On the Equivalence between Kernel Quadrature Rules and Random Feature Expansions

Francis Bach

On the Equivalence between Kernel Quadrature Rules and Random Feature Expansions

Francis Bach FRANCIS.BACH@ENS.FR
INRIA
Département d'Informatique de l'Ecole Normale Supérieure
Paris, France

Editor:

Abstract

We show that kernel-based quadrature rules for computing integrals can be seen as a special case of random feature expansions for positive definite kernels, for a particular decomposition that always exists for such kernels. We provide a theoretical analysis of the number of required samples for a given approximation error, leading to both upper and lower bounds that are based solely on the eigenvalues of the associated integral operator and match up to logarithmic terms. In particular, we show that the upper bound may be obtained from independent and identically distributed samples from a specific non-uniform distribution, while the lower bound if valid for any set of points. Applying our results to kernel-based quadrature, while our results are fairly general, we recover known upper and lower bounds for the special cases of Sobolev spaces. Moreover, our results extend to the more general problem of full function approximations (beyond simply computing an integral), with results in L_2- and L_∞-norm that match known results for special cases. Applying our results to random features, we show an improvement of the number of random features needed to preserve the generalization guarantees for learning with Lipshitz-continuous losses.

1. Introduction

The numerical computation of high-dimensional integrals is one of the core computational tasks in many areas of machine learning, signal processing and more generally applied mathematics, in particular in the context of Bayesian inference (Gelman, 2004), or the study of complex systems (Robert and Casella, 2005). In this paper, we focus on *quadrature rules*, that aim at approximating the integral of a certain function from only the (potentially noisy) knowledge of the function values at as few as possible well-chosen points. Key situations that remain active areas of research are problems where the measurable space where the function is defined on is either high-dimensional or structured (e.g., presence of discrete structures, or graphs). For these problems, techniques based on *positive definite kernels* have emerged as having the potential to efficiently deal with these situations, and to improve over plain Monte-Carlo integration (O'Hagan, 1991; Rasmussen and Ghahramani, 2003; Huszár and Duvenaud, 2012; Oates and Girolami, 2015). In particular, the quadrature problem may be cast as the one of approximating a fixed element, the mean element (Smola et al., 2007), of a Hilbert space as a linear combination of well chosen el-

ements, the goal being to minimize the number of these factors as it corresponds to the required number of function evaluations.

A seemingly unrelated problem on positive definite kernels have recently emerged, namely the representation of the corresponding infinite-dimensional feature space from *random sets of features*. If a certain positive definite kernel between two points may be represented as the expectation of the product of two random one-dimensional (typically non-linear) features computed on these two points, the full kernel (and hence its feature space) may be approximated by sufficiently many random samples, replacing the expectation by a sample average (Neal, 1995; Rahimi and Recht, 2007; Huang et al., 2006). When using these random features, the complexity of a regular kernel method such as the support vector machine or ridge regression goes from quadratic in the number of observations to linear in the number of observations, with a constant proportional to the number of random features, which thus drives the running time complexity of these methods.

In this paper, we make the following contributions:

– After describing the functional analysis framework our analysis is based on and presenting many examples in Section 2, we show in Section 3 that these two problems are strongly related; more precisely, optimizing weights in kernel-based quadrature rules can be seen as decomposing a certain function in a special class of random features for a particular decomposition that always exists for all positive definite kernels on a measurable space.

– We provide in Section 4 a theoretical analysis of the number of required samples for a given approximation error, leading to both upper and lower bounds that are based solely on the eigenvalues of the associated integral operator and match up to logarithmic terms. In particular, we show that the upper bound may be obtained as independent and identically distributed samples from a specific non-uniform distribution, while the lower bound if valid for any set of points.

– Applying our results to kernel quadrature, while our results are fairly general, we recover known upper and lower bounds for the special cases of Sobolev spaces (Section 4.4). Moreover, our results extend to the more general problem of full function approximations (beyond simply computing an integral), with results in L_2- and L_∞-norm that match known results for special cases (Section 5).

– Applying our results to random feature expansions, we show in Section 4.5 an improvement of the number of random features needed for preserving the generalization guarantees for learning with Lipshitz-continuous losses.

2. Random Feature Expansions of Positive Definite Kernels

Throughout this paper, we consider a topological space $\mathcal{X}$ equipped with a Borel probability measure $d\rho$, which we assume to have full support. This naturally defines the space of square-integrable functions[1].

1. For simplicity and following most of the literature on kernel methods, we identify functions and their equivalence classes for the equivalence relationship of being equal except for a zero-measure (for $d\rho$) subset of $\mathcal{X}$.

2.1 Reproducing kernel Hilbert spaces and integral operators

We consider a continuous positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, that is a symmetric function such that for all finite families of points in $\mathcal{X}$, the matrix of pairwise kernel evaluations is positive semi-definite. This thus defines a reproducing kernel Hilbert space (RKHS) $\mathcal{F}$ of functions from $\mathcal{X}$ to $\mathbb{R}$, which we also assume separable. This RKHS has two important characteristic properties (see, e.g., Berlinet and Thomas-Agnan, 2004):

(a) *Membership of kernel evaluations*: for any $x \in \mathcal{X}$, the function $k(\cdot, x) : y \mapsto k(y, x)$ is an element of $\mathcal{F}$.

(b) *Reproducing property*: for all $f \in \mathcal{F}$ and $x \in \mathcal{X}$, $f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{F}}$. In other words, function evaluations are equal to dot-products with a specific element of $\mathcal{F}$.

Moreover, throughout the paper, we assume that the function $x \mapsto k(x, x)$ is integrable with respect to $d\rho$ (which is weaker than $\sup_{x\in\mathcal{X}} k(x, x) < \infty$). This implies that $\mathcal{F}$ is a subset of $L_2(d\rho)$; that is, functions in the RKHS $\mathcal{F}$ are all square-integrable for $d\rho$. In general, $\mathcal{F}$ is strictly included in $L_2(d\rho)$, but, in this paper, we will always assume that it is *dense* in $L_2(d\rho)$, that is, any function in $L_2(d\rho)$ may be approximated arbitrarily closely by a function in $\mathcal{F}$. Finally, for simplicity of our notation (to make sure that the sequence of eigenvalues of integral operators is infinite) we will always assume that $L_2(d\rho)$ is infinite-dimensional, which excludes finite sets for $\mathcal{X}$. Note that the last two assumptions (denseness and infinite dimensionality) can easily be relaxed.

Integral operator. Reproducing kernel Hilbert spaces are often studied through a specific integral operator which leads to an isometry with $L_2(d\rho)$ (Smale and Cucker, 2001). Let $\Sigma : L_2(d\rho) \to L_2(d\rho)$ be defined as

$$(\Sigma f)(x) = \int_{\mathcal{X}} f(y)k(x, y)d\rho(y).$$

Since $\int_{\mathcal{X}} k(x, x)d\rho(x)$ is finite, Σ is self-adjoint, positive semi-definite and trace-class (Simon, 1979). Given that Σf is a linear combination of kernel functions $k(\cdot, y)$, it belongs to $\mathcal{F}$. More precisely, since we have assumed that $\mathcal{F}$ is dense in $L_2(d\rho)$, $\Sigma^{1/2}$, which is the unique positive self-adjoint square root of Σ, is a bijection from $L_2(d\rho)$ to our RKHS $\mathcal{F}$; that is, for any $f \in \mathcal{F}$, there exists a unique $g \in L_2(d\rho)$ such that $f = \Sigma^{1/2}g$ and $\|f\|_{\mathcal{H}} = \|g\|_{L_2(d\rho)}$ (Smale and Cucker, 2001). This justifies the notation $\Sigma^{-1/2}f$ for $f \in \mathcal{F}$ and means that $\Sigma^{1/2}$ is an isometry from $L_2(d\rho)$ to $\mathcal{F}$; in other words, for any functions f and g in $\mathcal{F}$, we have:

$$\langle f, g\rangle_{\mathcal{F}} = \langle \Sigma^{-1/2}f, \Sigma^{-1/2}g\rangle_{L_2(d\rho)}.$$

This justifies the view of $\mathcal{F}$ as the subspace of functions $f \in L_2(d\rho)$ such that $\|\Sigma^{-1/2}f\|^2_{L_2(d\rho)}$. This relationship is even more transparent when considering a spectral decomposition of Σ.

Mercer decomposition. From extensions of Mercer's theorem (König, 1986), there exists an orthonormal *basis* $(e_m)_{m\geqslant 1}$ of $L_2(d\rho)$ and a summable non-increasing sequence of strictly positive eigenvalues $(\mu_m)_{m\geqslant 1}$ such that $\Sigma e_m = \mu_m e_m$. Note that since we have assumed that $\mathcal{F}$ is dense in $L_2(d\rho)$, there are no zero eigenvalues.

Since $\Sigma^{1/2}$ is an isometry from $L_2(d\rho)$ to $\mathcal{F}$, $(\mu_m^{1/2} e_m)_{m \geqslant 1}$ is an orthonormal basis of $\mathcal{F}$. Moreover, we can use the eigendecomposition to characterize elements of $\mathcal{F}$ as the functions in $L_2(d\rho)$ such that

$$\|\Sigma^{-1/2} f\|^2_{L_2(d\rho)} = \sum_{m \geqslant 1} \mu_m^{-1} \langle f, e_m \rangle^2_{L_2(d\rho)}$$

is finite. In other words, once projected in the orthonormal basis $(e_m)_{m \geqslant 1}$, elements f of $\mathcal{F}$ correspond to a certain decay of its decomposition coefficients $(\langle f, e_m \rangle_{L_2(d\rho)})_{m \geqslant 1}$.

Finally, by decomposing the function $k(\cdot, y) : x \mapsto k(x, y)$, we obtain the Mercer decomposition:

$$k(x, y) = \sum_{m \geqslant 1} \mu_m e_m(x) e_m(y).$$

Properties of the spectrum. The sequence of eigenvalues $(\mu_m)_{m \geqslant 1}$ is an important quantity that appears in the analysis of kernel methods (Hastie and Tibshirani, 1990; Caponnetto and De Vito, 2007; Harchaoui et al., 2008; Bach, 2013; El Alaoui and Mahoney, 2014). It depends both on the kernel k and the chosen distribution $d\rho$.

Some modifications of the kernel k or the distribution $d\rho$ lead to simple behaviors for the spectrum. For example, if we have a second distribution so that $\frac{d\rho'}{d\rho}$ is upper-bounded by a constant c, then, as a consequence of the Courant-Fischer minimax theorem (Horn and Johnson, 2012), the eigenvalues for $d\rho'$ are less than than c times that the ones for $d\rho$. Similarly, if the kernel k' is such that $ck - k'$ is a positive definite kernel, then we have a similar bound between eigenvalues.

In this paper, for any strictly positive λ, we will also consider the quantity $m^*(\lambda)$ equal to the number of eigenvalues μ_m that are greater than or equal to λ. Since we have assume that the sequence m is non-increasing, we have $m^*(\lambda) = \max\{m \geqslant 1,\ \mu_m \geqslant \lambda\}$. This is a left-continuous non-increasing function, that tends to $+\infty$ when λ tends to zero (since we have assumed that there are infinitely many strictly positive eigenvalues), and characterizes the sequence $(\mu_m)_{m \geqslant 1}$, as we can recover μ_m as $\mu_m = \sup\{\lambda \geqslant 0,\ m^*(\lambda) \geqslant m\}$.

Potential confusion with covariance operator. Note that the operator Σ is a self-adjoint operator on $L_2(d\rho)$. It should not be confused with the (non-centered) covariance operator C (Baker, 1973), which is a self-adjoint operator on a different space, namely the RKHS $\mathcal{F}$, defined by $\langle g, Cf \rangle_{\mathcal{F}} = \int_{\mathcal{X}} f(x) g(x) d\rho(x)$. Given that $\Sigma^{1/2}$ is an isometry from $L_2(d\rho)$ to $\mathcal{F}$, the operator C may also be used to define an operator on $L_2(d\rho)$, which happens to be exactly Σ. Thus, the two operators have the same eigenvalues. Moreover, we have, for any $y \in \mathcal{X}$:

$$(Cf)(y) = \langle k(\cdot, y), Cf \rangle_{\mathcal{F}} = \int_{\mathcal{X}} k(x, y) f(x) d\rho(x) = (\Sigma f)(y),$$

that is, C is equal to the restriction of Σ on $\mathcal{F}$.

2.2 Kernels as expectations

On top of the generic assumptions made above, we assume that there is another measurable set $\mathcal{V}$ equipped with a probability measure $d\tau$. We consider a function $\varphi : \mathcal{V} \times \mathcal{X} \to \mathbb{R}$ which is square-integrable (for the measure $d\tau \otimes d\rho$), and assume that the kernel k may be written as, for all $x, y \in \mathcal{X}$:

$$k(x,y) = \int_{\mathcal{V}} \varphi(v,x)\varphi(v,y)d\tau(v) = \langle \varphi(\cdot,x), \varphi(\cdot,y)\rangle_{L_2(d\tau)}. \tag{1}$$

In other words, the kernel between x and y is simply the expectation of $\varphi(v,x)\varphi(v,y)$ for v following the probability distribution $d\tau$. In this paper, we see $x \mapsto \varphi(v,x) \in \mathbb{R}$ as a one-dimensional random feature and $\varphi(v,x)\varphi(v,y)$ is the dot-product associated with this random feature. We could consider extensions where $\varphi(v,x)$ has values in a Hilbert space (and not simply $\mathbb{R}$), but this is outside the scope of this paper.

Square-root of integral operator. Such additional structure allows to give an explicit characterization of the RKHS $\mathcal{F}$ in terms of the features φ. In terms of operators, the function φ leads to a specific square-root of the integral operator Σ defined in Section 2.1 (which is not the positive self-adjoint square-root $\Sigma^{1/2}$).

We consider the bounded linear operator $T : L_2(d\tau) \to L_2(d\rho)$ defined as

$$(Tg)(x) = \int_{\mathcal{V}} g(v)\varphi(v,x)d\tau(v) = \langle g, \varphi(\cdot,x)\rangle_{L_2(d\tau)}. \tag{2}$$

Given $T : L_2(d\tau) \to L_2(d\rho)$, the adjoint operator $T^* : L_2(d\rho) \to L_2(d\tau)$ is the unique operator such that $\langle g, T^*f\rangle_{L_2(d\tau)} = \langle Tg, f\rangle_{L_2(d\rho)}$ for all f, g. Given the definition of T in Eq. (2), we simply inverse the role of $\mathcal{V}$ and $\mathcal{X}$ and have:

$$(T^*f)(v) = \int_{\mathcal{X}} f(x)\varphi(v,x)d\rho(x).$$

This implies by Fubini's theorem that

$$\begin{aligned}(TT^*f)(y) &= \int_{\mathcal{V}} \left(\int_{\mathcal{X}} f(x)\varphi(v,y)d\rho(x) \right) \varphi(v,x)d\tau(v) \\ &= \int_{\mathcal{X}} f(x) \left(\int_{\mathcal{V}} \varphi(v,y)\varphi(v,x)d\tau(v) \right) d\rho(x) = \int_{\mathcal{X}} f(x)k(x,y)d\rho(x) = (\Sigma f)(y),\end{aligned}$$

that is we have an expression of the integral operator Σ as $\Sigma = TT^*$. Thus, the decomposition of the kernel k as an expectation corresponds to a particular *square root* T of the integral operator—there are many possible choices for such square roots, and thus many possible expansions like Eq. (1). It turns out that the positive self-adjoint square root $\Sigma^{1/2}$ will correspond to the equivalence with quadrature rules (see Section 3.2).

Decomposition of functions in $\mathcal{F}$. Since $\Sigma = TT^*$ and $\Sigma^{1/2}$ is an isometry between $L_2(d\rho)$ and $\mathcal{F}$, we can naturally expressed any elements of $\mathcal{F}$ through the operator T and thus the features φ.

As a linear operator, T defines a bijection from the orthogonal of its null space $(\operatorname{Ker} T)^\perp \subset L_2(d\tau)$ to its image $\operatorname{Im}(T) \subset L_2(d\rho)$, and this allows to define uniquely $T^{-1}f \in (\operatorname{Ker} T)^\perp$ for any $f \in \operatorname{Im}(T)$, and a dot-product on $\operatorname{Im}(T)$ as

$$\langle f, h\rangle_{\operatorname{Im}(T)} = \langle T^{-1}f, T^{-1}g\rangle_{L_2(d\tau)}.$$

As shown by Bach (2014, App. A), $\mathrm{Im}(T)$ turns out to be equal to our RKHS[2]. Thus, the norm $\|f\|_{\mathcal{F}}^2$ for $f \in \mathcal{F}$ is equal to the squared L_2-norm of $T^{-1}f \in (\mathrm{Ker}\ T)^\perp$, which is itself equal to the minimum of $\|g\|_{L_2(d\tau)}^2$ over all g such that $Tg = f$. The resulting g may also be defined through pseudo-inverses.

In other words, a function $f \in L_2(d\rho)$ is in $\mathcal{F}$ if and only if it may be written as

$$\forall x \in \mathcal{X},\ f(x) = \int_{\mathcal{V}} g(v)\varphi(v,x)d\tau(v) = \langle g, \varphi(\cdot,x)\rangle_{L_2(d\tau)},$$

for a certain function $g : \mathcal{V} \to \mathbb{R}$ such that $\|g\|_{L_2(d\tau)}^2$ is finite, with a norm $\|f\|_{\mathcal{F}}^2$ equal to the minimum (which is always attained) of $\|g\|_{L_2(d\tau)}^2$, over all possible decompositions of f.

Singular value decomposition. The operator T is an Hilbert-Schmidt operator, to which the singular value decopomposition can be applied (Kato, 1995). That is, there exists an orthonormal basis $(f_m)_{m\geqslant 1}$ of $(\mathrm{Ker}\ T)^\perp \subset L_2(d\tau)$, together with the orthonormal basis $(e_m)_{m\geqslant 1}$ of $L_2(d\rho)$ which we have from the eigenvalue decomposition of $\Sigma = TT^*$, such that $Tf_m = \mu_m^{1/2}e_m$. Moreover, we have:

$$\varphi(v,x) = \sum_{m\geqslant 1} \mu_m^{1/2} e_m(x) f_m(v), \tag{3}$$

with a convergence in $L_2(d\tau \otimes d\rho)$. This extends the Mercer decomposition of the kernel $k(x,y)$.

Integral operator as an expectation. Given the expansion of the kernel k in Eq. (1), we may express the integral operator Σ as follows, explicitly as an expectation:

$$\begin{aligned}\Sigma f &= \int_{\mathcal{X}} f(y)k(\cdot,y)d\rho(y) = \int_{\mathcal{X}}\int_{\mathcal{V}} f(y)\varphi(v,\cdot)\varphi(v,y)d\rho(y)d\tau(v) \\ &= \int_{\mathcal{V}} \varphi(v,\cdot)\langle\varphi(v,\cdot), f\rangle_{L_2(d\rho)}d\tau(v) = \left(\int_{\mathcal{V}} \varphi(v,\cdot)\otimes_{L_2(d\rho)}\varphi(v,\cdot)d\tau(v)\right)f, \end{aligned}\tag{4}$$

where $a \otimes_{L_2(d\rho)} b$ is the operator $L_2(d\rho) \to L_2(d\rho)$ so that $(a\otimes_{L_2(d\rho)} b)f = \langle b, f\rangle_{L_2(d\rho)}a$. This will be useful to define empirical versions, where the integral over $d\tau$ will be replaced by a finite average.

2.3 Examples

In this section, we provide examples of kernels and usual decompositions. We first start by decompositions that always exist, then focus on specific kernels based on Fourier components.

Mercer decompositions. The Mercer decomposition provides an expansion for all kernels, as follows:

$$k(x,y) = \sum_{m\geqslant 1} \frac{\mu_m}{\mathrm{tr}\,\Sigma}\Big[\big(\mathrm{tr}\,\Sigma\big)^{1/2}e_m(x)\Big]\cdot\Big[\big(\mathrm{tr}\,\Sigma\big)^{1/2}e_m(x)\Big],$$

2. The proof goes as follows: (a) for any $y \in \mathcal{X}$, $k(\cdot,y)$ can be expressed as $\int_{\mathcal{V}}\varphi(v,y)\varphi(v,\cdot)d\tau(v) = T\varphi(\cdot,y)$ and thus belongs to $\mathrm{Im}(T)$; (b) for any $f \in \mathrm{Im}(T)$, and $y \in \mathcal{X}$, we have $\langle f, k(\cdot,y)\rangle_{\mathrm{Im}(T)} = \langle T^{-1}f, \varphi(\cdot,y)\rangle_{L_2(d\tau)} = (TT^{-1}f)(y) = f(y)$, that is, the reproducing property is satisfied. These two properties are characteristic of $\mathcal{F}$.

which can be transformed in to an expectation with $\mathcal{V} = \mathbb{N}^*$. In Section 3.2, we provide another generic decomposition with $\mathcal{V} = \mathcal{X}$. Note that this decomposition is typically impossible to compute (except for special cases below, i.e., special pairs of kernels k and distributions $d\rho$).

Periodic kernels on $[0,1]$. We consider $\mathcal{X} = [0,1]$ and translation-invariant kernels $k(x,y)$ of the form $k(x,y) = t(x-y)$, where t is a square-integrable 1-periodic function. These kernels are positive definite if and only if the Fourier series of t is non-negative (Wahba, 1990). An orthonormal basis of $L_2([0,1])$ is composed of the constant function $c_0 : x \mapsto 1$ and the functions $c_m : x \mapsto \sqrt{2}\cos 2\pi mx$ and $s_m : x \mapsto \sqrt{2}\sin 2\pi mx$. A kernel may thus be expressed as

$$k(x,y) = \nu_0 c_0(x) + \sum_{m>0} \nu_m \big[c_m(x)c_m(y) + s_m(x)s_m(y)\big] = \nu_0 + 2\sum_{m>0} \nu_m \cos 2\pi m(x-y).$$

This can be put trivially as an expectation with $\mathcal{V} = \mathbb{Z}$ and leads to the usual Fourier features (Rahimi and Recht, 2007). This is also exactly a Mercer decomposition for k and the uniform distribution on $[0,1]$, with eigenvalues ν_0 and ν_m, $m > 0$ (each of these with multiplicity 2). The associated RKHS norm for a function f is then equal to

$$\|f\|_{\mathcal{F}}^2 = \nu_0^{-1}\Big(\int_0^1 f(x)dx\Big)^2 + 2\sum_{m>0} \nu_m^{-1}\left[\Big(\int_0^1 f(x)\cos 2\pi mx dx\Big)^2 + \Big(\int_0^1 f(x)\sin 2\pi mx dx\Big)^2\right].$$

A particularly interesting example is obtained through derivatives of f. If f is differentiable and has a derivative $f' \in L_2([0,1])$, then, on the Fourier series coefficients of f, taking the derivative corresponds to multiplying the two m-th coefficients by $2\pi m$ and swapping them. Sobolev spaces for periodic functions on $[0,1]$ (i.e., such that $f(0) = f(1)$) are defined through integrability of derivatives (Adams and Fournier, 2003). In the Hilbert space set-up, a function f belongs to the Sobolev space of order s if one can define a s-th order square-integrable derivative in L_2 (for the Lebesgue measure, which happens to be equal to $d\rho$), that is, $f^{(s)} \in L_2([0,1])$. The Sobolev squared norm is then defined as any positive linear combination of the quadratic forms $\int_0^1 f^{(t)}(x)^2 dx$, $t \in \{0,\dots,s\}$, with non-zero coefficients for $t = 0$ and $t = s$ (all of these norms are then equivalent). If only using $t = 0$ and $t = s$ with non-zero coefficients, we need $\nu_0^{-1} = 1$ and $\nu_m^{-1} = 1 + m^{2s}$. An equivalent (i.e, with upper and lower bounded ratios) sequence is obtained by replacing $\nu_m = (1+m^{2s})^{-1}$ by $\nu_m = m^{-2s}$, leading to a closed-form formula:

$$k(x,y) = 1 + \frac{(-1)^{s-1}(2\pi)^{2s}}{(2s)!} B_{2s}(\{x-y\}),$$

where $\{x-y\}$ denotes the fractional part of $x-y$, and B_{2s} is the $2s$-th Bernoulli polynomial (Wahba, 1990). The RKHS $\mathcal{F}$ is then the Sobolev space of order s on $[0,1]$, with a norm equivalent to any of the family of Sobolev norms; it will be used as a running example throughout this paper.

Extensions to $[0,1]^d$. In order to extend to $d > 1$, we may consider several extensions as described by Oates and Girolami (2015), and compute the resulting eigenvalues of the integral operators. For simplicity, we consider the Sobolev space on $[0,1]$, with $\nu_0 = 1$ and $\nu_m^{-1} = m^{2s}$ for $m > 0$. The first possibility to extend to $[0,1]^d$ is to take a kernel which is simply the pointwise product of individual kernels on $[0,1]$. That is, if $k(x,y)$ is the kernel on $[0,1]$, define $K(X,Y) = \prod_{j=1}^d k(x_j,y_j)$ between X and Y in $[0,1]^d$. As shown in Appendix A, this leads to eigenvalue decays bounded by

$(\log m)^{2s(d-1)} m^{-2s}$, and thus up to logarithmic terms at the same speed m^{-2s} as $d = 1$. While this sounds attractive in terms of generalization performance, it corresponds to a space a function which is not a Sobolev space in d dimensions. That is the associated squared norm on f would be equivalent to a linear combination of squared L_2-norm of partial derivatives

$$\int_{[0,1]^d} \Big(\frac{\partial^{t_1+\cdots+t_d} f}{\partial x_1^{t_1} \cdots \partial x_d^{t_d}} \Big)^2 dx$$

for all $t_1, \dots, t_d$ in $\{0, \dots, s\}$. This corresponds to functions which have square-integrable partial derivatives with all *individual* orders less than s. All values of $s \geqslant 1$ are allowed and lead to an RKHS.

This is thus to be contrasted with the usual multi-dimensional Sobolev space which is composed of functions which have square-integrable partial derivatives with all orders $(t_1, \dots, t_d)$ with *sum* $t_1 + \cdots + t_d$ less than s. Only $s > d/2$ is then allowed to get an RKHS. The Sobolev norm is then of the form

$$\sum_{t_1+\cdots+t_d \leqslant s} \int_{[0,1]^d} \Big(\frac{\partial^{t_1+\cdots+t_d} f}{\partial x_1^{t_1} \cdots \partial x_d^{t_d}} \Big)^2 dx.$$

In the expansion on the d-th order tensor product of the Fourier basis, the norm above is equivalent to putting a weight on the element $(m_1, \dots, m_d)$ asymptotically equivalent to $\big(\sum_{j=1}^d m_j \big)^{2s}$, which thus represent the inverse of the eigenvalues of the corresponding kernel for the uniform distribution $d\rho$ (this is simply an explicit Mercer decomposition). Thus, the number of eigenvalues which are greater than λ grows as the number of $(m_1, \dots, m_d)$ such that their sum is less than $\lambda^{-1/(2s)}$, which itself is less than a constant times $\lambda^{-d/(2s)}$ (see a proof in Appendix A). This leads to an eigenvalue decay of $m^{-2s/d}$, which is much worse because of the term in $1/d$ in the exponent.

Translation invariant kernels on $\mathbb{R}^d$. We consider $\mathcal{X} = \mathbb{R}^d$ and translation-invariant kernels $k(x, y)$ of the form $k(x, y) = t(x - y)$, where t is an integrable function from $\mathbb{R}^d$ to $\mathbb{R}$. It is known that these kernels are positive definite if and only if the Fourier transform of t is always a non-negative real number. More precisely, if $\hat{t}(\omega) = \int_{\mathbb{R}^d} t(x) e^{-i\omega^\top x} dx \in \mathbb{R}_+$, then

$$k(x, y) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{t}(\omega) e^{i\omega^\top (x-y)} d\omega = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{t}(\omega) \big[\cos \omega^\top x \cos \omega^\top y + \sin \omega^\top x \sin \omega^\top y \big] d\omega.$$

Following Rahimi and Recht (2007), by sampling ω from a density proportional to $\hat{t}(\omega) \in \mathbb{R}_+$ and b uniformly in $[0, 1]$ (and independently of ω), then by defining $\mathcal{V} = \mathbb{R}^d \times [0, 1]$ and $\varphi(\omega, b, x) = \sqrt{2} \cos(\omega^\top x + 2\pi b)$, we obtain the kernel k.

For these kernels, the decay of eigenvalues has been well-studied by Widom (1963), who relates the decay of eigenvalues to the tails of the distribution $d\rho$ and the decay of the Fourier transform of t. For example, for the Gaussian kernel where $k(x, y) = \exp(-\alpha \|x - y\|_2^2)$, on sub-Gaussian distributions, the decay of eigenvalues is geometric, and for kernels leading to Sobolev spaces of order s, such as the Matern kernel (Furrer and Nychka, 2007), the decay is of the form $m^{-2s/d}$. See also examples by Birman and Solomyak (1977); Harchaoui et al. (2008).

Finally, note that in terms of computation, there are extensions to avoid linear complexity in d (Le et al., 2013).

Kernels on hyperspheres. If $\mathcal{X} \subset \mathbb{R}^{d+1}$ is the d-dimensional hypersphere $\{x \in \mathbb{R}^{d+1}, \|x\|_2^2 = 1\}$, then specific kernels may be used, of the form $k(x,y) = t(x^\top y)$, where t has to have a positive Legendre expansion (Smola et al., 2001). Alternatively, kernels based on neural networks with random weights are directly in the form of random features (Cho and Saul, 2009; Bach, 2014): for example, the kernel $k(x,y) = \mathbb{E}(v^\top x)_+^s (v^\top x)_+^s$ for v uniformly distributed in the hypersphere corresponds to sampling weights in a one-hidden layer neural network with rectified linear units (Cho and Saul, 2009). It turns out that these kernels have a known decay for their spectrum.

As shown by Smola et al. (2001); Bach (2014), the equivalent of Fourier series (which corresponds to $d = 1$) is then the basis of spherical harmonics, which is organized by integer frequencies $k \geqslant 1$; instead of having 2 basis vectors (sine and cosine) per frequency, there are $O(k^{d-1})$ of them. As shown by Bach (2014, page 44), we have an explicit expansion of $k(x,y)$ in terms of spherical harmonics, leading to a sequence of eigenvalues equal to $k^{-d-2s-1}$ on the entire subspace associated with frequency k. Thus, by taking multiplicity into account, after $\sum_{j=1}^k j^{d-1} \approx k^d$ (up to constants) eigenvalues, we have an eigenvalue of $k^{-d-2s-1}$; this leads to an eigenvalue decay (where all eigenvalues are ordered in decreasing order and we consider the m-th one) as $(m^{1/d})^{-d-2s-1} = m^{-1-1/d-2s/d}$.

2.4 Approximation from randomly sampled features

Given the formulation of k as an expectation in Eq. (1), it is natural to consider sampling n elements $v_1, \dots, v_n \in \mathcal{V}$ from the distribution $d\tau$ and define the kernel approximation

$$\hat{k}(x,y) = \frac{1}{n} \sum_{i=1}^n \varphi(v_i, x) \varphi(v_i, y), \tag{5}$$

which defines a finite-dimensional RKHS $\hat{\mathcal{F}}$.

From the strong law of large numbers—which can be applied because we have the finite expectation $\mathbb{E}|\varphi(v,x)\varphi(v,y)| \leqslant \big(\mathbb{E}|\varphi(v,x)|^2 \mathbb{E}|\varphi(v,y)|^2\big)^{1/2}$, when n tends to infinity, $\hat{k}(x,y)$ tends to $k(x,y)$ almost surely, and thus we get as tight as desired approximations of the kernel k, for a given pair $(x,y) \in \mathcal{X} \times \mathcal{X}$. Rahimi and Recht (2007) show that for translation-invariant kernels on a Euclidean space, then the convergence is uniform over a compact subset of $\mathcal{X}$, with the traditional rate of convergence of $\sqrt{\frac{\log n}{n}}$.

In this paper, we rather consider *approximations of functions* in $\mathcal{F}$ by functions in $\hat{\mathcal{F}}$, the RKHS associated with $\hat{k}$. A key difficulty is that in general $\hat{\mathcal{F}}$ is not even included in $\mathcal{F}$, and therefore, we cannot use the norm in $\mathcal{F}$ to characterize approximations. In this paper, we choose the L_2-norm associated with the probability measure $d\rho$ on $\mathcal{X}$ to characterize the approximation. Given $f \in \mathcal{F}$ with norm $\|f\|_{\mathcal{F}}$ less than one, we look for a function $\hat{f} \in \hat{\mathcal{F}}$ of the smallest possible norm and so that $\|f - \hat{f}\|_{L_2(d\rho)}$ is as small as possible.

Note that the measure $d\tau$ is associated to the kernel k and the random features φ, while the measure $d\rho$ is associated to the way we want to measure errors (and leads to a specific defintion of the integral operator Σ).

Computation of error. Given the definition of the Hilbert space $\mathcal{F}$ in terms of φ in Section 2.2, given $g \in L_2(d\tau)$ with $\|g\|_{L_2(d\tau)} \leqslant 1$ and $f(x) = \int_{\mathcal{V}} g(v)\varphi(v,x)d\tau(v)$, we aim at finding an element of $\hat{\mathcal{F}}$ close to f. We can also represent $\hat{\mathcal{F}}$ through a similar decomposition, now with a finite number of features, i.e., through $\alpha \in \mathbb{R}^n$ such that $\hat{f} = \sum_{i=1}^n \alpha_i \varphi(v_i, \cdot)$ with norm[3] $\|\hat{f}\|_{\hat{\mathcal{F}}}^2 \leqslant n\|\alpha\|_2^2$ as small as possible and so that the following approximation error is also small:

$$\|\hat{f} - f\|_{L_2(d\rho)} = \left\| \sum_{i=1}^n \alpha_i \varphi(v_i, \cdot) - \int_{\mathcal{V}} g(v)\varphi(v,\cdot)d\tau(v) \right\|_{L_2(d\rho)}. \tag{6}$$

Note that with $\alpha_i = \frac{1}{n} g(v_i)$ and v_i sampled from $d\tau$ (independently), then, we have $\mathbb{E}(\|\alpha\|_2^2) = \sum_{i=1}^n \mathbb{E}\alpha_i^2 = \frac{1}{n}\mathbb{E}g(v)^2 \leqslant \frac{1}{n}$ and an expected error $\mathbb{E}(\|f - \hat{f}\|_{L_2(d\rho)}^2) = \frac{1}{n}\mathbb{E}\|g(v)\varphi(v,\cdot)\|_{L_2(d\rho)}^2 \leqslant \frac{1}{n}\sup_{v\in\mathcal{V}} \|\varphi(v,\cdot)\|_{L_2(d\rho)}^2$; our goal is to obtain an error rate with a better scaling in n, by (a) choosing a better distribution than $d\tau$ for the points $v_1, \dots, v_n$ and (b) by finding the best possible weights $\alpha \in \mathbb{R}^n$ (that should of course depend on the function g).

Goals. We thus aim at sampling n points $v_1, \dots, v_n \in \mathcal{V}$ from a distribution with density q with respect to $d\tau$. Then the kernel approximation using *importance weights* is equal to

$$\hat{k}(x,y) = \frac{1}{n}\sum_{i=1}^n \frac{1}{q(v_i)}\varphi(v_i,x)\varphi(v_i,y)$$

(so that the law of large numbers leads to an approximation converging to k), and we thus aim at minimizing $\left\| \sum_{i=1}^n \frac{\beta_i}{q(v_i)^{1/2}}\varphi(v_i,\cdot) - \int_{\mathcal{V}} g(v)\varphi(v,\cdot)d\tau(v) \right\|_{L_2(d\rho)}$, with $n\|\beta\|_2^2$ (which represents the norm of the approximation in $\hat{\mathcal{F}}$ because of our importance weights are taken into account) as small as possible.

3. Quadrature in RKHSs

Given a square-integrable (with respect to $d\rho$) function $g : \mathcal{X} \to \mathbb{R}$, the quadrature problem aims at approximating, for all $h \in \mathcal{F}$, integrals

$$\int_{\mathcal{X}} h(x)g(x)d\rho(x)$$

by linear combinations

$$\sum_{i=1}^n \alpha_i h(x_i)$$

of evaluations $h(x_1), \dots, h(x_n)$ of the function h at well-chosen points $x_1, \dots, x_n \in \mathcal{X}$. Of course, coefficients $\alpha \in \mathbb{R}^n$ are allowed to depend on g (they will in linear fashion in the next section), but not on h, as the so-called quadrature rule has to be applied to all functions in $\mathcal{F}$.

3. Note the factor n because our finite-dimensional kernel in Eq. (5) is an average of kernels and not a sum.

3.1 Approximation of the mean element

Following Smola et al. (2007), the error may be expressed using the reproducing property as:

$$\sum_{i=1}^{n} \alpha_i h(x_i) - \int_{\mathcal{X}} h(x) g(x) d\rho(x) = \left\langle h, \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) - \int_{\mathcal{X}} k(\cdot, x) g(x) d\rho(x) \right\rangle_{\mathcal{F}},$$

and by Cauchy-Schwarz inequality its supremum over $\|h\|_{\mathcal{F}} \leqslant 1$ is equal to

$$\left\| \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) - \int_{\mathcal{X}} k(\cdot, x) g(x) d\rho(x) \right\|_{\mathcal{F}}. \tag{7}$$

The goal of quadrature rules formulated in a RKHS is thus to find points $x_1, \dots, x_n \in \mathcal{X}$ and weights $\alpha \in \mathbb{R}^n$ so that the quantity in Eq. (7) is as small as possible (Smola et al., 2007). For $g = 1$, the function $\int_{\mathcal{X}} k(\cdot, x) d\rho(x)$ is usually referred to as the mean element of the distribution $d\rho$.

The standard Monte-Carlo solution is to consider $x_1, \dots, x_n$ sampled i.i.d. from $d\rho$ and the weights $\alpha_i = g(x_i)/n$, which leads to a decrease of the error in $1/\sqrt{n}$, with $\mathbb{E}\|\alpha\|_2^2 \leqslant \frac{1}{n}$ and an expected squared error which is equal to $\frac{1}{n}\mathbb{E}\|g(v)k(:, x)\|_{\mathcal{F}}^2 \leqslant \frac{1}{n}\|g\|_{L_2(d\rho)}^2 \sup_{x \in \mathcal{X}} k(x, x)$ (Smola et al., 2007). Note that when $g = 1$, Eq. (7) corresponds to a particular metric between the distribution $d\rho$ and its corresponding empirical distribution (Sriperumbudur et al., 2010).

In this paper, we explore sampling points x_i from a probability distribution on $\mathcal{X}$ with density q with respect to $d\rho$. