal poincaréinequality for convex domains. *Archive for Rational Mechanics and Analysis*, 5(1):286–292, 1960. doi: 10.1007/BF00252910. URL https://doi.org/10.1007/BF00252910.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[25] C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer, 2004.

[26] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *Proceedings of the 24th international conference on Machine learning*, pages 807–814, 2007.

[27] H.S. Tran, V. Petrovic, R. Bardenet, and S. Ghosh. Negative dependence as a toolbox for machine learning: review and new developments. *arXiv preprint arXiv:2502.07285*, 2025.

[28] N. Tremblay, S. Barthelmé, and P.-O. Amblard. Determinantal point processes for coresets. *Journal of Machine Learning Research*, 20(168):1–70, 2019.

[29] N. Tremblay, S. Barthelmé, K. Usevich, and P.-O. Amblard. Extended l-ensembles: a new representation for determinantal point processes. *The Annals of Applied Probability*, 33(1): 613–640, 2023.

[30] J. Tropp. An introduction to matrix concentration inequalities. *Foundations and trends® in machine learning*, 8(1-2):1–230, 2015.

A Technical appendices to discrete-to-continuous DPPs

In this section, we provide additional comments as well as the detailed proofs of our theorems in Section 3.

A.1 Additional comments on Theorem 3.2

On the order of $K_{\max}$ in (A1). Since $\int K(x,x)\mu(\mathrm{d}x) = n$, we have $K_{\max} \geq n$. In many examples of interest, one in fact has $K_{\max} \asymp n$. For instance, in orthogonal polynomial ensembles, if $\mathrm{supp}(\rho)$ is contained in a compact subset of the bulk and the reference measure μ is regular enough so that uniform bulk Christoffel asymptotics hold there, then $\sup_{x\in\mathrm{supp}(\rho)} K(x,x) \asymp n$ (see [3], Theorem A.4); in addition, if ρ is bounded below on its support, we will have $K_{\max} \asymp n$. Another family of examples is our novel class of wavelet DPPs introduced in Section 4. We will see that $K_{\max} \asymp n$ in these examples as well.

On the minibatch size. As remarked earlier, our algorithm produces a minibatch $\mathcal{S}_{\mathrm{d}}$ of size $m \leq n$, rather than one that is necessarily of size exactly n. However, we note that if N is sufficiently large so that $K_{\max}^2 \mathscr{E}(n, N, \delta, \delta') < 1$ (which is the relevant regime in practice), then on the same event appearing in the conclusion of Theorem 3.2, the minibatch has size exactly n. This follows from our proof technique; see Lemma A.4.

On explicit error terms. A notable feature of our result is that the error term $\mathscr{E}(n, N, \delta, \delta')$ is completely explicit and contains no hidden constants, which is particularly useful for quantitative analysis. It also distinguishes our result from many existing works in the literature, which often provide only asymptotic guarantees or estimates with implicit constants. In particular, existing discrete-to-continuous results typically treat the regime where $N \to \infty$ while the minibatch size n remains fixed. By contrast, our framework allows both the data set size N and the minibatch size n to grow simultaneously, with an explicit control of the corresponding error terms.

On the regime $N \gg n$. Theorem 3.2 establishes that our minibatch sampling algorithm satisfies a *variance-transfer guarantee*: the conditional variance of the linear statistic associated with the minibatch is bounded above and below by that of the underlying continuous DPP, up to *explicit* error terms. Therefore, our algorithm does not merely produce a repulsive minibatch; it approximately preserves the key second-order quantity (variance of linear statistics) of the underlying continuous DPPs, which is useful in sampling.

In the regime most relevant for applications, where the full data size N is much larger than the target minibatch size n, these error terms are negligible. Indeed, we observe that $K_{\max} \asymp n$ and $\mathscr{E}(n, N, \delta, \delta') = O(N^{-1/2})$ for large N. Thus, the error term is of order $O(n^2/N)$ in the expectation bound, and it is of order $O(n^2/\sqrt{N})$ in the high probability bound. Therefore, in the large N regime, the dominant term is $\mathrm{Var}[\Lambda_{\mathcal{S}_c}(f)]$. By choosing a continuous DPP kernel K that is known to enjoy variance reduction (i.e., $\mathrm{Var}[\Lambda_{\mathcal{S}_c}(f)] = O(n^{1-\alpha})$ for some $\alpha > 0$), the resulting minibatch $\mathcal{S}_{\mathrm{d}}$ automatically inherits this variance reduction property.

A.2 Proofs of results in Section 3

For simplicity of notation, we write the proofs in the real-valued case, namely, when ψ_k's are real-valued. The complex-valued case follows by the same argument after systematically replacing transposes by Hermitian adjoints and inserting complex conjugates where appropriate.

We set up some notations. Let $f : E \to \mathbb{R}$ be a bounded test function, we define

$$F(X_1, \dots, X_N) := \frac{1}{N^2} \sum_{1 \le i,j \le N} |f(X_i) - f(X_j)|^2 \frac{|K(X_i, X_j)|^2}{\rho(X_i)\rho(X_j)}. \tag{12}$$

For a function φ on $\mathcal{X}$, we set

$$\mathbf{D}(\varphi) := \mathrm{diag}(\varphi(X_1), \dots, \varphi(X_N))$$

where $\mathrm{diag}(a_1, \dots, a_N)$ denotes the $N \times N$-diagonal matrix with the diagonal entries are given by $(a_1, \dots, a_N)$. For a square matrix $\mathbf{A}$, we denote by $\|\mathbf{A}\|_{\mathrm{F}}$ the Frobenius norm of $\mathbf{A}$.

For $n, N \in \mathbb{N}$ and $\delta, \delta' \in (0, 1)$, we define

$$\mathscr{E}(n, N, \delta, \delta') = 4\sqrt{\frac{2\log(2/\delta')}{N}} + \frac{4}{9N^2} \log^2\Big(\frac{n^2+1}{\delta}\Big) + \frac{4}{N} \log\Big(\frac{n^2+1}{\delta}\Big). \tag{13}$$

A.2.1 Some technical lemmas

Lemma A.1 (Structure of r-DPP). *Let $[N] = \{1, 2, \dots, N\}$ and $r \le N$. Let $\mathcal{Y}$ be an L-ensemble on $[N]$ with an L-kernel $\mathbf{L}$ of rank r*

$$\mathbf{L} = \sum_{i=1}^{r} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top, \quad \lambda_i > 0,$$

where $\{\mathbf{v}_i : 1 \le i \le r\}$ is orthonormal in $\mathbb{R}^N$. Then:

(i) The r-DPP defined by $\mathbf{L}$, denoted by $\mathcal{Y}_{r\text{-DPP}}$, is also a projection DPP on $[N]$ with kernel given by

$$\mathbf{K} = \sum_{i=1}^{r} \mathbf{v}_i \mathbf{v}_i^\top.$$

(ii) Moreover, we have

$$|\mathbf{L}_{jj} - \mathbf{K}_{jj}| \leq \Big(\max_{1\leq i\leq r} |\lambda_i - 1| \Big) \mathbf{K}_{jj}, \quad \forall 1 \leq j \leq N.$$

Proof. (i) Let $J \subset [N]$ with $|J| = r$, we have the following formula:

$$\mathbb{P}(\mathcal{Y}_{r\text{-DPP}} = J) = \frac{1}{\lambda_1 \dots \lambda_r} \det[\mathbf{L} + \mathbf{I}] \mathbb{P}(\mathcal{Y} = J) = \prod_{i=1}^{r} \frac{\lambda_i + 1}{\lambda_i} \mathbb{P}(\mathcal{Y} = J).$$

Since $\mathcal{Y}$ is an L-ensemble on $[N]$ with the kernel $\mathbf{L}$, $\mathcal{Y}$ is also a DPP on $[N]$ with the kernel given by

$$\mathbf{L}(\mathbf{I} + \mathbf{L})^{-1} = \sum_{i=1}^{r} \frac{\lambda_i}{\lambda_i + 1} \mathbf{v}_i \mathbf{v}_i^\top.$$

Thus, $|\mathcal{Y}| \leq r$ almost surely. In particular, for any $J \subset [N]$ with $|J| = r$, we must have

$$\mathbb{P}(\mathcal{Y} = J) = \mathbb{P}(\mathcal{Y} \supset J).$$

Thus

$$\mathbb{P}(\mathcal{Y}_{r\text{-DPP}} = J) = \Big(\prod_{i=1}^{r} \frac{\lambda_i + 1}{\lambda_i} \Big) \cdot \det \Big[\Big(\sum_{i=1}^{r} \frac{\lambda_i}{\lambda_i + 1} \mathbf{v}_i \mathbf{v}_i^\top \Big)_{J \times J} \Big].$$

We now claim that for every $J \subset [N]$ with $|J| = r$

$$\Big(\prod_{i=1}^{r} \frac{\lambda_i + 1}{\lambda_i} \Big) \cdot \det \Big[\Big(\sum_{i=1}^{r} \frac{\lambda_i}{\lambda_i + 1} \mathbf{v}_i \mathbf{v}_i^\top \Big)_{J \times J} \Big] = \det \Big[\Big(\sum_{i=1}^{r} \mathbf{v}_i \mathbf{v}_i^\top \Big)_{J \times J} \Big]. \tag{14}$$

Indeed, let

$$\mathbf{D} := \operatorname{diag}\Big(\frac{\lambda_1}{\lambda_1 + 1}, \dots, \frac{\lambda_r}{\lambda_r + 1} \Big) \in \mathbb{R}^{r \times r} \quad , \quad \mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_r] \in \mathbb{R}^{N \times r}.$$

Then

$$\sum_{i=1}^{r} \frac{\lambda_i}{\lambda_i + 1} \mathbf{v}_i \mathbf{v}_i^\top = \mathbf{V} \mathbf{D} \mathbf{V}^\top = (\mathbf{D}^{1/2} \mathbf{V}^\top)^\top (\mathbf{D}^{1/2} \mathbf{V}^\top) \in \mathbb{R}^{N \times N}.$$

For $J \subset [N]$ with $|J| = r$, let $(\mathbf{D}^{1/2} \mathbf{V}^\top)_J$ denote the $r \times r$-submatrix of $\mathbf{D}^{1/2} \mathbf{V}^\top$ whose columns are indexed by J (note that $\mathbf{D}^{1/2} \mathbf{V}^\top \in \mathbb{R}^{r \times N}$). Then it is clear that

$$\Big(\sum_{i=1}^{r} \frac{\lambda_i}{\lambda_i + 1} \mathbf{v}_i \mathbf{v}_i^\top \Big)_{J \times J} = ((\mathbf{D}^{1/2} \mathbf{V}^\top)_J)^\top (\mathbf{D}^{1/2} \mathbf{V}^\top)_J.$$

Taking determinants both sides gives

$$\begin{aligned} \det \Big(\sum_{i=1}^{r} \frac{\lambda_i}{\lambda_i + 1} \mathbf{v}_i \mathbf{v}_i^\top \Big)_{J \times J} &= \Big(\det(\mathbf{D}^{1/2} \mathbf{V}^\top)_J \Big)^2 = \det[\mathbf{D}] \cdot \det[\mathbf{V} \mathbf{V}^\top]_{J \times J} \\ &= \Big(\prod_{i=1}^{r} \frac{\lambda_i}{\lambda_i + 1} \Big) \det \Big(\sum_{i=1}^{r} \mathbf{v}_i \mathbf{v}_i^\top \Big)_{J \times J}. \end{aligned}$$

as desired. The claim follows.

Therefore, $\mathbb{P}(\mathcal{Y}_{r\text{-DPP}} = J) = \det[\mathbf{K}_{J \times J}]$ where $\mathbf{K} := \sum_{i=1}^{r} \mathbf{v}_i \mathbf{v}_i^\top$ and $\mathbf{K}_{J \times J}$ is the submatrix of $\mathbf{K}$ whose rows and columns indexed by J. Since $\mathbf{K}$ is a projection matrix, $\mathcal{Y}_{r\text{-DPP}}$ is a projection DPP with kernel $\mathbf{K}$ as desired.

(ii) For each $1 \leq i \leq r$, we write $\mathbf{v}_i = (\mathbf{v}_{i1}, \dots, \mathbf{v}_{iN})^\top$. Then we have

$$\mathbf{L}_{jj} = \sum_{i=1}^{r} \lambda_i \mathbf{v}_{ij}^2 \quad , \quad \mathbf{K}_{jj} = \sum_{i=1}^{r} \mathbf{v}_{ij}^2 \qquad \forall 1 \leq j \leq N.$$

This implies

$$|\mathbf{L}_{jj} - \mathbf{K}_{jj}| \leq \sum_{i=1}^{r} |\lambda_i - 1| \mathbf{v}_{ij}^2 \leq \Big(\max_{1\leq i\leq r} |\lambda_i - 1| \Big) \sum_{i=1}^{r} \mathbf{v}_{ij}^2 = \Big(\max_{1\leq i\leq r} |\lambda_i - 1| \Big) \mathbf{K}_{jj}$$

as desired. □

Lemma A.2. *Let F be defined as in Equation (12). We have*

$$\mathbb{E}[F(X_1,\dots,X_N)] = \frac{2(N-1)}{N}\mathrm{Var}[\Lambda_{\mathcal{S}_c}(f)].$$

Moreover, for every $t \geq 0$

$$\mathbb{P}(|F(X_1,\dots,X_N) - \mathbb{E}[F(X_1,\dots,X_N)]| \geq t) \leq 2\exp\Big(-\frac{Nt^2}{32\|f\|_\infty^4 K_{\max}^4}\Big).$$

In particular, for every $\delta' \in (0,1)$, with probability at least $1-\delta'$ we have

$$|F(X_1,\dots,X_N) - \mathbb{E}[F(X_1,\dots,X_N)]| \leq 4\|f\|_\infty^2 K_{\max}^2\sqrt{\frac{2\log(2/\delta')}{N}}.$$

Proof. We have

$$\begin{aligned}\mathbb{E}[F(X_1,\dots,X_N)] &= \frac{N-1}{N}\mathbb{E}\Big[|f(X_1)-f(X_2)|^2\frac{|K(X_1,X_2)|^2}{\rho(X_1)\rho(X_2)}\Big]\\ &= \frac{N-1}{N}\int |f(x)-f(y)|^2|K(x,y)|^2\mu(\mathrm{d}x)\mu(\mathrm{d}y)\\ &= \Big(2-\frac{2}{N}\Big)\mathrm{Var}[\Lambda_{\mathcal{S}_c}(f)].\end{aligned}$$

For every $1 \leq i \leq N$, we have

$$\begin{aligned}&|F(x_1,\dots,x_i,\dots,x_N) - F(x_1,\dots,x_i',\dots,x_N)|\\ \leq\ & \frac{2}{N^2}\sum_{j=1}^N 4\|f\|_\infty^2 \max\Big(\frac{|K(x_i,x_j)|^2}{\rho(x_i)\rho(x_j)}, \frac{|K(x_i',x_j)|^2}{\rho(x_i')\rho(x_j)}\Big) \leq \frac{8\|f\|_\infty^2 K_{\max}^2}{N},\end{aligned}$$

where we used that $|K(x,y)|^2 \leq K(x,x)K(y,y)$. By the McDiarmid's inequality for functionals of independent random variables satisfying the bounded differences property (see [22]), we have for every $t \geq 0$

$$\mathbb{P}(|F(X_1,\dots,X_N) - \mathbb{E}[F(X_1,\dots,X_N)]| \geq t) \leq 2\exp\Big(-\frac{Nt^2}{32\|f\|_\infty^4 K_{\max}^4}\Big).$$

Let $\delta' = 2\exp\Big(-\frac{Nt^2}{32\|f\|_\infty^4 K_{\max}^4}\Big)$, we get

$$t = \|f\|_\infty^2 K_{\max}^2\sqrt{\frac{32\log(2/\delta')}{N}}$$

as desired. □

Lemma A.3 (Structures of $\mathbf{L}$ and $\widehat{\mathbf{L}}$)**.** *For $1 \leq i,j \leq N$, we have*

$$\mathbf{L}_{ij} = \frac{K(X_i,X_j)}{N\sqrt{\rho(X_i)\rho(X_j)}} \quad , \quad \widehat{\mathbf{L}}_{ij} = \frac{K(X_i,X_j)}{N\sqrt{\hat\rho(X_i)\hat\rho(X_j)}}.$$

Proof. Since $K(x,y) = \sum_{k=1}^n \psi_k(x)\psi_k(y)$, we have

$$\begin{pmatrix} K(X_1,X_1) & \dots & K(X_1,X_N)\\ \vdots & \ddots & \vdots\\ K(X_N,X_1) & \dots & K(X_N,X_N)\end{pmatrix} = \begin{pmatrix} \psi_1(X_1) & \dots & \psi_n(X_1)\\ \vdots & \ddots & \vdots\\ \psi_1(X_N) & \dots & \psi_n(X_N)\end{pmatrix}\begin{pmatrix} \psi_1(X_1) & \dots & \psi_n(X_1)\\ \vdots & \ddots & \vdots\\ \psi_1(X_N) & \dots & \psi_n(X_N)\end{pmatrix}^\top.$$

Since $\mathbf{L} = N^{-1}\mathbf{D}(\rho^{-1/2})\mathbf{\Psi}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1/2})$, we deduce that

$$\mathbf{L}_{ij} = N^{-1}\rho(X_i)^{-1/2}K(X_i,X_j)\rho(X_j)^{-1/2}.$$

A similar argument applies to $\widehat{\mathbf{L}}$. The lemma follows. □

Lemma A.4. *We have*

$$\mathbb{E}\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}^2\le\frac{nK_{\max}}{N}.$$

Moreover, for any $t\ge 0$, we have

$$\mathbb{P}\Big(\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}\ge t\Big)\le(n^2+1)\exp\Big(-\frac{Nt^2}{2K_{\max}(n+t/3)}\Big).$$

In particular, for every $\delta\in(0,1)$, with probability at least $1-\delta$ we have

$$\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}^2\le 4K_{\max}^2\Big(\frac{1}{9N^2}\log^2\Big(\frac{n^2+1}{\delta}\Big)+\frac{1}{N}\log\Big(\frac{n^2+1}{\delta}\Big)\Big).$$

Proof. We define the following random vectors

$$Y_i=\Big(\psi_1(X_i)/\sqrt{\rho(X_i)},\dots,\psi_n(X_i)/\sqrt{\rho(X_i)}\Big)^\top,\quad 1\le i\le N.$$

Then it is clear that

$$\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}=\frac{1}{N}\sum_{i=1}^N Y_iY_i^\top.$$

We have $\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n=\frac{1}{N}\sum_{i=1}^N(Y_iY_i^\top-\mathbf{I}_n)$. It is easy to see that

$$\begin{aligned}\|Y_iY_i^\top-\mathbf{I}_n\|_{\mathrm{F}}^2&=\sum_{k=1}^n\Big(\frac{\psi_k(X_i)^2}{\rho(X_i)}-1\Big)^2+2\sum_{1\le k<j\le n}\frac{\psi_k(X_i)^2}{\rho(X_i)}\frac{\psi_j(X_i)^2}{\rho(X_i)}\\&=\Big(\sum_{k=1}^n\frac{\psi_k(X_i)^2}{\rho(X_i)}\Big)^2-2\sum_{k=1}^n\frac{\psi_k(X_i)^2}{\rho(X_i)}+n\\&=\Big(\frac{K(X_i,X_i)}{\rho(X_i)}\Big)^2-2\frac{K(X_i,X_i)}{\rho(X_i)}+n.\end{aligned}$$

This particularly implies $\|Y_iY_i^\top-\mathbf{I}_n\|_{\mathrm{F}}\le K_{\max}$ almost surely for every $1\le i\le N$. Moreover, we have

$$\mathbb{E}\|Y_iY_i^\top-\mathbf{I}_n\|_{\mathrm{F}}^2\le nK_{\max}.$$

Since $Y_1,\dots,Y_N$ are independent, we have

$$\mathbb{E}\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}^2=\frac{1}{N^2}\sum_{i=1}^N\mathbb{E}\|Y_iY_i^\top-\mathbf{I}_n\|_{\mathrm{F}}^2\le\frac{nK_{\max}}{N}$$

as desired.

Now observe that $\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n=\frac{1}{N}\sum_{i=1}^N(Y_iY_i^\top-\mathbf{I}_n)$ and Y_i's are independent. By treating $Y_iY_i^\top-\mathbf{I}_n$ as a vector in $\mathbb{R}^{n^2+1}$ and applying Bernstein inequality for vector-valued random variables ([30], Theorem 1.6.2), we have for every $t\ge 0$

$$\mathbb{P}\Big(\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}\ge t\Big)\le(n^2+1)\exp\Big(-\frac{Nt^2}{2K_{\max}(n+t/3)}\Big).$$

Let

$$\delta=(n^2+1)\exp\Big(-\frac{Nt^2}{2K_{\max}(n+t/3)}\Big).$$

Solving for $t\ge 0$ we have

$$t=\frac{K_{\max}}{3N}\log\Big(\frac{n^2+1}{\delta}\Big)+\sqrt{\frac{K_{\max}^2}{9N^2}\log^2\Big(\frac{n^2+1}{\delta}\Big)+\frac{2nK_{\max}}{N}\log\Big(\frac{n^2+1}{\delta}\Big)}.$$

Using $(a+b)^2 \leq 2a^2 + 2b^2$, we have

$$\begin{aligned} t^2 &\leq 2\frac{K_{\max}^2}{9N^2}\log^2\Big(\frac{n^2+1}{\delta}\Big) + 2\frac{K_{\max}^2}{9N^2}\log^2\Big(\frac{n^2+1}{\delta}\Big) + \frac{4nK_{\max}}{N}\log\Big(\frac{n^2+1}{\delta}\Big) \\ &= \frac{4K_{\max}^2}{9N^2}\log^2\Big(\frac{n^2+1}{\delta}\Big) + \frac{4nK_{\max}}{N}\log\Big(\frac{n^2+1}{\delta}\Big) \\ &\leq 4K_{\max}^2\Big(\frac{1}{9N^2}\log^2\Big(\frac{n^2+1}{\delta}\Big) + \frac{1}{N}\log\Big(\frac{n^2+1}{\delta}\Big)\Big) \end{aligned}$$

where we used $K_{\max} \geq n$ in the last inequality. Therefore, with probability at least $1-\delta$ we have

$$\Big\|\frac{1}{N}\mathbf{\Psi}^\top \mathbf{D}(\rho^{-1})\mathbf{\Psi} - \mathbf{I}_n\Big\|_{\mathrm{F}}^2 \leq 4K_{\max}^2\Big(\frac{1}{9N^2}\log^2\Big(\frac{n^2+1}{\delta}\Big) + \frac{1}{N}\log\Big(\frac{n^2+1}{\delta}\Big)\Big).$$

□

A direct consequence of Lemma A.2 and Lemma A.4 is the following:

Corollary A.5. *With probability at least $1-\delta-\delta'$, we have*

$$|F - \mathbb{E}[F]| + \|f\|_\infty^2\Big\|\frac{1}{N}\mathbf{\Psi}^\top \mathbf{D}(\rho^{-1})\mathbf{\Psi} - \mathbf{I}_n\Big\|_{\mathrm{F}}^2 \leq \|f\|_\infty^2 K_{\max}^2 \mathscr{E}(n,N,\delta,\delta')$$

where $\mathscr{E}(n,N,\delta,\delta')$ defined as in (7), namely

$$\mathscr{E}(n,N,\delta,\delta') = 4\sqrt{\frac{2\log(2/\delta')}{N}} + \frac{4}{9N^2}\log^2\Big(\frac{n^2+1}{\delta}\Big) + \frac{4}{N}\log\Big(\frac{n^2+1}{\delta}\Big).$$

A.2.2 Proof of Theorem 3.2

Lemma A.6. *Given the data set $\mathcal{X}$, for every bounded test function $f : E \to \mathbb{R}$, we have*

$$\mathrm{Var}[\Lambda_{\mathcal{S}_{\mathrm{d}}}(f)|\mathcal{X}] \leq F(X_1,\ldots,X_N) + 4\|f\|_\infty^2\Big\|\frac{1}{N}\mathbf{\Psi}^\top \mathbf{D}(\rho^{-1})\mathbf{\Psi} - \mathbf{I}_n\Big\|_{\mathrm{F}}^2$$

and

$$\mathrm{Var}[\Lambda_{\mathcal{S}_{\mathrm{d}}}(f)|\mathcal{X}] \geq \frac{1}{4}F(X_1,\ldots,X_N) - 2\|f\|_\infty^2\Big\|\frac{1}{N}\mathbf{\Psi}^\top \mathbf{D}(\rho^{-1})\mathbf{\Psi} - \mathbf{I}_n\Big\|_{\mathrm{F}}^2.$$

Proof. By Lemma A.1, $\mathcal{S}_{\mathrm{d}}$ is a DPP on $\mathcal{X}$ with a projection kernel $\mathbf{K} = [\mathbf{K}_{ij}]$. Then we have

$$\mathrm{Var}[\Lambda_{\mathcal{S}_{\mathrm{d}}}(f)|\mathcal{X}] = \frac{1}{2}\sum_{i,j\in[N]} |f(X_i) - f(X_j)|^2 \mathbf{K}_{ij}^2.$$

Using the inequality $b^2 \leq 2a^2 + 2(a-b)^2$, we have

$$\mathbf{K}_{ij}^2 \leq 2\mathbf{L}_{ij}^2 + 2(\mathbf{K}_{ij} - \mathbf{L}_{ij})^2, \quad \forall i,j.$$

Thus

$$\begin{aligned} \mathrm{Var}[\Lambda_{\mathcal{S}_{\mathrm{d}}}(f)|\mathcal{X}] &\leq \sum_{i,j\in[N]} |f(X_i) - f(X_j)|^2 \mathbf{L}_{ij}^2 + \sum_{i,j\in[N]} |f(X_i) - f(X_j)|^2 (\mathbf{K}_{ij} - \mathbf{L}_{ij})^2 \\ &\leq \sum_{i,j\in[N]} |f(X_i) - f(X_j)|^2 \mathbf{L}_{ij}^2 + 4\|f\|_\infty^2 \|\mathbf{L} - \mathbf{K}\|_{\mathrm{F}}^2 \\ &= \frac{1}{N^2}\sum_{i,j\in[N]} |f(X_i) - f(X_j)|^2 \frac{K(X_i,X_j)^2}{\rho(X_i)\rho(X_j)} + 4\|f\|_\infty^2 \|\mathbf{L} - \mathbf{K}\|_{\mathrm{F}}^2 \\ &= F(X_1,\ldots,X_N) + 4\|f\|_\infty^2 \|\mathbf{L} - \mathbf{K}\|_{\mathrm{F}}^2. \end{aligned}$$

By Lemma A.1, $\mathbf{K}$ shares the same eigenvectors with $\mathbf{L}$. Thus

$$\|\mathbf{L} - \mathbf{K}\|_{\mathrm{F}}^2 = \sum_{\lambda(\mathbf{L})>0} (\lambda(\mathbf{L}) - 1)^2.$$

Moreover, $\mathbf{L} = N^{-1}\mathbf{D}(\rho^{-1/2})\mathbf{\Psi}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1/2})$ implies nonzero eigenvalues of $\mathbf{L}$ are the same as nonzero eigenvalues of $N^{-1}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}$. Hence

$$\sum_{\lambda(\mathbf{L})>0}(\lambda(\mathbf{L})-1)^2 \le \Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}^2.$$

Therefore

$$\mathrm{Var}[\Lambda_{\mathcal{S}_\mathrm{d}}(f)|\mathcal{X}] \le F(X_1,\dots,X_N) + 4\|f\|_\infty^2\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}^2.$$

Using the inequality $b^2 \ge \frac{1}{2}a^2-(a-b)^2$, we have

$$\mathbf{K}_{ij}^2 \ge \frac{1}{2}\mathbf{L}_{ij}^2 - (\mathbf{K}_{ij}-\mathbf{L}_{ij})^2, \quad \forall i,j.$$

By a similar argument, we can show that

$$\mathrm{Var}[\Lambda_{\mathcal{S}_\mathrm{d}}(f)|\mathcal{X}] \ge \frac{1}{4}F(X_1,\dots,X_N) - 2\|f\|_\infty^2\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}^2.$$

□

Proposition A.7 (Expectation bound)**.** *We have*

$$\begin{aligned}\mathbb{E}\Big[\mathrm{Var}[\Lambda_{\mathcal{S}_\mathrm{d}}(f)|\mathcal{X}]\Big] &\le 2\mathrm{Var}[\Lambda_{\mathcal{S}_\mathrm{c}}(f)] + \frac{4\|f\|_\infty^2 nK_{\max}}{N}, \quad \textit{and}\\ \mathbb{E}\Big[\mathrm{Var}[\Lambda_{\mathcal{S}_\mathrm{d}}(f)|\mathcal{X}]\Big] &\ge \frac{N-1}{2N}\mathrm{Var}[\Lambda_{\mathcal{S}_\mathrm{c}}(f)] - \frac{2\|f\|_\infty^2 nK_{\max}}{N}.\end{aligned}$$

Proof. By Lemma A.6, we have

$$\mathrm{Var}[\Lambda_{\mathcal{S}_\mathrm{d}}(f)|\mathcal{X}] \le F(X_1,\dots,X_N) + 4\|f\|_\infty^2\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}^2.$$

By Lemma A.2, we have

$$\mathbb{E}[F(X_1,\dots,X_N)] = \frac{2(N-1)}{N}\mathrm{Var}[\Lambda_{\mathcal{S}_\mathrm{c}}(f)] \le 2\mathrm{Var}[\Lambda_{\mathcal{S}_\mathrm{c}}(f)].$$

By Lemma A.4, we have

$$\mathbb{E}\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}^2 \le \frac{nK_{\max}}{N}.$$

Combining these ingredients, we deduce that

$$\mathbb{E}\Big[\mathrm{Var}[\Lambda_{\mathcal{S}_\mathrm{d}}(f)|\mathcal{X}]\Big] \le 2\mathrm{Var}[\Lambda_{\mathcal{S}_\mathrm{c}}(f)] + 4\|f\|_\infty^2\frac{nK_{\max}}{N}.$$

Similarly, we get the lower bound for the expectation. □

Now we are ready to complete the proof of Theorem 3.2.

End of proof of Theorem 3.2. By Lemma A.6

$$\begin{aligned}\mathrm{Var}[\Lambda_{\mathcal{S}_\mathrm{d}}(f)|\mathcal{X}] &\le F(X_1,\dots,X_N) + 4\|f\|_\infty^2\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}^2\\ &\le \mathbb{E}[F] + |F-\mathbb{E}[F]| + 4\|f\|_\infty^2\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}^2\\ &\le 2\mathrm{Var}[\Lambda_{\mathcal{S}_\mathrm{c}}(f)] + 4\Big(|F-\mathbb{E}[F]| + \|f\|_\infty^2\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}^2\Big)\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}[\Lambda_{\mathcal{S}_\mathrm{d}}(f)|\mathcal{X}] &\ge \frac{1}{4}F(X_1,\dots,X_N) - 2\|f\|_\infty^2\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}^2\\ &\ge \frac{1}{4}\mathbb{E}[F] - \frac{1}{4}|F-\mathbb{E}[F]| - 2\|f\|_\infty^2\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}^2\\ &\ge \frac{N-1}{2N}\mathrm{Var}[\Lambda_{\mathcal{S}_\mathrm{c}}(f)] - 2\Big(|F-\mathbb{E}[F]| + \|f\|_\infty^2\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}^2\Big).\end{aligned}$$

Applying Corollary A.5, the theorem follows. □

A.2.3 Proof of Theorem 3.3

Lemma A.8. *If* $\max_{1\le i\le N}\left|\frac{\rho(X_i)}{\hat\rho(X_i)}-1\right|\le\varepsilon$, *we have*

$$\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\hat\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}^2\le 2\varepsilon^2 n+8\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}^2.$$

Proof. We have

$$\begin{aligned}&\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\hat\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}\\ \le\;&\Big\|\frac{1}{N}\mathbf{\Psi}^\top\Big(\mathbf{D}(\hat\rho^{-1})-\mathbf{D}(\rho^{-1})\Big)\mathbf{\Psi}\Big\|_{\mathrm{F}}+\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}\\ =\;&\frac{1}{N}\Big\|\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1/2})\Big(\mathbf{D}(\rho/\hat\rho)-\mathbf{I}_N\Big)\mathbf{D}(\rho^{-1/2})\mathbf{\Psi}\Big\|_{\mathrm{F}}+\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}.\end{aligned}$$

Now note that

$$\Big\|\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1/2})\Big(\mathbf{D}(\rho/\hat\rho)-\mathbf{I}_N\Big)\mathbf{D}(\rho^{-1/2})\mathbf{\Psi}\Big\|_{\mathrm{F}}\le\|\mathbf{D}(\rho/\hat\rho)-\mathbf{I}_N\|_{\mathrm{op}}\|\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}\|_{\mathrm{F}}.$$

Since $|\rho(X_i)/\hat\rho(X_i)-1|\le\varepsilon$ for every i, one has $\|\mathbf{D}(\rho/\hat\rho)-\mathbf{I}_N\|_{\mathrm{op}}\le\varepsilon$. Therefore

$$\begin{aligned}\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\hat\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}&\le\varepsilon\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}\Big\|_{\mathrm{F}}+\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}\\ &\le\varepsilon\|\mathbf{I}_n\|_{\mathrm{F}}+(1+\varepsilon)\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}\\ &=\varepsilon\sqrt{n}+(1+\varepsilon)\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}\\ &\le\varepsilon\sqrt{n}+2\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}},\end{aligned}$$

where we used the fact that $\varepsilon<1$. Using the elementary inequality $(a+b)^2\le 2a^2+2b^2$, we have

$$\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\hat\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}^2\le 2\varepsilon^2 n+8\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}^2$$

as desired. □

Lemma A.9. *If* $\max_{1\le i\le N}\left|\frac{\rho(X_i)}{\hat\rho(X_i)}-1\right|\le\varepsilon$, *we have*

$$\mathrm{Var}[\Lambda_{\widehat{\mathcal{S}}_{\mathrm{d}}}(f)|\mathcal{X}]\le(1+\varepsilon)^2F(X_1,\dots,X_N)+32\|f\|_\infty^2\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}^2+8\|f\|_\infty^2\varepsilon^2 n,$$

and

$$\mathrm{Var}[\Lambda_{\widehat{\mathcal{S}}_{\mathrm{d}}}(f)|\mathcal{X}]\ge\frac{(1-\varepsilon)^2}{4}F(X_1,\dots,X_N)-16\|f\|_\infty^2\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}^2-4\|f\|_\infty^2\varepsilon^2 n.$$

Proof. By similar arguments as in Lemma A.6, we have

$$\mathrm{Var}[\Lambda_{\widehat{\mathcal{S}}_{\mathrm{d}}}(f)|\mathcal{X}]\le\frac{1}{N^2}\sum_{i,j\in[N]}|f(X_i)-f(X_j)|^2\frac{|K(X_i,X_j)|^2}{\hat\rho(X_i)\hat\rho(X_j)}+4\|f\|_\infty^2\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\hat\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}^2,$$

and

$$\mathrm{Var}[\Lambda_{\widehat{\mathcal{S}}_{\mathrm{d}}}(f)|\mathcal{X}]\ge\frac{1}{4N^2}\sum_{i,j\in[N]}|f(X_i)-f(X_j)|^2\frac{|K(X_i,X_j)|^2}{\hat\rho(X_i)\hat\rho(X_j)}-2\|f\|_\infty^2\Big\|\frac{1}{N}\mathbf{\Psi}^\top\mathbf{D}(\hat\rho^{-1})\mathbf{\Psi}-\mathbf{I}_n\Big\|_{\mathrm{F}}^2.$$

Since $1-\varepsilon\le|\rho(X_i)/\hat\rho(X_i)|\le1+\varepsilon$ for every i, we have

$$(1-\varepsilon)^2\frac{|K(X_i,X_j)|^2}{\rho(X_i)\rho(X_j)}\le\frac{|K(X_i,X_j)|^2}{\hat\rho(X_i)\hat\rho(X_j)}\le(1+\varepsilon)^2\frac{|K(X_i,X_j)|^2}{\rho(X_i)\rho(X_j)},\quad\forall\,i,j.$$

Therefore

$$\mathrm{Var}[\Lambda_{\widehat{\mathcal{S}}_{\mathrm{d}}}(f)|\mathcal{X}] \leq (1+\varepsilon)^2 F(X_1,\dots,X_N) + 4\|f\|_\infty^2 \left\| \frac{1}{N}\mathbf{\Psi}^\top \mathbf{D}(\hat{\rho}^{-1})\mathbf{\Psi} - \mathbf{I}_n \right\|_{\mathrm{F}}^2,$$

and

$$\mathrm{Var}[\Lambda_{\widehat{\mathcal{S}}_{\mathrm{d}}}(f)|\mathcal{X}] \geq \frac{(1-\varepsilon)^2}{4} F(X_1,\dots,X_N) - 2\|f\|_\infty^2 \left\| \frac{1}{N}\mathbf{\Psi}^\top \mathbf{D}(\hat{\rho}^{-1})\mathbf{\Psi} - \mathbf{I}_n \right\|_{\mathrm{F}}^2.$$

By Lemma A.8, we deduce that

$$\mathrm{Var}[\Lambda_{\widehat{\mathcal{S}}_{\mathrm{d}}}(f)|\mathcal{X}] \leq (1+\varepsilon)^2 F(X_1,\dots,X_N) + 32\|f\|_\infty^2 \left\| \frac{1}{N}\mathbf{\Psi}^\top \mathbf{D}(\rho^{-1})\mathbf{\Psi} - \mathbf{I}_n \right\|_{\mathrm{F}}^2 + 8\|f\|_\infty^2 \varepsilon^2 n,$$

and

$$\mathrm{Var}[\Lambda_{\widehat{\mathcal{S}}_{\mathrm{d}}}(f)|\mathcal{X}] \geq \frac{(1-\varepsilon)^2}{4} F(X_1,\dots,X_N) - 16\|f\|_\infty^2 \left\| \frac{1}{N}\mathbf{\Psi}^\top \mathbf{D}(\rho^{-1})\mathbf{\Psi} - \mathbf{I}_n \right\|_{\mathrm{F}}^2 - 4\|f\|_\infty^2 \varepsilon^2 n$$

as desired. □

End of proof of Theorem 3.3. If $\max_{1\leq i\leq N}\left|\frac{\rho(X_i)}{\hat{\rho}(X_i)} - 1\right| \leq \varepsilon$, by Lemma A.9

$$\begin{aligned}
&\mathrm{Var}[\Lambda_{\widehat{\mathcal{S}}_{\mathrm{d}}}(f)|\mathcal{X}] \\
\leq\ & (1+\varepsilon)^2 \mathbb{E}[F] + (1+\varepsilon)^2 |F - \mathbb{E}[F]| + 32\|f\|_\infty^2 \left\| \frac{1}{N}\mathbf{\Psi}^\top \mathbf{D}(\rho^{-1})\mathbf{\Psi} - \mathbf{I}_n \right\|_{\mathrm{F}}^2 + 8\|f\|_\infty^2 \varepsilon^2 n \\
\leq\ & (1+\varepsilon)^2 2\mathrm{Var}[\Lambda_{\mathcal{S}_{\mathrm{c}}}(f)] + 32\left(|F - \mathbb{E}[F]| + \|f\|_\infty^2 \left\| \frac{1}{N}\mathbf{\Psi}^\top \mathbf{D}(\rho^{-1})\mathbf{\Psi} - \mathbf{I}_n \right\|_{\mathrm{F}}^2\right) + 8\|f\|_\infty^2 \varepsilon^2 n
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{Var}[\Lambda_{\widehat{\mathcal{S}}_{\mathrm{d}}}(f)|\mathcal{X}] \\
\geq\ & \frac{(1-\varepsilon)^2}{4}\mathbb{E}[F] - \frac{(1-\varepsilon)^2}{4}|F - \mathbb{E}[F]| - 16\|f\|_\infty^2 \left\| \frac{1}{N}\mathbf{\Psi}^\top \mathbf{D}(\rho^{-1})\mathbf{\Psi} - \mathbf{I}_n \right\|_{\mathrm{F}}^2 - 4\|f\|_\infty^2 \varepsilon^2 n \\
\geq\ & (1-\varepsilon)^2 \frac{N-1}{2N}\mathrm{Var}[\Lambda_{\mathcal{S}_{\mathrm{c}}}(f)] - 16\left(|F - \mathbb{E}[F]| + \|f\|_\infty^2 \left\| \frac{1}{N}\mathbf{\Psi}^\top \mathbf{D}(\rho^{-1})\mathbf{\Psi} - \mathbf{I}_n \right\|_{\mathrm{F}}^2\right) - 4\|f\|_\infty^2 \varepsilon^2 n.
\end{aligned}$$

Applying Corollary A.5 completes the proof. □

B Technical Appendices to wavelet DPPs

In this section, we provide detailed discussions and proofs for our theorems in Section 4. We first recall a useful formula to compute variance of linear statistics of DPPs that will be used throughout this section.

Lemma B.1. *Let $\mathcal{S}$ be a DPP on a Polish space (E,μ) whose kernel K has the following orthonormal expansion*

$$K(x,y) = \sum_{k=1}^{n} \phi_k(x)\overline{\phi_k(y)}, \quad x,y \in E,$$

where $\{\phi_k : 1 \leq k \leq n\}$ is an orthonormal set in $L^2(\mu) := L^2(E,\mu)$. Then for any test function $f : E \to \mathbb{R}$, we have

$$\mathrm{Var}[\Lambda_{\mathcal{S}}(f)] = \sum_{k=1}^{n} \|\mathrm{proj}_{H^\perp}(f\phi_k)\|_{L^2(\mu)}^2$$

where $H^\perp$ is the orthogonal complement of the subspace $H = \mathrm{span}\{\phi_k : 1 \leq k \leq n\}$ in $L^2(\mu)$.

