Kernel Interpolation With Continuous Volume Sampling

Ayoub Belhadji [1] **Rémi Bardenet** [1] **Pierre Chainais** [1]

Abstract

A fundamental task in kernel methods is to pick nodes and weights, so as to approximate a given function from an RKHS by the weighted sum of kernel translates located at the nodes. This is the crux of kernel quadrature or kernel interpolation from discrete samples. Furthermore, RKHSs offer a convenient mathematical and computational framework, connecting the discrete and continuous worlds. We introduce and analyse continuous volume sampling (VS), the continuous counterpart – for choosing node locations – of a discrete distribution introduced in (Deshpande & Vempala, 2006). Our contribution is theoretical: we prove almost optimal bounds for interpolation and quadrature under VS. While similar bounds already exist for some specific RKHSs using ad-hoc node constructions, VS offers bounds that apply to any Mercer kernel and depend only on the spectrum of the associated integration operator. We emphasize that, unlike previous randomized approaches that rely on regularized leverage scores or determinantal point processes, evaluating the pdf of VS only requires pointwise evaluations of the kernel. VS is thus naturally amenable to MCMC samplers.

1. Introduction

Kernel approximation is a recurrent task in machine learning (Hastie, Tibshirani, and Friedman, 2009)[Chapter 5], signal processing (Unser, 2000) or numerical quadrature (Larkin, 1972). Expressed in its general form, we are given a reproducing kernel Hilbert space $\mathcal{F}$ (RKHS; Berlinet & Thomas-Agnan, 2011) of functions over $\mathcal{X}$, with a symmetric kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, and an element $\mu : \mathcal{X} \to \mathbb{R}$ of $\mathcal{F}$. We ask for conditions on a *design* $\boldsymbol{x} = (x_1, \dots, x_N) \in \mathcal{X}^N$, and on the corresponding weights $w_1, \dots, w_N$, such that the RKHS norm

$$\left\| \mu - \sum_{i=1}^{N} w_i k(x_i, .) \right\|_{\mathcal{F}} \tag{1}$$

is small. In other words, μ should be well reconstructed in $\mathcal{F}$ by the weighted sum.

Measuring the error in RKHS norm has a computational advantage. Indeed, minimizing (1) boils down to minimizing a quadratic form; and given a design $\boldsymbol{x}$ such that $\operatorname{Det} \boldsymbol{K}(\boldsymbol{x}) = \operatorname{Det}(k(x_i, x_j)) > 0$, Equation (1) has a unique set of minimizing weights. The minimizer corresponds to the weights $\hat{\boldsymbol{w}} = \boldsymbol{K}(\boldsymbol{x})^{-1}\mu(\boldsymbol{x})$, where $\mu(\boldsymbol{x}) \in \mathbb{R}^N$ contains the evaluation of μ at the N design nodes x_i. Whenever the weights are chosen to be $\hat{\boldsymbol{w}}$, the sum in (1) takes the same values as μ at the nodes x_i, and the optimal value of (1) is thus called *interpolation error*; otherwise we speak of *approximation error*. Note that when the kernel k is bounded, guarantees in RKHS norm translate to guarantees in the supremum norm.

In this work, we propose and analyze the interpolation based on a random design drawn from a distribution called *continuous volume sampling*, which favors designs $\boldsymbol{x}$ with a large value of $\operatorname{Det} \boldsymbol{K}(\boldsymbol{x})$. After introducing this new distribution, we prove non-asymptotic guarantees on the interpolation error which depend on the spectrum of the kernel k. Previous kernel-based randomized designs, both i.i.d. (Bach, 2017) and repulsive (Belhadji et al., 2019), can be hard to compute in practice since they require access to the Mercer decomposition of k. We show here that continuous volume sampling enjoys similar error bounds as well as some additional interpretable geometric properties, while having a joint density that can be evaluated as soon as one can evaluate the RKHS kernel k. In particular, this opens the possibility of Markov chain Monte Carlo samplers (Rezaei & Gharan, 2019).

Volume sampling was originally introduced on a finite domain (Deshpande et al., 2006), where it has been used in matrix subsampling for linear regression and low-rank approximations (Derezinski & Warmuth, 2017; Belhadji et al., 2018). Like (Belhadji et al., 2019), our work connects the discrete problem of subsampling columns from a matrix and the continuous problem of interpolating functions in an RKHS.

The rest of the article is organized as follows. Section 2

[1] Univ. Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL, Villeneuve d'Ascq, France. Correspondence to: Ayoub Belhadji <ayoub.belhadji@centralelille.fr>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

reviews kernel-based interpolation. In Section 3, we define continuous volume sampling and relate it to projection determinantal point processes, as used by (Belhadji et al., 2019). Section 4 contains our main results while Section 5 contains sketches of all proofs with pointers to the supplementary material for missing details. Section 6 numerically illustrates our main result. We conclude in Section 7, discussing some consequences beyond kernel interpolation.

Notation and assumptions. We assume that $\mathcal{X}$ is equipped with a Borel measure $\mathrm{d}\omega$, and that the support of $\mathrm{d}\omega$ is $\mathcal{X}$. Let $\mathbb{L}_2(\mathrm{d}\omega)$ be the Hilbert space of square integrable, real-valued functions on $\mathcal{X}$, with inner product $\langle \cdot, \cdot \rangle_{\mathrm{d}\omega}$, and associated norm $\|.\|_{\mathrm{d}\omega}$.

Assumption A $\int_{\mathcal{X}} k(x,x)\mathrm{d}\omega(x) < +\infty.$

Under Assumption A, define the integral operator

$$\boldsymbol{\Sigma} f(\cdot) = \int_{\mathcal{X}} k(\cdot, y) f(y) \mathrm{d}\omega(y), \quad f \in \mathbb{L}_2(\mathrm{d}\omega). \tag{2}$$

By construction, $\boldsymbol{\Sigma}$ is self-adjoint, positive semi-definite, and trace-class (Simon, 2005). For $m \in \mathbb{N}^*$, with $\mathbb{N}^* = \mathbb{N} \smallsetminus \{0\}$, denote by e_m the m-th eigenfunction of $\boldsymbol{\Sigma}$, normalized so that $\|e_m\|_{\mathrm{d}\omega} = 1$, and σ_m the corresponding eigenvalue. Assumption A implies that the embedding operator $I_{\mathcal{F}} : \mathcal{F} \longrightarrow \mathbb{L}_2(\mathrm{d}\omega)$ is compact; moreover, since $\mathrm{d}\omega$ is of full support in $\mathcal{X}$, $I_{\mathcal{F}}$ is injective (Steinwart & Christmann, 2008). This implies a Mercer-type decomposition of k,

$$k(x,y) = \sum_{m \in \mathbb{N}^*} \sigma_m e_m(x) e_m(y), \tag{3}$$

where the convergence is pointwise (Steinwart & Scovel, 2012). The eigenvalues (σ_m) are assumed to be non-increasing. Moreover, for $m \in \mathbb{N}^*$, we write $e_m^{\mathcal{F}} = \sqrt{\sigma_m} e_m$. Since $I_{\mathcal{F}}$ is injective, $(e_m^{\mathcal{F}})_{m \in \mathbb{N}^*}$ is an orthonormal basis of $\mathcal{F}$ (Steinwart & Scovel, 2012). Unless explicitly stated, we assume that $\mathcal{F}$ is dense in $\mathbb{L}_2(\mathrm{d}\omega)$, so that $(e_m)_{m \in \mathbb{N}^*}$ is an orthonormal basis of $\mathbb{L}_2(\mathrm{d}\omega)$. For more intuition, under these assumptions, $f \in \mathcal{F}$ if and only if $\sum_m \sigma_m^{-1} \langle f, e_m \rangle^2_{\mathbb{L}_2(\mathrm{d}\omega)}$ converges; and we denote for $r \geq 0$ $\|\boldsymbol{\Sigma}^{-r} f\|^2_{\mathcal{F}} = \sum_m \langle f, e_m^{\mathcal{F}} \rangle^2_{\mathcal{F}} / \sigma_m^{2r}$. For $\boldsymbol{x} \in \mathcal{X}^N$, we define $\boldsymbol{K}(\boldsymbol{x}) := k(x_i, x_j)_{i,j \in [N]}$. If $\operatorname{Det} \boldsymbol{K}(\boldsymbol{x}) > 0$, the subspace $\mathcal{T}(\boldsymbol{x}) = \operatorname{Span} k(x_i, .)_{i \in [N]}$ is of dimension N; we denote by $\Pi_{\mathcal{T}(\boldsymbol{x})}$ the $\langle ., . \rangle_{\mathcal{F}}$-orthogonal projection on $\mathcal{T}(\boldsymbol{x})$. Finally, for $N \in \mathbb{N}^*$, we will often sum over the sets

$$\mathcal{U}_N^m = \{U \subset \mathbb{N}^*, |U| = N, \; m \notin U\}, \tag{4}$$

$$\mathcal{U}_N = \{U \subset \mathbb{N}^*, |U| = N\}. \tag{5}$$

Finally, define the approximation error

$$\mathcal{E}(\mu; \boldsymbol{x}, \boldsymbol{w}) = \|\mu - \sum_{i \in [N]} w_i k(x_i, .)\|_{\mathcal{F}}, \tag{6}$$

where $[N] = \{1, \dots, N\}$. If $\operatorname{Det} \boldsymbol{K}(\boldsymbol{x}) > 0$, let $\hat{\boldsymbol{w}} = \boldsymbol{K}(\boldsymbol{x})^{-1} \mu(\boldsymbol{x})$ and define the interpolation error

$$\mathcal{E}(\mu; \boldsymbol{x}) = \|\mu - \sum_{i \in [N]} \hat{w}_i k(x_i, .)\|_{\mathcal{F}} \tag{7}$$

$$= \|\mu - \Pi_{\mathcal{T}(\boldsymbol{x})} \mu\|_{\mathcal{F}}. \tag{8}$$

2. Related Work

This section reviews some results on kernel interpolation to better situate our contributions. The literature on this topic is prolific and cannot be covered in details here. In particular, we start by reviewing results on optimal kernel quadrature, a particular case of kernel interpolation.

2.1. Interpolation for optimal kernel quadrature

Given $g \in \mathbb{L}_2(\mathrm{d}\omega)$, kernel quadrature deals with approximating the integrals

$$\int_{\mathcal{X}} f g \, \mathrm{d}\omega \approx \sum_{i \in [N]} w_i f(x_i), \quad f \in \mathcal{F}, \tag{9}$$

where the weights w_i do not depend on f. In principle, it is easy to control the integration error, i.e., the absolute value of the difference between the l.h.s. and r.h.s. of (9). Indeed,

$$\left| \int_{\mathcal{X}} f g \, \mathrm{d}\omega - \sum_{i \in [N]} w_i f(x_i) \right| \leq \|f\|_{\mathcal{F}} \, \mathcal{E}(\mu_g; \boldsymbol{x}, \boldsymbol{w}), \tag{10}$$

where $\mu_g = \int_{\mathcal{X}} g(x) k(x, .) \mathrm{d}\omega(x) = \boldsymbol{\Sigma} g$ is the so-called *embedding*[1] of g in the RKHS $\mathcal{F}$.

An upper bound on the approximation error of μ_g implies an upper bound on the integration error that is uniform over any bounded subset of $\mathcal{F}$. This observation sparked intense research on the kernel approximation of embeddings μ_g. Among kernel approximation results, we pay a particular attention to interpolation, i.e., approximation with optimal weights. In the sequel, we call *optimal kernel quadrature* the quadrature based on optimal weights $\hat{\boldsymbol{w}}$ minimizing (1) for a given set of nodes.

(Bojanov, 1981) proved that, for $g \equiv 1$, the interpolation of μ_g using the uniform grid over $\mathcal{X} = [0, 1]$ has an error in $\mathcal{O}(N^{-2s})$ if $\mathcal{F}$ is the periodic Sobolev space of order s, and that any set of nodes leads to that rate at least. A similar rate was proved for g not constant (Novak et al., 2015) even though it is only asymptotically optimal in that case.

In the quasi-Monte Carlo (QMC) literature, several designs were investigated for $\mathcal{X} = [0, 1]^d$, $g \equiv 1$ and $\mathcal{F}$ that may not

[1] When g is constant, μ_g is classically called the mean-element of the measure $\mathrm{d}\omega$ (Smola et al., 2007).

even be a Hilbert space; see (Dick & Pillichshammer, 2010). In this context, the term *QMC quadrature rule* means a low discrepancy sequence, loosely speaking a "well-spread" set of nodes, along with uniform weights $w_i = 1/N$. If $\mathcal{F}$ is a Korobov space of order $s \geq 1$, the Halton sequence of nodes (Halton, 1964) leads to $\mathcal{E}(\mu_g; \boldsymbol{x}, (1/N))^2$ in $\mathcal{O}(\log(N)^{2d} N^{-2})$ and higher-order digital nets converge faster as $\mathcal{O}(\log(N)^{2sd} N^{-2s})$ (Dick & Pillichshammer, 2014)[Theorem 5].

These rates are naturally inherited if the uniform weights are replaced by the respective optimal weights $\hat{\boldsymbol{w}}$, as observed by (Briol et al., 2019). In particular, (Briol et al., 2019) emphasize that the bound for higher-order digital nets attains the optimal rate in this RKHS. For optimal kernel quadrature based on Halton sequences, this inheritance argument does not explain the fast $\mathcal{O}(\log(N)^{2sd} N^{-2s})$ rates observed empirically by (Oettershagen, 2017).

Beside the hypercube, optimal kernel quadrature has been considered on the hypersphere equipped with the uniform measure (Ehler et al., 2019), or on $\mathbb{R}^d$ equipped with the Gaussian measure (Karvonen & Särkkä, 2019). In these works, the design construction is adhoc for the space $\mathcal{X}$ and g is usually assumed to be constant. Another approach is offered by optimization algorithms that we review in Section 2.3. Before that, we clarify the subtle difference between optimal kernel quadrature and kernel interpolation.

2.2. Kernel interpolation beyond embeddings

Besides the approximation of the embeddings μ_g discussed in Section 2.1, theoretical guarantees for the kernel interpolation of a general $\mu \in \mathcal{F}$ are sought *per se*. The Shannon reconstruction formula for bandlimited signals (Shannon, 1948) is implicitly an interpolation formula by the sinc kernel. The RKHS approach for sampling in signal processing was introduced in (Yao, 1967) for the Hilbert space of bandlimited signals; see also (Nashed & Walter, 1991) for generalizations. Remarkably, in those RKHSs, every $\mu \in \mathcal{F}$ is an embedding μ_g for some $g \in \mathbb{L}_2(\mathrm{d}\omega)$: k is a projection kernel of infinite rank. In general, for a trace-class kernel, the subspace spanned by the embeddings μ_g is strictly included in $\mathcal{F}$. More precisely, every μ_g satisfies

$$\|\boldsymbol{\Sigma}^{-1/2}\mu_g\|_{\mathcal{F}} = \|\boldsymbol{\Sigma}^{1/2} g\|_{\mathcal{F}} = \|g\|_{\mathbb{L}_2(\mathrm{d}\omega)} < +\infty.$$

This condition is more restrictive than what is required for a generic μ to belong to $\mathcal{F}$, i.e., $\|\mu\|_{\mathcal{F}} < +\infty$, so that kernel interpolation is more general than optimal kernel quadrature. The proposed approach will permit to deal with any $\mu \in \mathcal{F}$.

Scattered data approximation (Wendland, 2004) is another field where quantitative error bounds for kernel interpolation on $\mathcal{X} \subset \mathbb{R}^d$ are investigated; see (Schaback & Wendland, 2006) for a modern review. In a few words, these bounds typically depend on quantities such as the *fill-in distance* $\varphi(\boldsymbol{x}) = \sup_{y \in \mathcal{X}} \min_{i \in [N]} \|y - x_i\|_2$, so that the interpolation error converges to zero as $N \to \infty$ if $\varphi(\boldsymbol{x})$ goes to zero. Any node set can be considered, as long as $\varphi(\boldsymbol{x})$ is small. Using these techniques, (Oates & Girolami, 2016) proposed another application of kernel interpolation: the construction of functional control variates in Monte Carlo integration. Finally, note that the application of these techniques is restricted to compact domains: the fill-in distance is infinite if $\mathcal{X}$ is not compact, even for "well-spread" sets of nodes.

2.3. Optimization algorithms

Optimization approaches offer a variety of algorithms for the design of the interpolation nodes. (De Marchi, 2003) and (De Marchi et al., 2005) proposed greedily maximizing the so-called *power function*

$$p(x; \boldsymbol{x}) = \left[k(x,x) - k_{\boldsymbol{x}}(x)^{\intercal} \boldsymbol{K}(\boldsymbol{x})^{-1} k_{\boldsymbol{x}}(x)\right]^{1/2}, \quad (11)$$

where $k_{\boldsymbol{x}}(x) = (k(x, x_i))_{i \in [N]}$. This algorithm leads to an interpolation error that goes to zero with N for a kernel of class $\mathcal{C}^2$ (De Marchi et al., 2005). Later, (Santin & Haasdonk, 2017) proved better convergence rates for smoother kernels. Again, these results assume that the domain $\mathcal{X}$ is compact. Other greedy algorithms were proposed in the context of Bayesian quadrature (BQ) such as Sequential BQ (Huszár & Duvenaud, 2012), or Frank-Wolfe BQ (Briol et al., 2015). These algorithms sequentially minimize $\mathcal{E}(\mu_g; \boldsymbol{x})$, for a fixed $g \in \mathbb{L}_2(\mathrm{d}\omega)$. The nodes are thus adapted to one particular μ_g by construction. In general, each step of these greedy algorithms requires to solve a non-convex problem with many local minima (Oettershagen, 2017)[Chapter 5]. In practice, costly approximations must be employed such as local search in a random grid (Lacoste-Julien et al., 2015).

An alternative approach, that is very related to our contribution and has raised a lot of recent interest, is to observe that the squared power function (11) can be upper bounded by the inverse of Det $\boldsymbol{K}(\boldsymbol{x})$ (Schaback, 2005; Tanaka, 2019). Designs that maximize Det $\boldsymbol{K}(\boldsymbol{x})$ are called *Fekete points*; see e.g. (Bos & Maier, 2002; Bos & De Marchi, 2011). (Tanaka, 2019) proposed to approximate Det $\boldsymbol{K}(\boldsymbol{x})$ using the Mercer decomposition of k, followed by a rounding of the solution of a D-experimental design problem, yet without a theoretical analysis of the interpolation error. (Karvonen et al., 2019) proved that for the uni-dimensional Gaussian kernel, the approximate objective function of (Tanaka, 2019) is actually convex. Moreover, (Karvonen et al., 2019) analyze their interpolation error; see also Section 4.2. Finally, we emphasize that these algorithms require the knowledge of a Mercer-type decomposition of k so that they cannot be implemented for any kernel; moreover, the approximate objective function may be non-convex in general.

2.4. Random designs

In this section, we survey random node designs with uniform-in-g approximation guarantees for the embeddings μ_g in the RKHS norm. (Bach, 2017) studied the quadrature resulting from sampling i.i.d. nodes (x_j) from some proposal distribution q. He proved that when the proposal is chosen to be

$$q_\lambda^*(x) \propto \sum_{m\in\mathbb{N}^*} \frac{\sigma_m}{\sigma_m+\lambda} e_m(x)^2, \tag{12}$$

with $\lambda > 0$, and the number of points N satisfies $N \geq 5d_\lambda \log(16 d_\lambda/\delta)$ with $d_\lambda = \operatorname{Tr} \boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda \boldsymbol{I})^{-1}$, then with probability larger than $1-\delta$,

$$\sup_{\|g\|_{\mathrm{d}\omega}\leq 1} \inf_{\|\boldsymbol{w}\|^2 \leq \frac{4}{N}} \left\| \mu_g - \sum_{j\in[N]} \frac{w_j}{q_\lambda(x_j)^{1/2}} k(x_j,.) \right\|_{\mathcal{F}}^2 \leq 4\lambda. \tag{13}$$

The bound in (13) gives a control on the approximation error of μ_g by the subspace spanned by the $k(x_j,.)$, and this control is uniform over g in the unit ball of $\mathbb{L}_2(\mathrm{d}\omega)$. Note that for a fixed value of λ, the upper bound in (13) guarantees that the approximation error is smaller than 4λ. It does not however guarantee that the error goes to zero as N increases since it appears that λ should decrease as N increases. This coupling of N and λ combined with the condition $N \geq d_\lambda \log d_\lambda$ makes it intricate to derive a convergence rate from (13). Moreover, the optimal density q_λ^* is only implicitly available in general through the limit in (12), which makes sampling and pointwise evaluation difficult in practice.

(Belhadji et al., 2019) proposed a related kernel-based quadrature, but using nodes sampled from a repulsive joint distribution called a *projection determinantal point process* (DPP); see (Hough et al., 2006) and our Section 3. In particular, the repulsion is characterized by the first eigenfunctions $(e_n)_{n\in[N]}$ of the integration operator $\boldsymbol{\Sigma}$. The weights $\hat{\boldsymbol{w}}$ are chosen again by minimizing the residual error (1), which gives the uniform bound

$$\mathbb{E} \sup_{\|g\|_{\mathrm{d}\omega}\leq 1} \mathcal{E}(\mu_g; \boldsymbol{x})^2 \leq 2(N^2 r_N + o(N^2 r_N)), \tag{14}$$

where $r_N = \sum_{m\geq N+1} \sigma_m$. This result can be improved by further restricting g to be an eigenfunction of $\boldsymbol{\Sigma}$, leading to

$$\mathbb{E} \sup_{g\in\{e_n;\, n\geq 1\}} \mathcal{E}(\mu_g; \boldsymbol{x})^2 \leq 2(N r_N + o(N r_N)). \tag{15}$$

Now for smooth kernels, such as the Gaussian kernel or the Sobolev kernel with a large regularity parameter, the upper bounds in (14) and (15) do converge to 0 as N goes to $+\infty$. Furthermore, sampling from the recommended projection DPP can be implemented easily, although it still requires the knowledge of the Mercer decomposition of k, unlike the method that we introduce here in Section 3.

Since the bounds in (14) and (15) are uniform-in-g, they also concern interpolation. One downside of the analysis in (Belhadji et al., 2019) is that these upper bounds are rather pessimistic: experimental results suggest faster rates in $\mathcal{O}(\sigma_N)$. If one could prove these rates, then kernel quadrature or interpolation using DPPs would reach known lower bounds, which we now quickly survey.

2.5. Lower bounds

When investigating upper bounds for kernel interpolation errors, it is useful to remember existing lower bounds, so as to evaluate the tightness of one's results. In particular, N-widths theory (Pinkus, 2012) implies lower bounds for kernel interpolation errors, which once again show the importance of the spectrum of $\boldsymbol{\Sigma}$.

The N-width of $\mathcal{S} = \{\mu_g = \boldsymbol{\Sigma} g,\ \|g\|_{\mathbb{L}_2(\mathrm{d}\omega)} \leq 1\}$ with respect to the couple $(\mathbb{L}_2(\mathrm{d}\omega), \mathcal{F})$ (Pinkus, 2012, Chapter 1.7) is defined as the square root of

$$d_N(\mathcal{S})^2 = \inf_{\substack{Y\subset\mathcal{F} \\ \dim Y = N}} \sup_{\|g\|_{\mathrm{d}\omega}\leq 1} \inf_{y\in Y} \|\boldsymbol{\Sigma} g - y\|_{\mathcal{F}}^2.$$

In interpolation, we do use a subspace $Y \subset \mathcal{F}$ spanned by N independent functions $k(x_i,.)$, so that

$$\sup_{\|g\|_{\mathrm{d}\omega}\leq 1} \mathcal{E}(\boldsymbol{\Sigma} g; \boldsymbol{x})^2 \geq d_N(\mathcal{S})^2. \tag{16}$$

Applying (Pinkus, 2012, Theorem 2.2, Chapter 4) to the adjoint of the embedding operator $I_\mathcal{F}$ (Steinwart & Scovel, 2012)[Lemma 2.2], it comes $d_N(\mathcal{S})^2 = \sigma_{N+1}$. One may object that some QMC sequences seem to breach this lower bound. For example, in the Korobov space $(d = 2, s \geq 1)$, $\sigma_{N+1} = \mathcal{O}(\log(N)^{2s} N^{-2s})$ (Bach, 2017), while the interpolation of μ_g with $g \equiv 1$ using a Fibonacci lattice leads to an error in $\mathcal{O}(\log(N) N^{-2s}) = o(\sigma_{N+1})$ (Bilyk et al., 2012)[Theorem 4]. But this is the rate for one particular μ_g, and it cannot be achieved uniformly in g.

3. Volume Sampling and DPPs

In this section, we introduce a repulsive distribution that we call *continuous volume sampling* (VS) and compare it to projection determinantal point processes (DPPs; (Hough et al., 2006)). Both continuous VS and projection DPPs are parametrized using a reference measure $\mathrm{d}\omega$ and a repulsion kernel $\mathfrak{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$.

3.1. Continuous volume sampling

Definition 1 (Continuous volume sampling) *Let* $N \in \mathbb{N}^*$ *and* $\boldsymbol{x} = \{x_1, \dots, x_N\} \subset \mathcal{X}$. *We say that* $\boldsymbol{x}$ *follows the* *volume sampling distribution if $(x_1,\dots,x_N)$ is a random variable of $\mathcal{X}^N$, the law of which is absolutely continuous with respect to $\otimes_{i\in[N]}\mathrm{d}\omega(x_i)$, and the density writes*

$$f_{\mathrm{VS}}(x_1,\dots,x_N) \propto \mathrm{Det}\, \boldsymbol{K}(\boldsymbol{x}). \tag{17}$$

Two remarks are in order. First, under Assumption A, the density f_{VS} in (17) indeed integrates to 1. Indeed, Hadamard's inequality yields

$$\int_{\mathcal{X}^N} \mathrm{Det}\, \boldsymbol{K}(\boldsymbol{x}) \otimes \mathrm{d}\omega(\mathrm{x_i}) \leq \int_{\mathcal{X}^N} \prod_{i\in[N]} k(x_i,x_i) \otimes \mathrm{d}\omega(x_i)$$
$$= \left(\int_{\mathcal{X}} k(x,x)\mathrm{d}\omega(x)\right)^N < +\infty.$$

Second, the determinant in (17) is invariant to permutations, so that continuous volume sampling can indeed be seen as defining a random set $\boldsymbol{x} = \{x_1,\dots,x_N\}$.

In the following, we denote, for any symmetric and continuous kernel $\tilde{k}$ satisfying Assumption A,

$$Z_N(\tilde{k}) := \int_{\mathcal{X}^N} \mathrm{Det}\, \tilde{\boldsymbol{K}}(\boldsymbol{x}) \otimes \mathrm{d}\omega(x_i). \tag{18}$$

3.2. Continuous volume sampling as a mixture of DPPs

Definition 1 could be mistaken with the definition of a determinantal point process (DPP; Macchi, 1975). However, the cardinality of a DPP sample is a sum of Bernoulli random variables (Hough et al., 2006), while volume sampling is supported on subsets of $\mathcal{X}$ with cardinality exactly equal to N. This property is convenient for approximation tasks where the number of nodes N is fixed. While it is not a DPP, volume sampling is actually a mixture of DPPs.

Proposition 2 *For $U \subset \mathbb{N}^*$ define the projection kernel*

$$\mathfrak{K}_U(x,y) = \sum_{u\in U} e_u(x)e_u(y). \tag{19}$$

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

$$f_{\mathrm{VS}}(x_1,\dots,x_N) \propto \sum_{U\in\,\mathcal{U}_N} \prod_{u\in U} \sigma_u \,\mathrm{Det}(\mathfrak{K}_U(x_i,x_j))_{(i,j)}, \tag{20}$$

and the normalization constant is equal to

$$Z_N(k) = \mathrm{N}! \sum_{U\in\mathcal{U}_N} \prod_{u\in U} \sigma_u. \tag{21}$$

The proof of this proposition is given in Appendix 2.1. Observe that for every $U \subset \mathcal{U}_N$,

$$(x_1,\dots,x_N) \mapsto \frac{1}{N!}\mathrm{Det}(\mathfrak{K}_U(x_i,x_j))_{(i,j)\in[N]\times[N]}, \tag{22}$$

defines a well-normalized probability distribution on $\mathcal{X}^N$, called the *projection DPP* associated to the marginal kernel $\mathfrak{K}_U$ (Hough et al., 2006). Among all DPPs, only projection DPPs have a deterministic cardinality, equal to the rank of $\mathfrak{K}_U$ (Hough et al., 2006). Interestingly, the largest weight in the mixture (20) corresponds to the projection DPP of marginal kernel $\mathfrak{K}_{[N]}$ proposed in (Belhadji et al., 2019) for kernel quadrature. The following lemma gives an upper bound on this maximal weight using the eigenvalues of $\boldsymbol{\Sigma}$.

Lemma 3 *For $N \in \mathbb{N}^*$, define*

$$\delta_N = \prod_{n\in[N]} \sigma_n \Big/ \sum_{U\in\,\mathcal{U}_N} \prod_{u\in U} \sigma_u. \tag{23}$$

Then for all $N \in \mathbb{N}^*$, $\delta_N \leq \sigma_N/r_N$.

In particular, if the spectrum of k decreases polynomially, then $\delta_N = \mathcal{O}(1/N)$, so that as N grows, volume sampling becomes more different from the projection DPP of (Belhadji et al., 2019). In contrast, if the spectrum decays exponentially, then $\delta_N = \mathcal{O}(1)$.

3.3. Sampling algorithms

A projection DPP can be sampled exactly as long as one can evaluate the corresponding projection kernel $\mathfrak{K}$ (Hough et al., 2006). For kernel quadrature (Belhadji et al., 2019), evaluating $\mathfrak{K}$ requires the knowledge of the Mercer decomposition of the RKHS kernel k. The algorithm of (Hough et al., 2006) implements the chain rule for projection DPPs, and each conditional is sampled using rejection sampling. This suggests using the mixture in Proposition 2 to sample from the volume sampling distribution. Again, such an algorithm requires explicit knowledge of the Mercer decomposition of the kernel or at least a decomposition onto an orthonormal basis of $\mathcal{F}$ as in (Karvonen et al., 2019). This is a strong requirement that is undesirable in practice.

The fact that the joint pdf (17) only requires evaluating k pointwise suggests that volume sampling is *fully kernelized*, in the sense that a sampling algorithm should be able to bypass the need for a kernel decomposition, thus making the method very widely applicable. One could proceed by rejection sampling. Yet the acceptance ratio would likely scale poorly with N. A workaround would be to use an MCMC sampler similar to what was proposed in (Rezaei & Gharan, 2019). This MCMC algorithm is based on a Gibbs sampler chain: given a state $\boldsymbol{x} = \{x_1,\dots,x_N\}$, remove a node x_n chosen uniformly at random and add a node y with a probability proportional to $\mathrm{Det}\,\boldsymbol{K}(\boldsymbol{x}')$ where $\boldsymbol{x}' = \{x_1,\dots,x_{n-1},y,x_{n+1},\dots,x_N\}$, and the initial state of the Markov chain follow a distribution of density p_0 defined on $\mathcal{X}^N$ with respect to $\prod_{n\in[N]} \mathrm{d}\omega(x_n)$.

For this Markov chain, the authors were able to derive bounds for the mixing time:

$$\tau_{P_0}(\eta) = \min\{t|\ \|P_t - P_{\mathrm{VS}}\|_{\mathrm{TV}} \leq \eta\},$$

where $\|.\|_{\mathrm{TV}}$ is the total variation distance, P_t is the distribution of the Markov chain after t steps and P_{VS} is the distribution of the continuous volume sampling.

The author proposed the initialization of the Markov chain by a sequential algorithm and proved that the mixing time scales as $\mathcal{O}(N^5 \log(N))$. They also proved bounds on the expected number of rejections, which shows the feasibility of the implementation of the Gibbs steps. This sequential algorithm can be implemented in fully kernelized way without the need for the Mercer decomposition of k.

We leave investigating the efficiency of such an MCMC approach to volume sampling for future work.

4. Main Results

In this section, we give a theoretical analysis of kernel interpolation on nodes that follow the continuous volume sampling distribution. We state our main result in Section 4.1, an uniform-in-g upper bound of $\mathbb{E}_{\mathrm{VS}}\|\mu_g - \Pi_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2$. We give an upper bound for a general $\mu \in \mathcal{F}$ in Section 4.2.

4.1. The interpolation error for embeddings μ_g

The main theorem of this article decomposes the expected error for an embedding μ_g in terms of the expected errors $\epsilon_m(N)$ for eigenfunctions of the kernel.

Theorem 4 *Let* $g = \sum_{m\in\mathbb{N}^*} g_m e_m$ *satisfy* $\|g\|_{\mathrm{d}\omega} \leq 1$. *Then under Assumption A,*

$$\mathbb{E}_{\mathrm{VS}}\|\mu_g - \Pi_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2 = \sum_{m\in\mathbb{N}^*} g_m^2 \epsilon_m(N), \tag{24}$$

where $\epsilon_m(N) = \sigma_m \left(\sum_{U\in\mathcal{U}_N}\prod_{u\in U}\sigma_u\right)^{-1} \sum_{U\in\mathcal{U}_N^m}\prod_{u\in U}\sigma_u.$ *In particular, the sequence* $(\epsilon_m(N))_{m\in\mathbb{N}^*}$ *is non-increasing and*

$$\sup_{\|g\|_{\mathrm{d}\omega}\leq 1} \mathbb{E}_{\mathrm{VS}}\|\mu_g - \Pi_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2 \leq \sup_{m\in\mathbb{N}^*} \epsilon_m(N) = \epsilon_1(N). \tag{25}$$

Moreover,

$$\epsilon_1(N) \leq \sigma_N\left(1 + \beta_N\right), \tag{26}$$

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

In other words, under continuous volume sampling, $\epsilon_1(N)$ is a uniform upper bound on the expected squared interpolation error of *any* embedding μ_g such that $\|g\|_{\mathrm{d}\omega} \leq 1$. We shall see in Section 5.1 that $\epsilon_m(N) = \mathbb{E}_{\mathrm{VS}}\|\mu_{e_m} - \Pi_{\mathcal{T}(\boldsymbol{x})}\,\mu_{e_m}\|_{\mathcal{F}}^2$.

Now, for $N_0 \in \mathbb{N}^*$, a simple counting argument yields, for $m \geq N_0$, $\epsilon_m(N) \leq \sigma_{N_0}$. Actually, for $m \geq N_0$, $\|\mu_{e_m}\|_{\mathcal{F}}^2 \leq \sigma_{N_0}$, independently of the nodes.

Inequality (26) is less trivial and makes continuous volume sampling distribution worth of interest: the upper bound goes to 0 as $N \to +\infty$, below the initial error σ_{N_0}. Moreover, the convergence rate is $\mathcal{O}(\sigma_N)$, matching the lower bound of Section 2.5 if the sequence $(\beta_N)_{N\in\mathbb{N}^*}$ is bounded. In the following proposition, we prove that it is the case as soon as the spectrum decreases polynomially (e.g., Sobolev spaces of finite smoothness) or exponentially (e.g., the Gaussian kernel).

Proposition 5 *If* $\sigma_m = m^{-2s}$ *with* $s > 1/2$ *then*

$$\forall N \in \mathbb{N}^*,\ \beta_N \leq \left(1 + \frac{1}{2s-1}\right)\left(1 + \frac{1}{2s-1}\right)^{2s-1}. \tag{27}$$

If $\sigma_m = \alpha^m$, *with* $\alpha \in [0, 1[$, *then*

$$\forall N \in \mathbb{N}^*,\ \beta_N \leq \frac{\alpha}{1-\alpha}. \tag{28}$$

In both cases, the proof uses the fact that

$$\beta_N \leq [(N - M_N + 1)\sigma_N]^{-1} \sum_{m\geq M_N} \sigma_m, \tag{29}$$

for a well designed sequence M_N. For example, if $\sigma_m = m^{-2s}$, we take $M_N = \lceil N/c \rceil$ with $c > 1$; if $\sigma_m = \alpha^m$ we take $M_N = N$. We give a detailed proof in the supplementary.

For a general kernel, if an asymptotic equivalent of σ_N is known (Widom, 1963; 1964), it should be possible to give an explicit construction of M_N. Indeed,

$$\beta_N \leq \frac{\sigma_{M_N}}{\sigma_N} + [(N - M_N + 1)\sigma_N]^{-1} \sum_{m\geq N+1} \sigma_m, \tag{30}$$

and M_N should be chosen to control both terms in the RHS. Figure 1 illustrates the upper bound of Theorem 4 and the constant of Proposition 5 in case of the periodic Sobolev space of order $s = 3$. We observe that $\mathbb{E}_{\mathrm{VS}}\,\mathcal{E}(\mu_{e_m};\boldsymbol{x})^2$ respects the upper bound: it starts from the initial error level σ_m and decreases according to the upper bound for $N \geq m$.

4.2. The interpolation error of any element of $\mathcal{F}$

Theorem 4 dealt with the interpolation of an embedding μ_g of some function $g \in \mathbb{L}_2(\mathrm{d}\omega)$. We now give a bound on the interpolation error for any $\mu \in \mathcal{F}$. We need the following assumption, which is relatively weak; see Proposition 5 and the discussion that follows.

Assumption B *There exists $B > 0$ such that $\beta_N \leq B$.*

Theorem 6 *Let $\mu \in \mathcal{F}$. Assume that $\|\boldsymbol{\Sigma}^{-r}\mu\|_{\mathcal{F}} < +\infty$ for some $r \in [0, 1/2]$. Then, under Assumption B,*

$$\mathbb{E}_{\mathrm{VS}}\, \mathcal{E}(\mu; \boldsymbol{x})^2 \leq (2 + B)\sigma_N^{2r} \|\boldsymbol{\Sigma}^{-r}\mu\|_{\mathcal{F}}^2 = \mathcal{O}(\sigma_N^{2r}).$$

In other words, the expected interpolation error depends on the smoothness parameter r. For $r = 1/2$, we exactly recover the rate of Theorem 4. In contrast, for $r < 1/2$, the rate $\mathcal{O}(\sigma_N^{2r})$ is slower. For $r = 0$, our bound is constant with N. Note that assuming more smoothness ($r > 1/2$) does not seem to improve the rate $\mathcal{O}(\sigma_N)$.

Let us comment on this bound in two classical cases. First, consider the uni-dimensional Sobolev space of order s. Assumption B is satisfied by Proposition 5 and the squared error scales as $\mathcal{O}(N^{-4sr})$. Moreover, for this family of RKHSs, $\|\boldsymbol{\Sigma}^{-r}.\|_{\mathcal{F}}$ can be seen as the norm in the Sobolev space of order $(2r + 1)s$, and we recover a result in (Schaback & Wendland, 2006)[Theorem 7.8] for quasi-uniform designs. By using the norm in the RKHS $\mathcal{F}$ of rougher functions, we upper bound the interpolation error of μ belonging to the smoother RKHS $\boldsymbol{\Sigma}^r \mathcal{F}$. Second, we emphasize again that our result is agnostic to the choice of the kernel, as long as Assumption B holds. In particular, Theorem 6 applies to the Gaussian kernel: the rate is slower $\mathcal{O}(\sigma_N^{2r})$ yet still exponential. Finally, recall that for $f \in \mathcal{F}$

$$|f(x)|^2 = |\langle f, k(x, .)\rangle_{\mathcal{F}}|^2 \leq \|f\|_{\mathcal{F}}^2 k(x, x), \tag{31}$$

so that, bounds on the RKHS norm imply bounds on the uniform norm if the kernel k is bounded. Therefore, for $r \in [0, 1/2]$, our result improves on the rate $\mathcal{O}(N^2\sigma_N^{2r})$ of approximate Fekete points (Karvonen et al., 2019).

4.3. Asymptotic unbiasedness of kernel quadrature

As explained in Section 2.1, kernel interpolation is widely used for the design of quadratures. In that setting, one more advantage of continuous volume sampling is the consistency of its estimator. This is the purpose of the following result.

Theorem 7 *Let $f \in \mathcal{F}$, and $g \in \mathbb{L}_2(\mathrm{d}\omega)$. Then*

$$\begin{aligned}\mathcal{B}_N(f, g) &\triangleq \mathbb{E}_{\mathrm{VS}} \left(\int_{\mathcal{X}} fg \,\mathrm{d}\omega - \sum_{i \in N} \hat{w}_i f(x_i) \right). \\ &= \sum_{n \in \mathbb{N}^*} \langle f, e_n\rangle_{\mathrm{d}\omega} \langle g, e_n\rangle_{\mathrm{d}\omega} \left(1 - \mathbb{E}_{\mathrm{VS}}\, \tau_n^{\mathcal{F}}(\boldsymbol{x})\right).\end{aligned}$$

Moreover, $\mathcal{B}_N(f, g) \to 0$ as $N \to +\infty$.

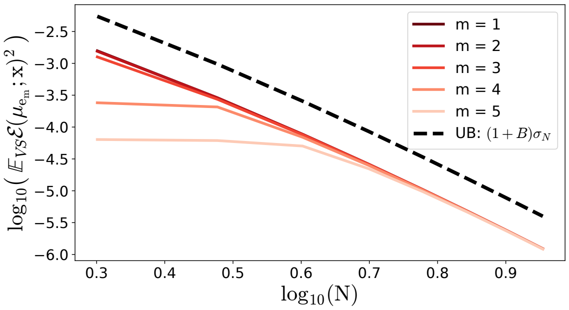

Figure 1. The value of $\mathbb{E}_{\mathrm{VS}}\, \mathcal{E}(\mu_{e_m}; \boldsymbol{x})^2$ for $m \in \{1, 2, 3, 4, 5\}$ for the periodic Sobolev space ($s = 3, d = 1$) compared to the theoretical upper bound (UB) of Theorem 4.

Compared to the upper bound on the integration error given by (10), the bias term in Theorem 7 takes into account the interaction between f and g. For example, if for all $n \in \mathbb{N}^*$, $\langle f, e_n\rangle_{\mathrm{d}\omega} \langle g, e_n\rangle_{\mathrm{d}\omega} = 0$, the quadrature is unbiased for every N. Theorem 7 is a generalization of a known property of regression estimators based on volume sampling in the discrete setting (Ben-Tal & Teboulle, 1990; Derezinski & Warmuth, 2017).

5. Sketch of the Proofs

The proof of Theorem 4 decomposes into three steps. First, in Section 5.1, we write $\mathcal{E}(\mu_g; \boldsymbol{x})^2$ as a function of the square of the interpolation errors $\mathcal{E}(\mu_{e_m}; \boldsymbol{x})^2$ of the embeddings μ_{e_m}. Then, in Section 5.2, we give closed formulas for $\mathbb{E}_{\mathrm{VS}}\, \mathcal{E}(\mu_{e_m}; \boldsymbol{x})^2$ in terms of the eigenvalues of $\boldsymbol{\Sigma}$. Finally, the inequality (26) is proved using an upper bound on the ratio of symmetric polynomials (Guruswami & Sinop, 2012). The details are given in Appendix 2.4.3. Finally, the proofs of Theorem 6 and Theorem 7 are straightforward consequences of Theorem 4. The details are given in Appendix 2.9 and Appendix 2.10.

5.1. Decomposing the interpolation error

Let $\boldsymbol{x} \in \mathcal{X}^N$ such that $\mathrm{Det}\, \boldsymbol{K}(\boldsymbol{x}) > 0$. For $m_1, m_2 \in \mathbb{N}^*$, let the *cross-leverage score* between m_1 and m_2 associated to $\boldsymbol{x}$ be

$$\tau_{m_1, m_2}^{\mathcal{F}}(\boldsymbol{x}) = e_{m_1}^{\mathcal{F}}(\boldsymbol{x})^{\intercal} \boldsymbol{K}(\boldsymbol{x})^{-1} e_{m_2}^{\mathcal{F}}(\boldsymbol{x}). \tag{32}$$

When $m_1 = m_2 = m$, we speak of the m-th leverage score[2] associated to $\boldsymbol{x}$, and simply write $\tau_m^{\mathcal{F}}(\boldsymbol{x})$. By Lemma S6, the m-th leverage score is related to the interpolation error of the m-th eigenfunction $e_m^{\mathcal{F}}$. Indeed,

$$\|e_m^{\mathcal{F}} - \Pi_{\mathcal{T}(\boldsymbol{x})} e_m^{\mathcal{F}}\|_{\mathcal{F}}^2 = 1 - \tau_m^{\mathcal{F}}(\boldsymbol{x}) \in [0, 1]. \tag{33}$$

[2] Our definition is consistent with the leverage scores used in matrix subsampling (Drineas et al., 2006). Loosely speaking, $\tau_m^{\mathcal{F}}(\boldsymbol{x})$ is the leverage score of the m-th column of the semi-infinite matrix $(e_n^{\mathcal{F}}(x_i))_{(i,n) \in [N] \times \mathbb{N}^*}$.

Similarly, for the cross-leverage score,

$$\langle \Pi_{\mathcal{T}(\boldsymbol{x})} e_{m_1}^{\mathcal{F}}, \Pi_{\mathcal{T}(\boldsymbol{x})} e_{m_2}^{\mathcal{F}} \rangle_{\mathcal{F}} = \tau_{m_1,m_2}^{\mathcal{F}}(\boldsymbol{x}) \in [-1,1]. \quad (34)$$

For $g \in \mathbb{L}_2(\mathrm{d}\omega)$, the interpolation error of the embedding μ_g can be expressed using the (cross-)leverage scores.

Lemma 8 *If* $\operatorname{Det} \boldsymbol{K}(\boldsymbol{x}) > 0$, *then,*

$$\begin{aligned} \mathcal{E}(\mu_g;\boldsymbol{x})^2 = & \sum_{m\in\mathbb{N}^*} g_m^2 \sigma_n \Big(1 - \tau_m^{\mathcal{F}}(\boldsymbol{x})\Big) \\ & - \sum_{m_1 \neq m_2 \in \mathbb{N}^*} g_{m_1} g_{m_2} \sqrt{\sigma_{m_1}} \sqrt{\sigma_{m_2}} \tau_{m_1,m_2}^{\mathcal{F}}(\boldsymbol{x}). \end{aligned} \quad (35)$$

In particular, with probability one, a design sampled from the continuous volume sampling distribution in Definition 1 satisfies (35). Furthermore, we shall see that the expected value of the (cross-) leverage scores has a simple expression.

5.2. Explicit formulas for expected leverage scores

Proposition 9 expresses expected leverage scores in terms of the spectrum of the integration operator.

Proposition 9 *For* $m \in \mathbb{N}^*$,

$$\mathbb{E}_{\mathrm{VS}}\, \tau_m^{\mathcal{F}}(\boldsymbol{x}) = \frac{1}{\sum\limits_{U \in\, \mathcal{U}_N} \prod\limits_{u \in U} \sigma_u} \sum_{\substack{U \in\, \mathcal{U}_N \\ m \in U}} \prod_{u\in U} \sigma_u. \quad (36)$$

Moreover, for $m_1, m_2 \in \mathbb{N}^*$ *such that* $m_1 \neq m_2$, *we have*

$$\mathbb{E}_{\mathrm{VS}}\, \tau_{m_1,m_2}^{\mathcal{F}}(\boldsymbol{x}) = 0. \quad (37)$$

In Appendix 2.4, we combine Lemma 8 with Proposition 9. This concludes the proof of Theorem 4 by Beppo Levi's monotone convergence theorem.

It remains to prove Proposition 9. Again, we proceed in two steps. First, our Proposition 10 yields a characterization of $\mathbb{E}_{\mathrm{VS}}\, \tau_m^{\mathcal{F}}(\boldsymbol{x})$ and $\mathbb{E}_{\mathrm{VS}}\, \tau_{m_1,m_2}^{\mathcal{F}}(\boldsymbol{x})$ in terms of the spectrum of three perturbed versions of the integration operator $\boldsymbol{\Sigma}$. Second, we give explicit forms of these spectra in Proposition 11 below. The idea is to express $\mathbb{E}_{\mathrm{VS}}\, \tau_m(\boldsymbol{x})^{\mathcal{F}}$ as the normalization constant (21) of a perturbation of the kernel k. The same goes for $\mathbb{E}_{\mathrm{VS}}\, \tau_{m_1,m_2}^{\mathcal{F}}(\boldsymbol{x})$.

Let $t \in \mathbb{R}_+$ and $\boldsymbol{\Sigma}_t$, $\boldsymbol{\Sigma}_t^+$ and $\boldsymbol{\Sigma}_t^-$ be the integration operators[3] on $\mathbb{L}_2(\mathrm{d}\omega)$, respectively associated with the kernels

$$k_t(x,y) = k(x,y) + t e_m^{\mathcal{F}}(x) e_m^{\mathcal{F}}(y), \quad (38)$$

$$\begin{aligned} k_t^+(x,y) = & k(x,y) \\ & + t\left(e_{m_1}^{\mathcal{F}}(x) + e_{m_2}^{\mathcal{F}}(x)\right)\left(e_{m_1}^{\mathcal{F}}(y) + e_{m_2}^{\mathcal{F}}(y)\right), \end{aligned} \quad (39)$$

$$\begin{aligned} k_t^-(x,y) = & k(x,y) \\ & + t\left(e_{m_1}^{\mathcal{F}}(x) - e_{m_2}^{\mathcal{F}}(x)\right)\left(e_{m_1}^{\mathcal{F}}(y) - e_{m_2}^{\mathcal{F}}(y)\right). \end{aligned} \quad (40)$$

By Assumption A, and by the fact that $(e_m)_{m\in\mathbb{N}^*}$ is an orthonormal basis of $\mathbb{L}_2(\mathrm{d}\omega)$, all three kernels also have integrable diagonals (see Assumption A). In particular, they define RKHSs that can be embedded in $\mathbb{L}_2(\mathrm{d}\omega)$. Moreover, recalling the definition (21) of the normalization constant Z_N of volume sampling, the following quantities are finite

$$\phi_m(t) = \frac{1}{N!} Z_N(k_t), \quad \phi_{m_1,m_2}^+(t) = \frac{1}{N!} Z_N(k_t^+),$$
$$\text{and} \quad \phi_{m_1,m_2}^-(t) = \frac{1}{N!} Z_N(k_t^-). \quad (41)$$

Remember that by Proposition 2,

$$\phi_m(t) = N! \sum_{U \in\, \mathcal{U}_N} \prod_{u\,\in U} \tilde{\sigma}_u(t), \quad (42)$$

where $\{\tilde{\sigma}_u(t),\ u \in \mathbb{N}^*\}$ is the set of eigenvalues[4] of $\boldsymbol{\Sigma}_t$. Similar identities are valid for $\phi_{m_1,m_2}^+(t)$ and $\phi_{m_1,m_2}^-(t)$ with the eigenvalues of $\boldsymbol{\Sigma}_t^+$ and $\boldsymbol{\Sigma}_t^-$ respectively.

Proposition 10 *The functions* ϕ_m, ϕ_{m_1,m_2}^+ *and* ϕ_{m_1,m_2}^- *are right differentiable in zero. Furthermore,*

$$\mathbb{E}_{\mathrm{VS}}\, \tau_m^{\mathcal{F}}(\boldsymbol{x}) = \frac{1}{Z_N(k)} \frac{\partial \phi_m}{\partial t}\bigg|_{t=0^+},$$

and

$$\mathbb{E}_{\mathrm{VS}}\, \tau_{m_1,m_2}^{\mathcal{F}}(\boldsymbol{x}) = \frac{1}{4 Z_N(k)} \left(\frac{\partial \phi_{m_1,m_2}^+}{\partial t} - \frac{\partial \phi_{m_1,m_2}^-}{\partial t}\right)\bigg|_{t=0^+}.$$

The details of the proof are postponed to Appendix 2.7. We complete this proposition with a description of the spectrum of the operators $\boldsymbol{\Sigma}_t$, $\boldsymbol{\Sigma}_t^+$ and $\boldsymbol{\Sigma}_t^-$ using the spectrum of $\boldsymbol{\Sigma}$.

Proposition 11 *The eigenvalues of* $\boldsymbol{\Sigma}_t$ *write*

$$\tilde{\sigma}_u(t) = \begin{cases} \sigma_u & \text{if } u \neq m, \\ (1+t)\sigma_u & \text{if } u = m. \end{cases} \quad (43)$$

Moreover, the eigenvalues of $\boldsymbol{\Sigma}_t^+$ *and* $\boldsymbol{\Sigma}_t^-$ *satisfy*

$$\{\tilde{\sigma}_u^+(t),\ u \in \mathbb{N}^*\} = \{\tilde{\sigma}_u^-(t),\ u \in \mathbb{N}^*\}. \quad (44)$$

The proof is based on the observation that the perturbations in (38), (39), and (40) only affect a principal subspace of dimension 1 or 2; see Appendix 2.6.

Combining the characterization of $\mathbb{E}_{\mathrm{VS}}\, \tau_m^{\mathcal{F}}(\boldsymbol{x})$ and $\mathbb{E}_{\mathrm{VS}}\, \tau_{m_1,m_2}^{\mathcal{F}}(\boldsymbol{x})$ given in Proposition 10, and Proposition 11, we prove Proposition 9; see details in Appendix 2.8.

[3] We drop from the notation the dependencies on m, m_1 and m_2 for simplicity.

[4] For a given value of t, the eigenvalues $\tilde{\sigma}_u(t)$ are not necessarily decreasing in u. We give explicit formulas for these eigenvalues in Proposition 11, and the order satisfied for $t = 0$ is not necessarily preserved for $t > 0$. This does not change anything to the argument since these eigenvalues only appear in quantities such as $\phi_m(t)$ which are invariant under permutation of the eigenvalues.

6. A Numerical Simulation

To illustrate Theorem 4, we let $\mathcal{X} = [0, 1]$ and $\mathrm{d}\omega$ be the uniform measure on $\mathcal{X}$. Let the RKHS kernel be (Berlinet & Thomas-Agnan, 2011)

$$k_s(x, y) = 1 + 2 \sum_{m \in \mathbb{N}^*} \frac{1}{m^{2s}} \cos(2\pi m(x - y)), \quad (45)$$

so that $\mathcal{F} = \mathcal{F}_s$ is the Sobolev space of order s on $[0, 1]$. The Mercer decomposition (3) of k_s is such that $\sigma_1 = 1$, $e_1 \equiv 1$ is constant and, for $n \geq 1$, $\sigma_{2n} = \sigma_{2n+1} = 1/n^{2s}$,

$$\begin{cases} e_{2n}(x) & = \sqrt{2}\cos(2\pi n x), \\ e_{2n+1}(x) & = \sqrt{2}\sin(2\pi n x). \end{cases} \quad (46)$$

Note that k_s can be expressed in closed form using Bernoulli polynomials (Wahba, 1990). In Theorem 4, (24) decomposes the expected interpolation error of any μ_g in terms of the interpolation error $\epsilon_m(N)$ of the μ_{e_m}. Therefore, it is sufficient to numerically check the values of the $\epsilon_m(N)$. As an illustration we consider $g \in \{e_1, e_5, e_7\}$ in (24), so that $\mu_{e_m} = \boldsymbol{\Sigma} e_m = \sigma_m e_m$, with $m \in \{1, 5, 7\}$. We use the Gibbs sampler proposed by (Rezaei & Gharan, 2019) to approximate continuous volume sampling.

We consider various numbers of points $N \in [2, 20]$. Figure 2 shows log-log plots of the theoretical (th) value of $\mathbb{E}_{\mathrm{VS}} \|\mu_g - \Pi_{\mathcal{T}(\boldsymbol{x})}\mu_g\|_{\mathcal{F}}^2$ compared to its empirical (emp) counterpart, vs. N, for $s \in \{1, 2\}$. For each N, the estimate is an average over 50 independent samples, each sample resulting from 500 individual Gibbs iterations.

For both values of the smoothness parameter s, we observe a close fit of the estimate with the actual expected error.

7. Discussion

We deal with interpolation in RKHSs using random nodes and optimal weights. This problem is intimately related to kernel quadrature, though interpolation is more general. We introduced continuous volume sampling (VS), a repulsive point process that is a mixture of DPPs, although not a DPP itself. VS comes with a set of advantages. First, interpretable bounds on the interpolation error can be derived under minimalistic assumptions. Our bounds are close to optimal since they share the same decay rate as known lower bounds. Moreover, we provide explicit evaluations of the constants appearing in our bounds for some particular RKHSs (e.g., Sobolev, Gaussian). Second, while the eigendecomposition of the integration operator plays an important role in the analysis, the definition of the density function of volume sampling only involves kernel evaluations. In that sense, VS is a fully kernelized approach. Unlike previous work on random design, this may permit sampling without knowing the Mercer decomposition of the kernel (Rezaei & Gharan, 2019), as demonstrated in Section 6. Investigating efficient samplers and their impact on bounds is deferred to future work; the current paper is a theoretical motivation for further methodological research.

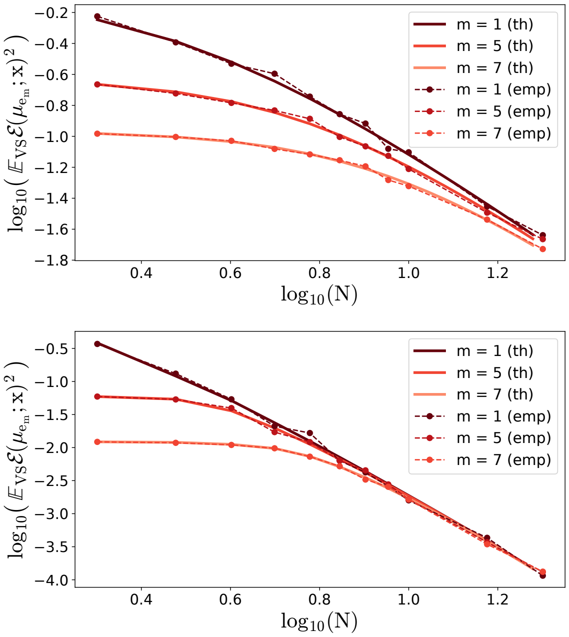

Figure 2. The empirical estimate of $\mathbb{E}_{\mathrm{VS}}\, \mathcal{E}(\mu_{e_m}; \boldsymbol{x})^2$ for $m \in \{1, 5, 7\}$ compared to its expression (24) in the case of the periodic Sobolev space. The smoothness is $s = 1$ (top), $s = 2$ (bottom).

We conclude with a few general remarks. In the context of interpolation, (Karvonen et al., 2019) propose to use approximate Fekete points. In comparison, our analysis yields a sharper upper bound while circumventing the analysis of the Lebesgue constant, a central quantity in interpolation theory. Yet, it would be interesting to analyze the distribution of the Lebesgue constant under continuous volume sampling, to provide an assessment of the numerical stability of our approach. A related extension of our work would be the analysis of interpolation under regularization as in (Bach, 2017). Finally, while optimal kernel quadrature may be the main application of our results, Section 4.2 hints that more generic elements than embeddings can also be well approximated using VS. This connects to kernel quadrature in misspecified settings (Kanagawa et al., 2016), or experimental design problems for Gaussian process approximations (Wynne et al., 2020).

Acknowledgements

AB acknowledges support from ANR grant BoB (ANR-16-CE23-0003), Région Hauts-de-France, and Centrale Lille. RB acknowledges support from ERC grant BLACKJACK (ERC-2019-STG-851866).

References

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

Belhadji, A., Bardenet, R., and Chainais, P. A determinantal point process for column subset selection. *arXiv preprint arXiv:1812.09771*, 2018.

Belhadji, A., Bardenet, R., and Chainais, P. Kernel quadrature with DPPs. In *Advances in Neural Information Processing Systems 32*, pp. 12907–12917. 2019.

Ben-Tal, A. and Teboulle, M. A geometric property of the least squares solution of linear equations. *Linear algebra and its applications*, 139:165–170, 1990.

Berlinet, A. and Thomas-Agnan, C. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

Bilyk, D., Temlyakov, V. N., and Yu, R. The L_2 discrepancy of two-dimensional lattices. In *Recent Advances in Harmonic Analysis and Applications*, pp. 63–77. Springer, 2012.

Bojanov, B. D. Uniqueness of the optimal nodes of quadrature formulae. *Mathematics of computation*, 36(154): 525–546, 1981.

Bos, L. P. and De Marchi, S. On optimal points for interpolation by univariate exponential functions. *Dolomites Research Notes on Approximation*, 4(1), 2011.

Bos, L. P. and Maier, U. On the asymptotics of Fekete-type points for univariate radial basis interpolation. *Journal of Approximation Theory*, 119(2):252–270, 2002.

Briol, F. X., Oates, C., Girolami, M., and Osborne, M. A. Frank-Wolfe Bayesian quadrature: Probabilistic integration with theoretical guarantees. In *Advances in Neural Information Processing Systems*, pp. 1162–1170, 2015.

Briol, F. X., Oates, C. J., Girolami, M., Osborne, M. A., Sejdinovic, D., et al. Probabilistic integration: A role in statistical computation? *Statistical Science*, 34(1):1–22, 2019.

De Marchi, S. On optimal center locations for radial basis function interpolation: computational aspects. *Rend. Splines Radial Basis Functions and Applications*, 61(3): 343–358, 2003.

De Marchi, S., Schaback, R., and Wendland, H. Near-optimal data-independent point locations for radial basis function interpolation. *Advances in Computational Mathematics*, 23(3):317–330, 2005.

Derezinski, M. and Warmuth, M. K. Unbiased estimates for linear regression via volume sampling. In *Advances in Neural Information Processing Systems*, pp. 3084–3093, 2017.

Deshpande, A. and Vempala, S. Adaptive sampling and fast low-rank matrix approximation. In *Proceedings of the 9th International Conference on Approximation Algorithms for Combinatorial Optimization Problems, and 10th International Conference on Randomization and Computation*, APPROX'06/RANDOM'06, pp. 292–303. Springer-Verlag, 2006.

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

Dick, J. and Pillichshammer, F. *Digital nets and sequences: discrepancy theory and quasi–Monte Carlo integration*. Cambridge University Press, 2010.

Dick, J. and Pillichshammer, F. Discrepancy theory and quasi-Monte Carlo integration. In *A panorama of discrepancy theory*, pp. 539–619. Springer, 2014.

Drineas, P., Mahoney, M. W., and Muthukrishnan, S. Sampling algorithms for $\ell 2$ regression and applications. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pp. 1127–1136. Society for Industrial and Applied Mathematics, 2006.

Ehler, M., Gräf, M., and Oates, C. J. Optimal Monte Carlo integration on closed manifolds. *Statistics and Computing*, 29(6):1203–1214, 2019.

Guruswami, V. and Sinop, A. K. Optimal column-based low-rank matrix reconstruction. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pp. 1207–1214. SIAM, 2012.

Halton, J. H. Algorithm 247: Radical-inverse quasi-random point sequence. *Communications of the ACM*, 7(12): 701–702, 1964.

Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

Hough, J. B., Krishnapur, M., Peres, Y., and Virág, B. Determinantal processes and independence. *Probability surveys*, 3:206–229, 2006.

Huszár, F. and Duvenaud, D. Optimally-weighted herding is Bayesian quadrature. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI'12, pp. 377–386. AUAI Press, 2012.

Kanagawa, M., Sriperumbudur, B. K., and Fukumizu, K. Convergence guarantees for kernel-based quadrature rules in misspecified settings. In *Advances in Neural Information Processing Systems*, pp. 3288–3296, 2016.

Karvonen, T. and Särkkä, S. Gaussian kernel quadrature at scaled Gauss-Hermite nodes. *BIT Numerical Mathematics*, pp. 1–26, 2019.

Karvonen, T., Särkkä, S., and Tanaka, K. Kernel-based interpolation at approximate Fekete points. *arXiv preprint arXiv:1912.07316*, 2019.

Lacoste-Julien, S., Lindsten, F., and Bach, F. Sequential kernel herding: Frank-wolfe optimization for particle filtering. *arXiv preprint arXiv:1501.02056*, 2015.

Larkin, F. M. Gaussian measure in Hilbert space and applications in numerical analysis. *The Rocky Mountain Journal of Mathematics*, pp. 379–421, 1972.

Macchi, O. The coincidence approach to stochastic point processes. 7:83–122, 03 1975.

Nashed, M. Z. and Walter, G. G. General sampling theorems for functions in reproducing kernel Hilbert spaces. *Mathematics of Control, Signals and Systems*, 4(4):363, 1991.

Novak, E., Ullrich, M., and Woźniakowski, H. Complexity of oscillatory integration for univariate sobolev spaces. *Journal of Complexity*, 31(1):15–41, 2015.

Oates, C. and Girolami, M. Control functionals for quasi-monte carlo integration. In Gretton, A. and Robert, C. C. (eds.), *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 56–65, Cadiz, Spain, 09–11 May 2016. PMLR.

Oettershagen, J. *Construction of optimal cubature algorithms with applications to econometrics and uncertainty quantification*. PhD Thesis, University of Bonn, 2017.

Pinkus, A. *N-widths in Approximation Theory*, volume 7. Springer Science & Business Media, 2012.

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

Santin, G. and Haasdonk, B. Convergence rate of the data-independent p-greedy algorithm in kernel-based approximation. *Dolomites Research Notes on Approximation*, 10 (Special_Issue), 2017.

Schaback, R. Multivariate interpolation by polynomials and radial basis functions. *Constructive Approximation*, 21 (3):293–317, 2005.

Schaback, R. and Wendland, H. Kernel techniques: from machine learning to meshless methods. *Acta numerica*, 15:543–639, 2006.

Shannon, C. E. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

Simon, B. *Trace Ideals and Their Applications*. American Mathematical Society, 2005.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. A Hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pp. 13–31. Springer, 2007.

Steinwart, I. and Christmann, A. *Support Vector Machines*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387772413.

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

Tanaka, K. Generation of point sets by convex optimization for interpolation in reproducing kernel Hilbert spaces. *Numerical Algorithms*, pp. 1–31, 2019.

Unser, M. Sampling-50 years after shannon. *Proceedings of the IEEE*, 88(4):569–587, 2000.

Wahba, G. *Spline Models for Observational Data*, volume 59. SIAM, 1990.

Wendland, H. *Scattered Data Approximation*. Cambridge University Press, 2004.

Widom, H. Asymptotic behavior of the eigenvalues of certain integral equations. I. *Transactions of the American Mathematical Society*, 109(2):278–295, 1963.

Widom, H. Asymptotic behavior of the eigenvalues of certain integral equations. II. *Archive for Rational Mechanics and Analysis*, 17(3):215–229, 1964.

Wynne, G., Briol, F. X., and Girolami, M. Convergence guarantees for gaussian process approximations under several observation models. *arXiv preprint arXiv:2001.10818*, 2020.

Yao, K. Applications of reproducing kernel Hilbert spaces-bandlimited signal models. *Information and Control*, 11 (4):429–444, 1967.