Proof. We have

$$\begin{aligned}
\mathrm{Var}[\Lambda_{\mathcal{S}}(f)] &= \int |f(x)|^2 K(x,x)\mathrm{d}\mu(x) - \iint f(x)\overline{f(y)}|K(x,y)|^2\mathrm{d}\mu(x)\mathrm{d}\mu(y) \\
&= \int |f(x)|^2 \sum_{k=1}^{n} |\phi_k(x)|^2 \mathrm{d}\mu(x) - \iint f(x)\overline{f(y)}\Big|\sum_{k=1}^{n}\phi_k(x)\overline{\phi_k(y)}\Big|^2 \mathrm{d}\mu(x)\mathrm{d}\mu(y) \\
&= \sum_{k=1}^{n} \|f\phi_k\|_{L^2(\mu)}^2 - \sum_{k,k'=1}^{n} |\langle f\phi_k, \phi_{k'}\rangle_{L^2(\mu)}|^2 \\
&= \sum_{k=1}^{n}\Big(\|f\phi_k\|_{L^2(\mu)}^2 - \sum_{k'=1}^{n} |\langle f\phi_k, \phi_{k'}\rangle_{L^2(\mu)}|^2\Big) = \sum_{k=1}^{n} \|\mathrm{proj}_{H^\perp}(f\phi_k)\|_{L^2(\mu)}^2.
\end{aligned}$$

□

B.1 Brief introduction to wavelets

We briefly recall some standard facts from wavelet theory, following the notation of [21, Chapter 7].

Definition B.1. *A multiresolution approximation of $L^2(\mathbb{R})$ is a sequence of closed subspaces $(V_j)_{j\in\mathbb{Z}}$ satisfying*

$$\cdots \subset V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset V_{-2} \subset \cdots,$$

together with

$$\overline{\bigcup_{j\in\mathbb{Z}} V_j} = L^2(\mathbb{R}), \qquad \bigcap_{j\in\mathbb{Z}} V_j = \{0\},$$

and the scaling relation

$$f(x) \in V_j \quad \Longleftrightarrow \quad f(2x) \in V_{j-1}.$$

The space V_j should be understood as the approximation space at scale 2^j; thus, smaller values of j correspond to finer resolution.

Definition B.2. *Given a multiresolution approximation $(V_j)_{j\in\mathbb{Z}}$ of $L^2(\mathbb{R})$, a function $\phi \in L^2(\mathbb{R})$ is called a scaling function if its integer translates $\{\phi(\cdot - k) : k \in \mathbb{Z}\}$ generate the reference space V_0. In addition, ϕ is called an orthonormal scaling function if $\{\phi(\cdot - k) : k \in \mathbb{Z}\}$ is an orthonormal basis of V_0.*

The corresponding normalized dilates and translates are

$$\phi_{j,k}(x) = 2^{-j/2}\phi(2^{-j}x - k), \qquad j \in \mathbb{Z},\ k \in \mathbb{Z}.$$

Then, for each $j \in \mathbb{Z}$,

$$V_j = \overline{\mathrm{span}}\{\phi_{j,k} : k \in \mathbb{Z}\}.$$

The functions $\phi_{j,k}$ therefore describe the low-frequency or approximation component of a function at scale 2^j.

Lemma B.2. *Let $\phi \in L^\infty(\mathbb{R})$ be a compactly supported normalized orthonormal scaling function associated with a multiresolution analysis, normalized so that $\int_{\mathbb{R}} \phi(x)\mathrm{d}x = 1$. Then*

$$\sum_{k\in\mathbb{Z}} \phi(x-k) = 1 \qquad \textit{for a.e. } x \in \mathbb{R}.$$

Consequently,

$$\sum_{\mathbf{k}\in\mathbb{Z}^d} 2^{-dj/2}\Phi_{-j,\mathbf{k}}(\mathbf{x}) = 1 \quad \textit{for a.e. } \mathbf{x} \in \mathbb{R}^d.$$

Moreover, there exists $c_\phi > 0$ such that $\sum_{k\in\mathbb{Z}} \phi(x-k)^2 \geq c_\phi$ for a.e. $x \in \mathbb{R}$.

Proof. We use the Fourier transform convention $\widehat{f}(\xi) = \int_{\mathbb{R}} f(x)e^{-i\xi x}\mathrm{d}x$. Since ϕ is compactly supported and belongs to $L^\infty(\mathbb{R})$, we have $\phi \in L^1(\mathbb{R}) \cap L^2(\mathbb{R})$, and hence $\widehat{\phi}$ is continuous. The normalization gives

$$\widehat{\phi}(0) = \int_{\mathbb{R}} \phi(x)\mathrm{d}x = 1.$$

Because the integer translates $\{\phi(\cdot - k) : k \in \mathbb{Z}\}$ are orthonormal in $L^2(\mathbb{R})$, the standard Fourier characterization of orthonormal translates gives

$$\sum_{\ell \in \mathbb{Z}} \left|\widehat{\phi}(\xi + 2\pi\ell)\right|^2 = 1 \qquad \text{for a.e. } \xi \in \mathbb{R}.$$

We claim that

$$\widehat{\phi}(2\pi m) = 0 \quad \text{for all } m \in \mathbb{Z} \setminus \{0\}.$$

Indeed, suppose that for some $m \neq 0$ we have $\left|\widehat{\phi}(2\pi m)\right| > 0$, then by continuity of $\widehat{\phi}$, for ξ sufficiently close to 0, both $\left|\widehat{\phi}(\xi)\right|$ is close to 1, and $\left|\widehat{\phi}(\xi + 2\pi m)\right|$ is bounded away from 0. Hence, on a set of positive measure near $\xi = 0$, we would have $\sum_{\ell \in \mathbb{Z}} \left|\widehat{\phi}(\xi + 2\pi\ell)\right|^2 > 1$, contradicting the orthonormal-translate identity. The claim follows.

We define the periodization $P_\phi(x) := \sum_{k \in \mathbb{Z}} \phi(x - k)$. Since ϕ is compactly supported, this sum is locally finite. Hence P_ϕ is a well-defined 1-periodic function in L^2_{loc}. For $m \in \mathbb{Z}$, its m-th Fourier coefficient is

$$\int_0^1 P_\phi(x) e^{-2\pi i m x} \mathrm{d}x = \sum_{k \in \mathbb{Z}} \int_0^1 \phi(x - k) e^{-2\pi i m x} \mathrm{d}x = \int_{\mathbb{R}} \phi(u) e^{-2\pi i m u} \mathrm{d}u = \widehat{\phi}(2\pi m).$$

Hence, $\int_0^1 P_\phi(x) \mathrm{d}x = \widehat{\phi}(0) = 1$, while for $m \neq 0$, we have $\int_0^1 P_\phi(x) e^{-2\pi i m x} \mathrm{d}x = \widehat{\phi}(2\pi m) = 0$. Thus P_ϕ has the same Fourier coefficients as the constant function 1. Hence $P_\phi(x) = 1$ for a.e. $x \in \mathbb{R}$, that is,

$$\sum_{k \in \mathbb{Z}} \phi(x - k) = 1 \qquad \text{for a.e. } x \in \mathbb{R}.$$

Next, recall that $\phi_{-j,k}(x) = 2^{j/2}\phi(2^j x - k)$, which implies that for $\mathbf{x} = (x_1, \ldots, x_d)$ and $\mathbf{k} = (k_1, \ldots, k_d)$, we have

$$2^{-dj/2}\Phi_{-j,\mathbf{k}}(x) = \prod_{r=1}^{d} \phi(2^j x_r - k_r).$$

Summing over $\mathbf{k} \in \mathbb{Z}^d$, we get

$$\sum_{\mathbf{k} \in \mathbb{Z}^d} 2^{-dj/2}\Phi_{-j,\mathbf{k}}(\mathbf{x}) = \sum_{k_1, \ldots, k_d \in \mathbb{Z}} \prod_{r=1}^{d} \phi(2^j x_r - k_r) = \prod_{r=1}^{d} \left(\sum_{k_r \in \mathbb{Z}} \phi(2^j x_r - k_r) \right) = 1$$

for a.e. $\mathbf{x} \in \mathbb{R}^d$ as desired.

Finally, since ϕ is compactly supported, there exists an integer $M_\phi < \infty$ such that, for every $x \in \mathbb{R}$, at most M_ϕ terms in the sum $\sum_{k \in \mathbb{Z}} \phi(x - k)$ are nonzero. By Cauchy-Schwarz, we have

$$1 = \Big(\sum_{k \in \mathbb{Z}} \phi(x - k) \Big)^2 \leq (\#\{k \in \mathbb{Z} : \phi(x - k) \neq 0\}) \sum_{k \in \mathbb{Z}} \phi(x - k)^2 \leq M_\phi \sum_{k \in \mathbb{Z}} \phi(x - k)^2$$

for a.e. $x \in \mathbb{R}$. Let $c_\phi := M_\phi^{-1}$, we then have $\sum_{k \in \mathbb{Z}} \phi(x - k)^2 \geq c_\phi > 0$ for a.e. $x \in \mathbb{R}$. This completes the proof. □

B.2 Spaces of test functions

In this section, we present the definitions of Hölder continuous functions and Sobolev spaces that we mentioned in Section 4.

Definition B.3 (Hölder continuous functions)**.** *Let $s \in (0, 1]$. The space of s-Hölder continuous functions on $\mathbb{R}^d$, denoted by $C^{0,s}$, consist of all functions $f : \mathbb{R}^d \to \mathbb{R}$ such that*

$$|f|_{C^{0,s}} := \sup_{\mathbf{x} \neq \mathbf{y}} \frac{|f(\mathbf{x}) - f(\mathbf{y})|}{\|\mathbf{x} - \mathbf{y}\|^s} < \infty.$$

Remark 2. *If f satisfies (∗), namely, $f(\mathbf{x}) = 0$ on $\mathbb{R}^d \setminus (0, 1)^d$, then we have*

$$|f|_{C^{0,s}} = \sup_{\mathbf{x}, \mathbf{y} \in [0,1]^d, \mathbf{x} \neq \mathbf{y}} \frac{|f(\mathbf{x}) - f(\mathbf{y})|}{\|\mathbf{x} - \mathbf{y}\|^s}.$$

We also consider the Sobolev spaces $H^s(\Omega)$, where Ω is an open domain in $\mathbb{R}^d$. For $s \in \mathbb{N}$, the common definition of Sobolev spaces is as follows:

Definition B.4 (Sobolev spaces)**.** *Let $s \in \mathbb{N}, s \geq 1$, the Sobolev space $H^s(\Omega)$ is defined by*

$$H^s(\Omega) := \Big\{ f \in L^2(\Omega) : D^\alpha f \in L^2(\Omega) \text{ for every multi-index } \alpha \text{ with } |\alpha| \leq s \Big\}$$

where $D^\alpha f$ denotes the weak derivative of f. The Sobolev norm is defined by

$$\|f\|^2_{H^s(\Omega)} := \sum_{|\alpha| \leq s} \|D^\alpha f\|^2_{L^2(\Omega)}.$$

In particular, for $s = 1$

$$\|f\|^2_{H^1(\Omega)} = \|f\|^2_{L^2(\Omega)} + \|\nabla f\|^2_{L^2(\Omega)}.$$

For $s < 1$, we take the following definition for the fractional Sobolev space $H^s(\Omega)$:

Definition B.5 (Fractional Sobolev spaces)**.** *Let $s \in (0, 1)$, the fractional Sobolev space $H^s(\Omega)$ is defined by*

$$H^s(\Omega) := \Big\{ f \in L^2(\Omega) : |f|_{H^s(\Omega)} < \infty \Big\},$$

where

$$|f|_{H^s(\Omega)} := \Big(\iint_{\Omega \times \Omega} \frac{(f(\mathbf{x}) - f(\mathbf{y}))^2}{\|\mathbf{x} - \mathbf{y}\|^{d+2s}} \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \Big)^{1/2},$$

called the Gagliardo seminorm of f.

We are particularly interested in the case $\Omega = (0, 1)^d$. For simplicity of notation, we set

$$H^s := H^s((0, 1)^d).$$

B.3 Proof of Theorem 4.2

In this section, we present the detailed proof of Theorem 4.2. Let ϕ be a bounded, compactly supported, orthonormal scaling function associated with a multiresolution analysis, and normalized by $\int \phi(x)\mathrm{d}x = 1$. We consider the wavelet DPP kernel K defined in (8), namely,

$$K(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{k} \in \mathcal{I}} \Phi_{-j,\mathbf{k}}(\mathbf{x}) \Phi_{-j,\mathbf{k}}(\mathbf{y}).$$

Let $H := \mathrm{span}\{\Phi_{-j,\mathbf{k}} : \mathbf{k} \in \mathcal{I}\}$ in $L^2(\mathbb{R}^d, \mathrm{d}\mathbf{x})$. Let $\mathcal{S}$ be the DPP on $\mathbb{R}^d$ defined by K with respect to the Lebesgue measure $\mathrm{d}\mathbf{x}$. Let f be a test function satisfying ($*$), then by Lemma B.1 we have

$$\mathrm{Var}[\Lambda_{\mathcal{S}}(f)] = \sum_{\mathbf{k} \in \mathcal{I}} \|\mathrm{proj}_{H^\perp}(f\Phi_{-j,\mathbf{k}})\|^2_{L^2(\mathrm{d}\mathbf{x})}.$$

Remark 3. *In what follows, C will denote a constant depending only on d and ϕ, which possibly changes from line by line.*

B.3.1 The case of $C^{0,s}$.

We first prove for the case $f \in C^{0,s}$ and f satisfies ($*$). In particular, we will show that

$$\mathrm{Var}[\Lambda_{\mathcal{S}}(f)] \leq C|f|^2_{C^{0,s}} n^{1-2s/d},$$

where $C > 0$ depending only on ϕ and d.

To this end, for each $\mathbf{k} \in \mathcal{I}$, we fix a point $\mathbf{x}_\mathbf{k} \in \mathrm{supp}(\Phi_{-j,\mathbf{k}})$. We define a remainder function

$$R_\mathbf{k}(\mathbf{x}) := f(\mathbf{x}) - f(\mathbf{x}_\mathbf{k}).$$

Then

$$\mathrm{proj}_{H^\perp}(f\Phi_{-j,\mathbf{k}}) = f(\mathbf{x}_\mathbf{k})\mathrm{proj}_{H^\perp}(\Phi_{-j,\mathbf{k}}) + \mathrm{proj}_{H^\perp}(R_\mathbf{k}\Phi_{-j,\mathbf{k}}) = \mathrm{proj}_{H^\perp}(R_\mathbf{k}\Phi_{-j,\mathbf{k}})$$

since $\Phi_{-j,\mathbf{k}} \in H$. Therefore

$$\begin{aligned}\|\mathrm{proj}_{H^\perp}(f\Phi_{-j,\mathbf{k}})\|^2_{L^2(\mathrm{d}\mathbf{x})} &= \|\mathrm{proj}_{H^\perp}(R_{\mathbf{k}}\Phi_{-j,\mathbf{k}})\|^2_{L^2(\mathrm{d}\mathbf{x})} \le \|R_{\mathbf{k}}\Phi_{-j,\mathbf{k}}\|^2_{L^2(\mathrm{d}\mathbf{x})} \\ &= \int_{\mathrm{supp}(\Phi_{-j,\mathbf{k}})} |f(\mathbf{x}) - f(\mathbf{x}_{\mathbf{k}})|^2 |\Phi_{-j,\mathbf{k}}(\mathbf{x})|^2 \mathrm{d}\mathbf{x} \\ &\le \sup_{\mathbf{x}\in\mathrm{supp}(\Phi_{-j,\mathbf{k}})} |f(\mathbf{x}) - f(\mathbf{x}_{\mathbf{k}})|^2\end{aligned}$$

since $\int |\Phi_{-j,\mathbf{k}}(\mathbf{x})|^2 \mathrm{d}\mathbf{x} = 1$. Note that $f \in C^{0,s}$, we have

$$|f(\mathbf{x}) - f(\mathbf{x}_{\mathbf{k}})|^2 \le |f|^2_{C^{0,s}} \|\mathbf{x} - \mathbf{x}_{\mathbf{k}}\|^{2s}, \quad \forall \mathbf{x}.$$

This implies

$$\begin{aligned}\|\mathrm{proj}_{H^\perp}(f\Phi_{-j,\mathbf{k}})\|^2_{L^2(\mathrm{d}\mathbf{x})} &\le |f|^2_{C^{0,s}} \sup_{\mathbf{x}\in\mathrm{supp}(\Phi_{-j,\mathbf{k}})} \|\mathbf{x} - \mathbf{x}_{\mathbf{k}}\|^{2s} \\ &\le |f|^2_{C^{0,s}} \mathrm{diam}(\mathrm{supp}(\Phi_{-j,\mathbf{k}}))^{2s},\end{aligned}$$

where $\mathrm{diam}(A) := \sup_{\mathbf{x},\mathbf{y}\in A} \|\mathbf{x}-\mathbf{y}\|$ for $A \subset \mathbb{R}^d$. Note that $\Phi_{-j,\mathbf{k}}(\mathbf{x}) = \phi_{-j,k_1}(x_1)\ldots\phi_{-j,k_d}(x_d)$, which implies

$$\mathrm{diam}(\mathrm{supp}(\Phi_{-j,\mathbf{k}})) = \sqrt{d}\cdot 2^{-j}\cdot \mathrm{diam}(\mathrm{supp}(\phi)).$$

Thus

$$\|\mathrm{proj}_{H^\perp}(f\Phi_{-j,\mathbf{k}})\|^2_{L^2(\mathrm{d}\mathbf{x})} \le \mathrm{diam}(\mathrm{supp}(\phi))^{2s} |f|^2_{C^{0,s}} d^s 2^{-2sj}, \quad \forall \mathbf{k} \in \mathcal{I}.$$

Therefore,

$$\mathrm{Var}[\Lambda_{\mathcal{S}}(f)] \le n \cdot \mathrm{diam}(\mathrm{supp}(\phi))^{2s} |f|^2_{C^{0,s}} d^s 2^{-2sj} \le C(\phi, d) \cdot |f|^2_{C^{0,s}} n^{1-2s/d},$$

as desired.

B.3.2 The case of H^s with $s < 1$.

Let $f \in H^s = H^s((0,1)^d)$ with $s \in (0,1)$. We now show that

$$\mathrm{Var}[\Lambda_{\mathcal{S}}(f)] \le C|f|^2_{H^s} n^{1-2s/d},$$

where $C > 0$ depending only on ϕ and d. To this end, let $\mathrm{Conv}[A]$ denote the convex hull of a set A. For each $\mathbf{k} \in \mathcal{I}$, let $S_{\mathbf{k}} := \mathrm{Conv}[\mathrm{supp}(\Phi_{-j,\mathbf{k}})] \cap (0,1)^d$, and define

$$\bar{f}_{\mathbf{k}} := \frac{1}{\mathrm{Vol}(S_{\mathbf{k}})} \int_{S_{\mathbf{k}}} f(\mathbf{y}) \mathrm{d}\mathbf{y}.$$

By the same argument as above, we have

$$\begin{aligned}&\|\mathrm{proj}_{H^\perp}(f\Phi_{-j,\mathbf{k}})\|^2_{L^2(\mathrm{d}\mathbf{x})} = \|\mathrm{proj}_{H^\perp}((f - \bar{f}_{\mathbf{k}})\Phi_{-j,\mathbf{k}})\|^2_{L^2(\mathrm{d}\mathbf{x})} \\ \le\ & \int_{S_{\mathbf{k}}} |f(\mathbf{x}) - \bar{f}_{\mathbf{k}}|^2 |\Phi_{-j,\mathbf{k}}(\mathbf{x})|^2 \mathrm{d}\mathbf{x} \\ \le\ & 2^{dj}\|\phi\|^{2d}_\infty \int_{S_{\mathbf{k}}} \Big| f(\mathbf{x}) - \frac{1}{\mathrm{Vol}(S_{\mathbf{k}})} \int_{S_{\mathbf{k}}} f(\mathbf{y}) \mathrm{d}\mathbf{y} \Big|^2 \mathrm{d}\mathbf{x}.\end{aligned}$$

By the fractional Poincaré-Wirtinger inequality (see e.g. [1], Proposition 2.2), we have for $f \in H^s((0,1)^d)$

$$\int_{S_{\mathbf{k}}} \Big| f(\mathbf{x}) - \frac{1}{\mathrm{Vol}(S_{\mathbf{k}})} \int_{S_{\mathbf{k}}} f(\mathbf{y}) \mathrm{d}\mathbf{y} \Big|^2 \mathrm{d}\mathbf{x} \le C \cdot \mathrm{diam}(S_{\mathbf{k}})^{2s} |f|^2_{H^s(S_{\mathbf{k}})}$$

where C depends only on ϕ and d, and

$$|f|^2_{H^s(\Omega)} := \iint_{\Omega\times\Omega} \frac{(f(\mathbf{x}) - f(\mathbf{y}))^2}{\|\mathbf{x} - \mathbf{y}\|^{d+2s}} \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}$$

for a domain $\Omega \subset \mathbb{R}^d$. Now note that

$$\mathrm{diam}(S_{\mathbf{k}}) \le \mathrm{diam}(\mathrm{supp}(\Phi_{-j,\mathbf{k}})) = \mathrm{diam}(\mathrm{supp}(\phi)) \cdot 2^{-j}\sqrt{d}.$$

Thus

$$\int_{S_{\mathbf{k}}} \Big| f(\mathbf{x}) - \frac{1}{\mathrm{Vol}(S_{\mathbf{k}})} \int_{S_{\mathbf{k}}} f(\mathbf{y}) \mathrm{d}\mathbf{y} \Big|^2 \mathrm{d}\mathbf{x} \leq C \cdot 2^{-2sj} |f|^2_{H^s(S_{\mathbf{k}})}$$

for some constant C depending only on d and ϕ. Hence

$$\mathrm{Var}[\Lambda_{\mathcal{S}}(f)] \leq C \cdot 2^{dj-2sj} \sum_{\mathbf{k}\in\mathcal{I}} |f|^2_{H^s(S_{\mathbf{k}})}.$$

We observe that

$$\begin{aligned}
\sum_{\mathbf{k}\in\mathcal{I}} |f|^2_{H^s(S_{\mathbf{k}})} &= \iint_{(0,1)^d\times(0,1)^d} \frac{(f(\mathbf{x})-f(\mathbf{y}))^2}{\|\mathbf{x}-\mathbf{y}\|^{d+2s}} \Big(\sum_{\mathbf{k}\in\mathcal{I}} \mathbf{1}_{S_{\mathbf{k}}}(\mathbf{x}) \mathbf{1}_{S_{\mathbf{k}}}(\mathbf{y}) \Big) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \\
&\leq \iint_{(0,1)^d\times(0,1)^d} \frac{(f(\mathbf{x})-f(\mathbf{y}))^2}{\|\mathbf{x}-\mathbf{y}\|^{d+2s}} \Big(\sum_{\mathbf{k}\in\mathcal{I}} \mathbf{1}_{S_{\mathbf{k}}}(\mathbf{x}) \Big) \Big(\sum_{\mathbf{k}\in\mathcal{I}} \mathbf{1}_{S_{\mathbf{k}}}(\mathbf{y}) \Big) \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}.
\end{aligned}$$

For $\mathbf{k}, \mathbf{k}' \in \mathcal{I}$, we have $S_{\mathbf{k}} \cap S_{\mathbf{k}'} \neq \varnothing$ only if

$$\mathrm{Conv}[\mathrm{supp}(\phi_{-j,k_i})] \cap \mathrm{Conv}[\mathrm{supp}(\phi_{-j,k'_i})] \neq \varnothing \quad \text{for all } 1 \leq i \leq d.$$

Assuming that $\mathrm{supp}(\phi) \subset [a,b]$, we have

$$\mathrm{Conv}[\mathrm{supp}(\phi_{-j,k})] \subset [2^{-j}(a+k), 2^{-j}(b+k)]$$

which implies that

$$\#\{k' \in \mathbb{Z} : \mathrm{Conv}[\mathrm{supp}(\phi_{-j,k})] \cap \mathrm{Conv}[\mathrm{supp}(\phi_{-j,k'})] \neq \varnothing\} \leq 2(b-a)+1.$$

Thus, for every $\mathbf{x} \in (0,1)^d$, we have

$$\sum_{\mathbf{k}\in\mathcal{I}} \mathbf{1}_{S_{\mathbf{k}}}(\mathbf{x}) \leq (2(b-a)+1)^d,$$

which implies

$$\sum_{\mathbf{k}\in\mathcal{I}} |f|^2_{H^s(S_{\mathbf{k}})} \leq (2(b-a)+1)^{2d} |f|^2_{H^s((0,1)^d)}.$$

Combining all ingredients, we deduce that

$$\mathrm{Var}[\Lambda_{\mathcal{S}}(f)] \leq C |f|^2_{H^s((0,1)^d)} n^{1-2s/d}$$

where C depending only on d and ϕ as desired.

B.3.3 The case of H^1.

Let $f \in H^1 = H^1((0,1)^d)$. We now show that

$$\mathrm{Var}[\Lambda_{\mathcal{S}}(f)] \leq C \|\nabla f\|^2_{L^2((0,1)^d)} n^{1-2s/d},$$

where $C > 0$ depending only on ϕ and d. The proof follows exactly as the case H^s with $s < 1$ above, where the Poincaré-Wirtinger inequality replaced by the Payne–Weinberger Poincaré inequality [23], which states that for every bounded, convex open domain $\Omega \subset \mathbb{R}^d$

$$\int_{\Omega} \Big| f(\mathbf{x}) - \frac{1}{\mathrm{Vol}(\Omega)} \int_{\Omega} f(\mathbf{y}) \mathrm{d}\mathbf{y} \Big|^2 \mathrm{d}\mathbf{x} \leq \frac{\mathrm{diam}(\Omega)^2}{\pi^2} \|\nabla f\|^2_{L^2(\Omega)}, \quad \forall f \in H^1(\Omega).$$

B.4 Proof of Proposition 4.1

Consider the kernel (8), for some $j \geq 1$ and $\phi = \phi_{\mathrm{Haar}}$ the Haar scaling function. Note that in that case, the rank of the kernel is $n = 2^{dj}$, that $K(x,x) = n$ for all $x \in [0,1]^d$, and that for $k \neq \ell$,

$$\mathrm{supp}(\mathbf{\Phi}_{-j,k}) \cap \mathrm{supp}(\mathbf{\Phi}_{-j,\ell}) = \emptyset.$$

Actually, the supports of $\mathbf{\Phi}_{-\mathbf{j},\mathbf{k}}$, $k \in \mathcal{I}$ are a partition of $[0,1]^d$. Moreover, by construction (8), for $k \in \mathcal{I}$,

$$x, y \in \mathrm{supp}(\mathbf{\Phi}_{-j,k}) \Rightarrow K(x,y) = n,$$

while $K(x,y) = 0$ as soon as x and y belong to the supports of two different wavelets. Now consider the pdf (3) defining DPP(K,dx). If any two points x_i, x_j, $i \neq j$ are in the support of the same $\mathbf{\Phi}_{-j,k}$, then two columns of the matrix in the determinant are equal, and the determinant is thus zero. Meanwhile, the RHS of (3) is constant on configurations $x_1, \ldots, x_n$ that belong to n different supports. This proves the claim.

C Technical Appendices to applications

C.1 A quadrature for the Haar DPP

The simple linear statistic (9) has fast-decaying mean squared error when the scaling function ϕ is the Haar scaling function. In light of Proposition 4.1, the result is expected since the rate had been known for stratified sampling since [11]. For completeness, we still state the proposition and prove it as a consquence of our more general result for wavelet-based DPPs.

Proposition C.1 (Haar wavelet DPPs as quadratures)**.** *Let $\mathcal{S}$ be the DPP on $(\mathbb{R}^d, \mathrm{d}\mathbf{x})$ defined by the kernel K in (8) with ϕ_{Haar} the Haar scaling function. Let f be a function satisfying ($*$). Then, there exists a constants $C > 0$ depending only on d and ω such that*

$$\mathbb{E}\Big[\Big|\mathcal{L}_{\mathcal{S}}(f) - \int f(\mathbf{x})\omega(\mathbf{x})\mathrm{d}\mathbf{x})\Big|^2\Big] \leq C\|f\|_s^2 n^{-1-2s/d},$$

where

$$\|f\|_s^2 := \begin{cases} \|f\|_\infty^2 + |f|_{C^{0,s}}^2 & \text{if } f \in C^{0,s} \\ \|f\|_{L^2((0,1)^d)}^2 + |f|_{H^s}^2 & \text{if } f \in H^s, s < 1 \\ \|f\|_{L^2((0,1)^d)}^2 + \|\nabla f\|_{L^2((0,1)^d)}^2 & \text{if } f \in H^s, s = 1. \end{cases}$$

Proof. Note that for the Haar scaling function $\phi = \phi_{\mathrm{Haar}}$, we have for a.e. $\mathbf{x} \in (0,1)^d$

$$K(\mathbf{x}, \mathbf{x}) = \sum_{\mathbf{k} \in \mathcal{I}} \Phi_{-j,\mathbf{k}}(\mathbf{x})^2 = n.$$

Let $f \in C^{0,s}$ for $s \in (0, 1]$. Since $\omega \in C^1$, we have

$$\begin{aligned} |f(\mathbf{x})\omega(\mathbf{x}) - f(\mathbf{y})\omega(\mathbf{y})| &\leq \|\omega\|_\infty |f(\mathbf{x}) - f(\mathbf{y})| + \|f\|_\infty |\omega(\mathbf{x}) - \omega(\mathbf{y})| \\ &\leq \|\omega\|_\infty |f|_{C^{0,s}} \|\mathbf{x} - \mathbf{y}\|^s + \|f\|_\infty \|\nabla\omega\|_\infty \|\mathbf{x} - \mathbf{y}\| \\ &\leq \Big(\|\omega\|_\infty |f|_{C^{0,s}} + d^{(1-s)/2}\|f\|_\infty \|\nabla\omega\|_\infty\Big)\|\mathbf{x} - \mathbf{y}\|^s. \end{aligned}$$

This implies $f\omega \in C^{0,s}$, and

$$|f\omega|_{C^{0,s}}^2 \leq C(|f|_{C^{0,s}}^2 + \|f\|_\infty^2) = C\|f\|_s^2$$

for some constant $C > 0$ depending only on d and ω. Applying Theorem 4.2 yields

$$\mathbb{E}\Big[\Big|\mathcal{L}_{\mathcal{S}}(f) - \int f(\mathbf{x})\omega(\mathbf{x})\mathrm{d}\mathbf{x})\Big|^2\Big] = \mathrm{Var}\Big[\Lambda_{\mathcal{S}}\Big(\frac{f\omega}{n}\Big)\Big] \lesssim \|f\|_s^2 n^{-1-2s/d}.$$

By a similar argument, we also can show that if $f \in H^s$ and $\omega \in C^1$, then $f\omega \in H^s$ with

$$|f\omega|_{H^s}^2 \leq C\|f\|_s^2$$

for some constant $C > 0$ depending only on d and ω. Applying Theorem 4.2 yields the desired result. □

C.2 Proof of Theorem 5.1

Remark 4. *In what follows, we use the notations $\gtrsim, \lesssim$ to denote inequalities up to multiplicative constants which depend only on d and ϕ.*

Proof. Let $\{\mathbf{x}_\mathbf{k} : \mathbf{k} \in \mathcal{I}\}$ be the nodes designed as in (10), i.e., be such that for each $\mathbf{k} \in \mathcal{I}$, $\mathbf{x}_\mathbf{k}$ is an arbitrarily chosen but fixed point in $\mathrm{supp}(\Phi_{-j,\mathbf{k}})$. For a function f with $\mathrm{supp}(f) \Subset (0,1)^d$, we define the following operator Q acting on f as

$$Qf(\mathbf{x}) := \frac{1}{2^{dj/2}} \sum_{\mathbf{k} \in \mathcal{I}} \Phi_{-j,\mathbf{k}}(\mathbf{x}) f(\mathbf{x}_\mathbf{k}). \tag{15}$$

We then have

$$\int Qf(\mathbf{x})\mathrm{d}\mathbf{x} = \frac{1}{2^{dj/2}} \sum_{\mathbf{k} \in \mathcal{I}} f(\mathbf{x}_\mathbf{k}) \int \Phi_{-j,\mathbf{k}}(\mathbf{x})\mathrm{d}\mathbf{x} = \frac{1}{2^{dj}} \sum_{\mathbf{k} \in \mathcal{I}} f(\mathbf{x}_\mathbf{k}),$$

where we used the fact that

$$\int \Phi_{-j,\mathbf{k}}(\mathbf{x})\mathrm{d}\mathbf{x} = \prod_{i=1}^{d}\Big(\int \phi_{-j,k_i}(x_i)\mathrm{d}x_i\Big) = \prod_{i=1}^{d}\Big(\int 2^{j/2}\phi(2^j x_i - k_i)\mathrm{d}x_i\Big) = 2^{-dj/2}.$$

We now show that:

$$\mathrm{Var}\Big[\Lambda_{\mathcal{S}}\Big(\frac{f-Qf}{K}\Big)\Big] = O(n^{-1-2s/d}) \tag{16}$$

for all function $f \in C^{0,s}$ with $\mathrm{supp}(f) \Subset (0,1)^d$.

To this end, we first observe that

$$\mathrm{Var}\Big[\Lambda_{\mathcal{S}}\Big(\frac{f-Qf}{K}\Big)\Big] \le \int \Big|\frac{f(\mathbf{x})-Qf(\mathbf{x})}{K(\mathbf{x},\mathbf{x})}\Big|^2 K(\mathbf{x},\mathbf{x})\mathrm{d}\mathbf{x}.$$

Since $\mathrm{supp}(f)$ is bounded away from the boundary of the hypercube $[0,1]^d$, there exists $\delta > 0$ such that $\mathrm{supp}(f) \subset (\delta, 1-\delta)^d$. Note that for each $\delta > 0$, there exists j_0 such that for $j \ge j_0$ we have

$$K(\mathbf{x},\mathbf{x}) = 2^{dj}\prod_{i=1}^{d}\Big(\sum_{k\in\mathbb{Z}}\phi(2^j x_i - k)^2\Big), \quad \forall \mathbf{x} \in (\delta, 1-\delta)^d.$$

By Lemma B.2, there exists $c_\phi > 0$ so that $\sum_{k\in\mathbb{Z}}\phi(x-k)^2 \ge c_\phi > 0$ for a.e. $x \in \mathbb{R}$. Thus we have $K(\mathbf{x},\mathbf{x}) \gtrsim n$ for all $j \ge j_0$ and a.e. $\mathbf{x} \in (\delta, 1-\delta)^d$. Thus, for $j \ge j_0$ we would have

$$\mathrm{Var}\Big[\Lambda_{\mathcal{S}}\Big(\frac{f-Qf}{K}\Big)\Big] \lesssim \frac{1}{n}\int_{(\delta,1-\delta)^d} |f(\mathbf{x})-Qf(\mathbf{x})|^2\mathrm{d}\mathbf{x}.$$

To control $|f(\mathbf{x}) - Qf(\mathbf{x})|$, we note that

$$\sum_{\mathbf{k}\in\mathbb{Z}^d} 2^{-dj/2}\Phi_{-j,\mathbf{k}}(\mathbf{x}) = 1, \quad \text{for a.e. } \mathbf{x} \in \mathbb{R}^d.$$

This implies

$$f(\mathbf{x}) = \sum_{\mathbf{k}\in\mathbb{Z}^d} 2^{-dj/2}\Phi_{-j,\mathbf{k}}(\mathbf{x})f(\mathbf{x}), \quad \text{for a.e. } \mathbf{x} \in \mathbb{R}^d.$$

Note that if $\mathrm{supp}(f) \subset (\delta, 1-\delta)^d$, for large enough j (depending on $\mathrm{supp}(\phi)$ and δ) we would have $(\delta, 1-\delta)^d \cap \mathrm{supp}(\Phi_{-j,\mathbf{k}}) = \varnothing$ if $\mathbf{k} \notin \mathcal{I}$. Hence, for large enough j, we would have

$$f(\mathbf{x}) = \sum_{\mathbf{k}\in\mathcal{I}} 2^{-dj/2}\Phi_{-j,\mathbf{k}}(\mathbf{x})f(\mathbf{x}), \quad \text{for a.e. } \mathbf{x} \in (\delta, 1-\delta)^d.$$

This implies that for a.e. $\mathbf{x} \in (\delta, 1-\delta)^d$

$$\begin{aligned} |f(\mathbf{x}) - Qf(\mathbf{x})| &\le 2^{-dj/2}\sum_{\mathbf{k}\in\mathcal{I}} |\Phi_{-j,\mathbf{k}}(\mathbf{x})||f(\mathbf{x}) - f(\mathbf{x}_\mathbf{k})| \\ &\le \|\phi\|_\infty^d \sum_{\mathbf{k}\in\mathcal{I}} \mathbf{1}_{S_\mathbf{k}}(\mathbf{x})|f(\mathbf{x}) - f(\mathbf{x}_\mathbf{k})|, \end{aligned}$$

where $S_\mathbf{k} := \mathrm{supp}(\Phi_{-j,\mathbf{k}}) \cap (0,1)^d$. For each $\mathbf{k} \in \mathcal{I}$, we have

$$|f(\mathbf{x}) - f(\mathbf{x}_\mathbf{k})| \lesssim |f|_{C^{0,s}} 2^{-sj}.$$

By the proof of Theorem 4.2 (ii), the sum $\sum_{\mathbf{k}\in\mathcal{I}} \mathbf{1}_{S_\mathbf{k}}(\mathbf{x})$ is finite and depends only on d and ϕ. Therefore, for a.e. $\mathbf{x} \in (\delta, 1-\delta)^d$

$$|f(\mathbf{x}) - Qf(\mathbf{x})|^2 \lesssim |f|^2_{C^{0,s}} 2^{-2sj}.$$

Thus,

$$\mathrm{Var}\Big[\Lambda_{\mathcal{S}}\Big(\frac{f-Qf}{K}\Big)\Big] \le C|f|^2_{C^{0,s}} n^{-1-2s/d}$$

for some constant $C > 0$ depending only on ϕ and d. The claim (16) follows.

Now, we observe that

$$\begin{aligned}\widetilde{\mathcal{L}}_{\mathcal{S}}(f) &= \mathcal{L}_{\mathcal{S}}(f) - \frac{1}{2^{dj/2}}\sum_{i=1}^{n}\sum_{\mathbf{k}\in\mathcal{I}}\frac{\Phi_{-j,\mathbf{k}}(Y_i)}{K(Y_i,Y_i)}f(\mathbf{x}_{\mathbf{k}})\omega(\mathbf{x}_{\mathbf{k}}) + \frac{1}{2^{dj}}\sum_{\mathbf{k}\in\mathcal{I}}f(\mathbf{x}_{\mathbf{k}})\omega(\mathbf{x}_{\mathbf{k}})\\ &= \Lambda_{\mathcal{S}}\Big(\frac{f\omega}{K}\Big) - \sum_{i=1}^{n}\frac{Q(f\omega)(Y_i)}{K(Y_i,Y_i)} + \int Q(f\omega)(\mathbf{x})\mathrm{d}\mathbf{x}\\ &= \Lambda_{\mathcal{S}}\Big(\frac{f\omega}{K}\Big) - \Lambda_{\mathcal{S}}\Big(\frac{Q(f\omega)}{K}\Big) + \int Q(f\omega)(\mathbf{x})\mathrm{d}\mathbf{x}\\ &= \Lambda_{\mathcal{S}}\Big(\frac{f\omega - Q(f\omega)}{K}\Big) + \int Q(f\omega)(\mathbf{x})\mathrm{d}\mathbf{x}.\end{aligned}$$

Taking expectation gives

$$\mathbb{E}[\widetilde{\mathcal{L}}_{\mathcal{S}}(f)] = \int\frac{f(\mathbf{x})\omega(\mathbf{x}) - Q(f\omega)(\mathbf{x})}{K(\mathbf{x},\mathbf{x})}K(\mathbf{x},\mathbf{x})\mathrm{d}\mathbf{x} + \int Q(f\omega)(\mathbf{x})\mathrm{d}\mathbf{x} = \int f(\mathbf{x})\omega(\mathbf{x})\mathrm{d}\mathbf{x}.$$

This shows $\widetilde{\mathcal{L}}_{\mathcal{S}}(f)$ is an unbiased estimator of $\int f(\mathbf{x})\omega(\mathbf{x})\mathrm{d}\mathbf{x}$.

Finally, notice that if $f \in C^{0,s}$ and $\omega \in C^1$, then $f\omega \in C^{0,s}$ with

$$|f\omega|^2_{C^{0,s}} \le C\Big(|f|^2_{C^{0,s}} + \|f\|^2_{\infty}\Big)$$

for some constant $C > 0$ depending only on d and ω. Moreover, we have $\mathrm{supp}(f\omega) \subset \mathrm{supp}(\omega) \Subset (0,1)^d$. Applying the claim (16) for the function $f\omega$ yields the desired result. □

C.3 Proof of Theorem 5.2

Proof of Theorem 5.2. Notice that

$$\mathbb{E}[L_{\mathcal{S}_{\mathrm{d}}}(f)|\mathcal{X}] = \sum_{\mathbf{x}\in\mathcal{X}}\frac{f(\mathbf{x})}{\mathbb{P}(\mathbf{x}\in\mathcal{S}_{\mathrm{d}}|\mathcal{X})}\mathbb{P}(\mathbf{x}\in\mathcal{S}_{\mathrm{d}}|\mathcal{X}) = \sum_{\mathbf{x}\in\mathcal{X}}f(\mathbf{x}) = L(f),$$

which implies $\mathbb{E}[|L_{\mathcal{S}_{\mathrm{d}}}(f) - L(f)|^2|\mathcal{X}] = \mathrm{Var}[L_{\mathcal{S}_{\mathrm{d}}}(f)|\mathcal{X}]$. We define

$$\widetilde{L}_{\mathcal{S}_{\mathrm{d}}}(f) := \sum_{X_i\in\mathcal{S}_{\mathrm{d}}}\frac{f(X_i)}{\mathbf{L}_{ii}} = \sum_{\mathbf{x}\in\mathcal{S}_{\mathrm{d}}}\frac{Nf(\mathbf{x})\rho(\mathbf{x})}{K(\mathbf{x},\mathbf{x})} = N\Lambda_{\mathcal{S}_{\mathrm{d}}}\Big(\frac{f\rho}{K}\Big).$$

Then, we have

$$\begin{aligned}\mathbb{E}\Big[\Big|\frac{L_{\mathcal{S}_{\mathrm{d}}}(f) - L(f)}{N}\Big|^2\Big|\mathcal{X}\Big] &= \mathrm{Var}\Big[\frac{1}{N}L_{\mathcal{S}_{\mathrm{d}}}(f)\Big|\mathcal{X}\Big]\\ &\le 2\mathrm{Var}\Big[\frac{1}{N}\widetilde{L}_{\mathcal{S}_{\mathrm{d}}}(f)\Big|\mathcal{X}\Big] + 2\mathrm{Var}\Big[\frac{L_{\mathcal{S}_{\mathrm{d}}}(f) - \widetilde{L}_{\mathcal{S}_{\mathrm{d}}}(f)}{N}\Big|\mathcal{X}\Big]\\ &= 2\mathrm{Var}\Big[\Lambda_{\mathcal{S}_{\mathrm{d}}}\Big(\frac{f\rho}{K}\Big)\Big|\mathcal{X}\Big] + 2\mathrm{Var}\Big[\frac{L_{\mathcal{S}_{\mathrm{d}}}(f) - \widetilde{L}_{\mathcal{S}_{\mathrm{d}}}(f)}{N}\Big|\mathcal{X}\Big]. \quad (17)\end{aligned}$$

We note that $K(\mathbf{x},\mathbf{y})$ is built by the Haar scaling function, which yields $K(\mathbf{x},\mathbf{x}) = n$ for a.e. $\mathbf{x} \in (0,1)^d$. This implies

$$\sup_{\mathbf{x}\in\mathrm{supp}(\rho)}\frac{K(\mathbf{x},\mathbf{x})}{\rho(\mathbf{x})} \le \frac{n}{\rho_{\min}}.$$

Thus we can choose $K_{\max} = n/\rho_{\min}$. By Theorem 3.2, there is an event $\mathcal{A}$ with $\mathbb{P}(\mathcal{A}) \ge 1 - \delta - \delta'$ such that on $\mathcal{A}$ we have the following inequality

$$\mathrm{Var}\Big[\Lambda_{\mathcal{S}_{\mathrm{d}}}\Big(\frac{f\rho}{K}\Big)\Big|\mathcal{X}\Big] \le 2\mathrm{Var}\Big[\Lambda_{\mathcal{S}}\Big(\frac{f\rho}{K}\Big)\Big] + 4\Big\|\frac{f\rho}{K}\Big\|^2_{\infty}\frac{n^2}{\rho^2_{\min}}\mathscr{E}(n,N,\delta,\delta')$$

where $\mathscr{E}(n,N,\delta,\delta')$ is defined as (7). Note that $\|f\rho/K\|_{\infty} = \|f\rho\|_{\infty}/n \le \|f\|_{\infty}\|\rho\|_{\infty}/n$. Thus

$$\mathrm{Var}\Big[\Lambda_{\mathcal{S}_{\mathrm{d}}}\Big(\frac{f\rho}{K}\Big)\Big|\mathcal{X}\Big] \le 2\mathrm{Var}\Big[\Lambda_{\mathcal{S}}\Big(\frac{f\rho}{K}\Big)\Big] + 4\Big(\frac{\|f\|_{\infty}\|\rho\|_{\infty}}{\rho_{\min}}\Big)^2\mathscr{E}(n,N,\delta,\delta'). \quad (18)$$

Now observe that if $f \in C^{0,s}$ and $\rho \in C^1$, we have $f\rho \in C^{0,s}$. By Theorem 4.2, there exists $C > 0$ depending only on d and ϕ such that

$$\mathrm{Var}\Big[\Lambda_{\mathcal{S}}\Big(\frac{f\rho}{K}\Big)\Big] \le C|f\rho|^2_{C^{0,s}} n^{-1-2s/d}. \tag{19}$$

By (18) and (19), we deduce that on the event $\mathcal{A}$

$$\mathrm{Var}\Big[\Lambda_{\mathcal{S}_{\mathrm{d}}}\Big(\frac{f\rho}{K}\Big)\Big|\mathcal{X}\Big] \le 2C|f\rho|^2_{C^{0,s}} n^{-1-2s/d} + 4\Big(\frac{\|f\|_\infty \|\rho\|_\infty}{\rho_{\min}}\Big)^2 \mathscr{E}(n, N, \delta, \delta'). \tag{20}$$

On the other hand, we recall that by Lemma A.1, for each realization of the data set $\mathcal{X}$, we have $\mathcal{S}_{\mathrm{d}}$ is a discrete DPP on $\mathcal{X}$ defined by some projection matrix $\mathbf{K}$. Then $\mathbb{P}(X_i \in \mathcal{S}_{\mathrm{d}}|\mathcal{X}) = \mathbf{K}_{ii}$ and thus

$$L_{\mathcal{S}_{\mathrm{d}}}(f) = \sum_{X_i \in \mathcal{S}_{\mathrm{d}}} \frac{f(X_i)}{\mathbf{K}_{ii}}.$$

Hence

$$\begin{aligned}
\mathrm{Var}\Big[\frac{L_{\mathcal{S}_{\mathrm{d}}}(f) - \widetilde{L}_{\mathcal{S}_{\mathrm{d}}}(f)}{N}\Big|\mathcal{X}\Big] &= \frac{1}{N^2}\mathrm{Var}\Big[\sum_{X_i \in \mathcal{S}_{\mathrm{d}}} f(X_i)\Big(\frac{\mathbf{L}_{ii} - \mathbf{K}_{ii}}{\mathbf{K}_{ii}\mathbf{L}_{ii}}\Big)\Big|\mathcal{X}\Big] \\
&\le \frac{1}{N^2}\sum_{i=1}^N f(X_i)^2\Big(\frac{\mathbf{L}_{ii} - \mathbf{K}_{ii}}{\mathbf{K}_{ii}\mathbf{L}_{ii}}\Big)^2 \mathbf{K}_{ii}.
\end{aligned}$$

By Lemma A.1, if we write $\lambda_1(\mathbf{L}) \ge \lambda_2(\mathbf{L}) \ge \ldots \ge \lambda_N(\mathbf{L})$ for the eigenvalues of $\mathbf{L}$, then

$$|\mathbf{L}_{ii} - \mathbf{K}_{ii}| \le \Big(\max_{1\le j\le n} |\lambda_j(\mathbf{L}) - 1|\Big)\mathbf{K}_{ii}, \quad \forall i.$$

Thus

$$\mathrm{Var}\Big[\frac{L_{\mathcal{S}_{\mathrm{d}}}(f) - \widetilde{L}_{\mathcal{S}_{\mathrm{d}}}(f)}{N}\Big|\mathcal{X}\Big] \le \frac{\|f\|^2_\infty}{N^2}\Big(\max_{1\le j\le n} |\lambda_j(\mathbf{L}) - 1|^2\Big)\sum_{i=1}^N \frac{\mathbf{K}_{ii}}{\mathbf{L}^2_{ii}}.$$

We note that

$$\mathbf{L}_{ii} = \frac{K(X_i, X_i)}{N\rho(X_i)} = \frac{n}{N\rho(X_i)} \quad \text{almost surely}$$

since K is built by the Haar scaling function. By Lemma A.3, we have

$$\max_{1\le j\le n} |\lambda_j(\mathbf{L}) - 1|^2 = \|N^{-1}\Phi^\top \mathbf{D}(\rho^{-1})\Phi - \mathbf{I}_n\|^2_{\mathrm{op}} \le \|N^{-1}\Phi^\top \mathbf{D}(\rho^{-1})\Phi - \mathbf{I}_n\|^2_{\mathrm{F}}.$$

Note that on $\mathcal{A}$ we have

$$\|N^{-1}\Phi^\top \mathbf{D}(\rho^{-1})\Phi - \mathbf{I}_n\|^2_{\mathrm{F}} \le \frac{n^2}{\rho^2_{\min}}\Big(\frac{4}{9N^2}\log^2\Big(\frac{n^2+1}{\delta}\Big) + \frac{4}{N}\log\Big(\frac{n^2+1}{\delta}\Big)\Big).$$

Combining all ingredients with $\sum_{i=1}^N \mathbf{K}_{ii} \le n$, we deduce that on $\mathcal{A}$

$$\mathrm{Var}\Big[\frac{L_{\mathcal{S}_{\mathrm{d}}}(f) - \widetilde{L}_{\mathcal{S}_{\mathrm{d}}}(f)}{N}\Big|\mathcal{X}\Big] \le 4\Big(\frac{\|f\|_\infty \|\rho\|_\infty}{\rho_{\min}}\Big)^2\Big(\frac{n}{9N^2}\log^2\Big(\frac{n^2+1}{\delta}\Big) + \frac{n}{N}\log\Big(\frac{n^2+1}{\delta}\Big)\Big). \tag{21}$$

Combining (17), (20) and (21), on the event $\mathcal{A}$, we have

$$\begin{aligned}
\mathrm{Var}\Big[\frac{1}{N}L_{\mathcal{S}_{\mathrm{d}}}(f)\Big|\mathcal{X}\Big] &\le 4C|f\rho|^2_{C^{0,s}} n^{-1-2s/d} + 8\Big(\frac{\|f\|_\infty \|\rho\|_\infty}{\rho_{\min}}\Big)^2 \mathscr{E}(n, N, \delta, \delta') \\
&+ 8\Big(\frac{\|f\|_\infty \|\rho\|_\infty}{\rho_{\min}}\Big)^2\Big(\frac{n}{9N^2}\log^2\Big(\frac{n^2+1}{\delta}\Big) + \frac{n}{N}\log\Big(\frac{n^2+1}{\delta}\Big)\Big) \\
&= 4C|f\rho|^2_{C^{0,s}} n^{-1-2s/d} + 8\Big(\frac{\|f\|_\infty \|\rho\|_\infty}{\rho_{\min}}\Big)^2 \widetilde{\mathscr{E}}(n, N, \delta, \delta')
\end{aligned}$$

where

$$\begin{aligned}\widetilde{\mathscr{E}}(n,N,\delta,\delta') &:= \mathscr{E}(n,N,\delta,\delta') + \frac{n}{9N^2}\log^2\Big(\frac{n^2+1}{\delta}\Big) + \frac{n}{N}\log\Big(\frac{n^2+1}{\delta}\Big)\\ &= 4\sqrt{\frac{2\log(2/\delta')}{N}} + \frac{n+4}{9N^2}\log^2\Big(\frac{n^2+1}{\delta}\Big) + \frac{n+4}{N}\log\Big(\frac{n^2+1}{\delta}\Big).\end{aligned}$$

Therefore, for N large enough such that $\widetilde{\mathscr{E}}(n,N,\delta,\delta') \le n^{-1-2s/d}$, with probability at least $1-\delta-\delta'$ we have

$$\mathbb{E}\Big[\Big|\frac{L_{\mathcal{S}_\mathrm{d}}(f)-L(f)}{N}\Big|^2\Big|\mathcal{X}\Big] = \mathrm{Var}\Big[\frac{1}{N}L_{\mathcal{S}_\mathrm{d}}(f)\Big|\mathcal{X}\Big] \le Cn^{-1-2s/d}$$

for some C depends only on d and ρ as desired. Here we used the fact that $\|f\|_\infty \le 1$ and $|f|_{C^{0,s}} \le 1$ to decouple the dependence on f. □

C.4 Coreset guarantees for discrete wavelet DPPs

Theorem C.2 (Discrete wavelet DPPs as coresets)**.** *Let $\mathcal{F}$ be a class of test functions such that*

$$\|f\|_\infty \le 1, |f|_{C^{0,\alpha}([0,1]^d)} \le 1 \text{ and } |L(f)| \ge cN \text{ for some } c>0. \tag{F1}$$

We assume either

$$\dim \mathrm{span}_{\mathbb{R}}(\mathcal{F}) = \mathrm{d}_{\mathcal{F}} < \infty, \tag{F2}$$

or there exists a bounded subset $\Theta \subset \mathbb{R}^{\mathrm{d}_{\mathcal{F}}}$ such that

$$\mathcal{F} = \{f_\theta : \theta \in \Theta\} \text{ and } \|f_\theta - f_{\theta'}\|_\infty \le \ell\|\theta-\theta'\|, \forall \theta,\theta' \in \Theta. \tag{F3}$$

Then there exist $C_0, C_1 > 0$ depending only on c, d, and ρ such that for all $\delta,\delta' \in (0,1)$, with probability at least $1-\delta-\delta'$ in the data set $\mathcal{X}$, we have for $0 \le \varepsilon \le C_0 n^{-2\alpha/d}$:

(i) under (F1) and (F2):

$$\mathbb{P}\Big(\exists f \in \mathcal{F} : \Big|\frac{L_{\mathcal{S}_\mathrm{d}}(f)}{L(f)} - 1\Big| \ge \varepsilon \Big| \mathcal{X}\Big) \le 2\exp\Big(6\mathrm{d}_{\mathcal{F}} - C_1 n^{1+2\alpha/d}\varepsilon^2\Big),$$

(ii) under (F1) and (F3):

$$\mathbb{P}\Big(\exists f \in \mathcal{F} : \Big|\frac{L_{\mathcal{S}_\mathrm{d}}(f)}{L(f)} - 1\Big| \ge \varepsilon \Big| \mathcal{X}\Big) \le 2\exp\Big((C_2 - \log\varepsilon)\mathrm{d}_{\mathcal{F}} - C_1 n^{1+2\alpha/d}\varepsilon^2\Big)$$

for some $C_2 > 0$ depends only on Θ, ℓ, ρ and c,

provided that $\widetilde{\mathscr{E}}(n,N,\delta,\delta') \le n^{-1-2\alpha/d}$ and $n^{2\alpha/d}\rho_{\min}^2 \ge 8$.

Proof of Theorem C.2. Let us fix $\delta,\delta' \ge 0$ such that $\delta+\delta' < 1$. Let N be large enough such that $\widetilde{\mathscr{E}}(n,N,\delta,\delta') \le n^{-1-2\alpha/d}$. By Theorem 5.2, there exists an event $\mathcal{A}$ (measurable w.r.t. the data set $\mathcal{X}$) with probability at least $1-\delta-\delta'$, we have

$$\mathrm{Var}\Big[\frac{L_{\mathcal{S}_\mathrm{d}}(f)}{N}\Big|\mathcal{X}\Big] \le Cn^{-1-2\alpha/d}, \quad \forall\, f \in C^{0,\alpha}([0,1]^d) : \|f\|_\infty \le 1, |f|_{C^{0,\alpha}([0,1]^d)} \le 1.$$

Following the proof of Theorem 5.2, on the same event $\mathcal{A}$ we have

$$\begin{aligned}\|N^{-1}\Phi^\top \mathbf{D}(\rho^{-1})\Phi - \mathbf{I}_n\|_{\mathrm{F}}^2 &\le \frac{n^2}{\rho_{\min}^2}\Big(\frac{4}{9N^2}\log^2\Big(\frac{n^2+1}{\delta}\Big) + \frac{4}{N}\log\Big(\frac{n^2+1}{\delta}\Big)\Big)\\ &\le \frac{4n}{\rho_{\min}^2}\widetilde{\mathscr{E}}(n,N,\delta,\delta') \le \frac{4n^{-2\alpha/d}}{\rho_{\min}^2} \le \frac{1}{2}\end{aligned}$$

since $n^{2\alpha/d}\rho_{\min}^2 \ge 8$ by our assumption. This implies that $\mathrm{rank}(\Phi) = n$, and thus $|\mathcal{S}_\mathrm{d}| = n$ on this event $\mathcal{A}$.

Moreover, following the proof of Theorem 5.2, we have for every $i \in \{1,2,\ldots,N\}$

$$\mathbf{K}_{ii} \ge \mathbf{L}_{ii} - |\mathbf{K}_{ii} - \mathbf{L}_{ii}| \ge \mathbf{L}_{ii} - \|N^{-1}\Phi^\top\mathbf{D}(\rho^{-1})\Phi - \mathbf{I}_n\|_{\mathrm{F}}\mathbf{K}_{ii}.$$

Thus, on event $\mathcal{A}$ we have $\mathbf{L}_{ii} \leq 2\mathbf{K}_{ii}$ for every i. Now note that

$$\mathbf{L}_{ii} = \frac{n}{N\rho(X_i)} \geq \|\rho\|_\infty^{-1} \frac{n}{N} \quad \text{almost surely.}$$

Therefore, on this event $\mathcal{A}$, we have

$$\mathbf{K}_{ii} \geq \frac{1}{2\|\rho\|_\infty} \frac{n}{N}.$$

Applying Theorem C.3 with $V := Cn^{-1-2\alpha/d}$ completes the proof. □

Remark 5. *The parameter ε should be interpreted as the relative error tolerance of the coreset approximation: smaller values of ε correspond to higher accuracy. For example, for $d \geq 4$, we have $-1/2 \leq -2\alpha/d$; hence we can choose $\varepsilon = O(n^{-1/2-\alpha'/d})$ for any $\alpha' < \alpha$. Theorem C.2 then says that the discrete wavelet DPP construction provides a high-probability coreset guarantee with accuracy rate of order $n^{-1/2-\alpha'/d}$, for any $\alpha' < \alpha$. Thus, when $\alpha = 1$, we can achieve the accuracy rate close to $n^{-1/2-1/d}$, which is clearly better than the rate $n^{-1/2}$ of independent sampling, and the rate $n^{-1/2-1/(2d)}$ of discrete OPE-based DPPs (see Remark 6.1 in [5]).*

For reader's convenience, we cite here Theorem 4 in [5], after harmonizing notations.

Theorem C.3 (Theorem 4 in [5])**.** *Let $\mathcal{S}$ be a DPP with a Hermitian kernel $\mathbf{K}$ on a finite set $\mathcal{X} = \{x_1, \ldots, x_N\}$ with $\mathbb{E}[|\mathcal{S}|] = n$. Assuming that there exists $\eta > 0$ independent of n, N such that:*

$$\mathbf{K}_{ii} \geq \eta \frac{n}{N}, \quad \forall 1 \leq i \leq N.$$

Let $\mathcal{F}$ be a class of bounded functions on $\mathcal{X}$ such that

$$\|f\|_\infty \leq 1, |L(f)| \geq cN \text{ for some } c > 0 \text{ uniformly on } \mathcal{F} \tag{C1}$$

and either

$$\dim \operatorname{span}_{\mathbb{R}}(\mathcal{F}) = \mathrm{d}_{\mathcal{F}} < \infty, \tag{C2}$$

or there exists a bounded subset $\Theta \subset \mathbb{R}^{\mathrm{d}_{\mathcal{F}}}$ such that

$$\mathcal{F} = \{f_\theta : \theta \in \Theta\} \text{ and } \|f_\theta - f_{\theta'}\|_\infty \leq \ell\|\theta - \theta'\|, \forall \theta, \theta' \in \Theta. \tag{C3}$$

Let $L_{\mathcal{S}}(f) := \sum_{x \in \mathcal{S}} f(x)/\mathbb{P}(x \in \mathcal{S})$, and let

$$V \geq \sup_{f \in \mathcal{F}} \operatorname{Var}\Big[\frac{L_{\mathcal{S}}(f)}{N}\Big].$$

Then there exists a universal constant $A > 0$ such that for $0 \leq \varepsilon \leq \frac{4A\eta}{3c} nV$:

(i) under (C1) and (C2) we have

$$\mathbb{P}\Big(\exists f \in \mathcal{F} : \Big|\frac{L_{\mathcal{S}}(f)}{L(f)} - 1\Big| \geq \varepsilon\Big) \leq 2\exp\Big(6\mathrm{d}_{\mathcal{F}} - \frac{c^2\varepsilon^2}{16AV}\Big),$$

(ii) under (C1) and (C3), if $|\mathcal{S}| \leq Bn$, we have

$$\mathbb{P}\Big(\exists f \in \mathcal{F} : \Big|\frac{L_{\mathcal{S}}(f)}{L(f)} - 1\Big| \geq \varepsilon\Big) \leq 2\exp\Big(C\mathrm{d}_{\mathcal{F}} - \mathrm{d}_{\mathcal{F}}\log\varepsilon - \frac{c^2\varepsilon^2}{16AV}\Big),$$

where $C > 0$ depends only on Θ, ℓ, B, η and c.

D Additional numerical experiments

D.1 Integrating rough functions

We compare the variance of five quadrature methods for integrating functions on the unit interval. These include vanilla iid sampling (IID), the wavelet-based DPPs HAAR and DB2 introduced in this paper (the latter using the adjusted linear statistic $\widetilde{\mathcal{L}}$ in (10)), the Dirichlet DPP DIR considered in [8] and the OPE-based DPP OPE built from the first n Legendre polynomials.

We consider four test functions of varying regularity:

- $\varphi_\gamma(x) = \frac{1}{d}\sum_{i=1}^{d} \frac{f_\gamma(x_i)}{\int_0^1 f_\gamma}, \quad f_\gamma(t) = \frac{1}{d}\sum_{i=1}^{d} |t-1/2|^\gamma$ for $\gamma \in \{0.25, 0.75\}$,
- $\varphi_{\text{mixcos}}(x) = \frac{1}{d}\sum_{i=1}^{d} \frac{f_{\text{mc}}(x_i)}{\int_0^1 f_{\text{mc}}}, \quad f_{\text{mc}}(t) = 0.1\,|\cos(5\pi(t-1/2))| + (t-1/2)^2$,
- $\varphi_{\text{bump}}(x) = \prod_{i=1}^{d} \frac{f_{\text{bump}}(x_i)}{\int_0^1 f_{\text{bump}}}, \quad f_{\text{bump}}(t) = \exp(-0.1/(t(1-t)))$.

We note that the function φ_γ lies in the Hölder space $C^{0,\gamma}([0,1]^d)$ and in the Sobolev space $H^s((0,1)^d)$ for all $s < \gamma + 1/2$. The function φ_{mixcos}, even though non-differentiable, is Lipschitz (i.e $C^{0,1}([0,1]^d)$ and belongs to the space $H^s((0,1)^d)$ for all $s < 3/2$. The last function we consider is φ_{bump}, which is infinitely smooth.

For each function, we generate 100 independent realizations of every estimator and report the empirical variance against the number of samples n on a log-log scale. The results illustrate that all DPP-based estimators improve over the Monte Carlo rate n^{-1}. In $d = 1$, HAAR and DB2 offer the most significant variance reduction, comfortably outperforming both DIR and OPE. In $d = 2$, HAAR still offers the best performance. On the other hand, while DB2 still outperforms the non-wavelet DPP samplers for $\varphi_\gamma(x)$ for both $\gamma \in \{0.25, 0.75\}$, this advantage does not persist for the other two functions.

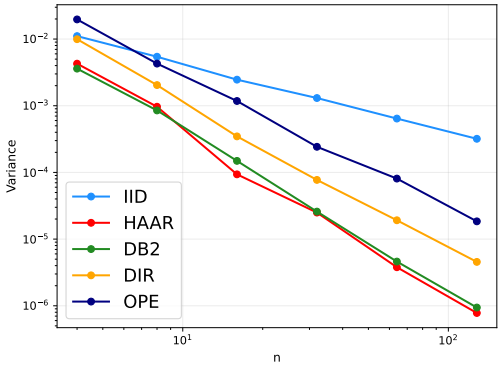

Figure 9: d = 1: $\varphi_{0.25}(x)$

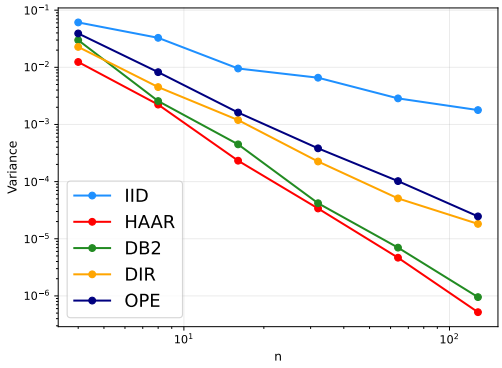

Figure 10: d = 1: $\varphi_{0.75}(x)$

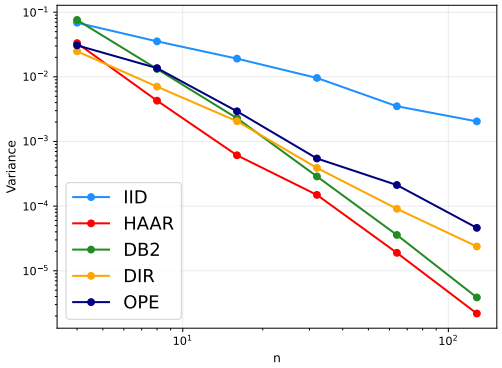

Figure 11: d = 1: $\varphi_{\text{mixcos}}(x)$

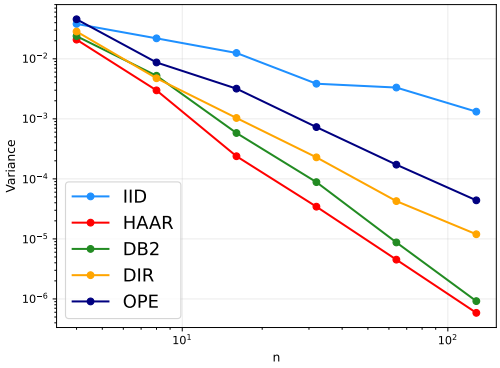

Figure 12: d = 1: $\varphi_{\text{bump}}(x)$

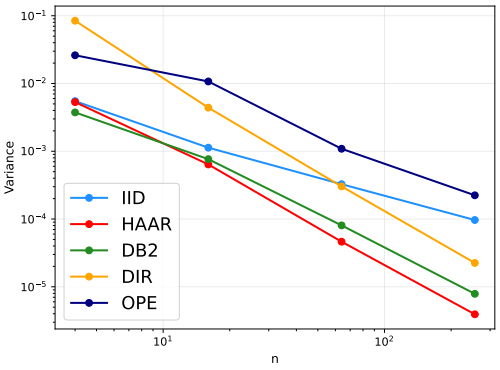

Figure 13: d = 2: $\varphi_{0.25}(x)$

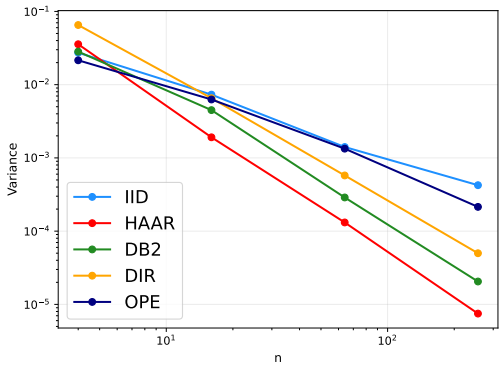

Figure 14: d = 2: $\varphi_{0.75}(x)$

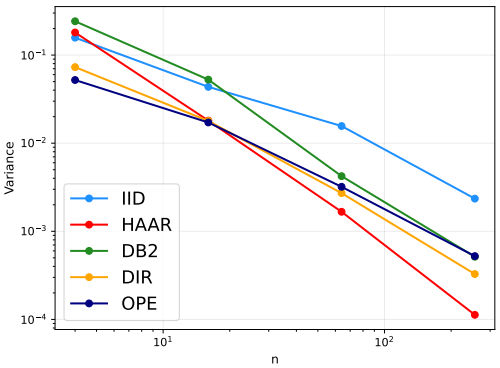

Figure 15: d = 2: $\varphi_{\text{mixcos}}(x)$

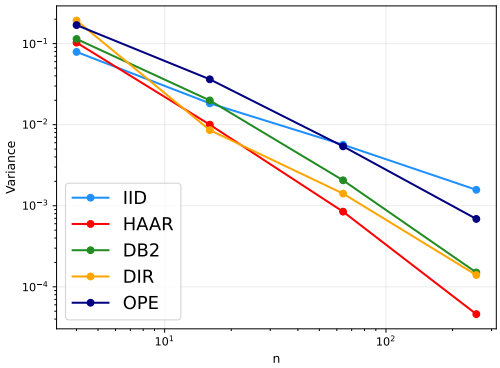

Figure 16: d = 2: $\varphi_{\text{bump}}(x)$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have provided all the relevant proofs and experimental results in Section 3, Section 4, Section 5 and Section 7.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed limitations and questions for future work in Section 8.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All the theorems are stated with full set of assumptions, and the complete proofs of these results are fully given in details in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The details of the experimental setup are provided in Section 7. Upon publication, we will release the code to reproduce our experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our code will be made publicly available upon publication.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The details are provided in Section 7.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Error bars, wherever applicable, are provided; c.f. Section 7.2.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper is mostly theoretical in nature, and the experiments we provide were run a single modern laptop; there shouldn't be any special computational need for re-executing out code.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our results do not violate any of the code of ethics provided by NeurIPS.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: Our work is primarily theoretical in nature and does not have any significant immediate societal impact beyond advancing knowledge.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We provide details on the dataset we use in Section 7.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: There are no new assets, but we have provided all the codes used for our empirical evaluation along with proper documentation as supplementary material.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.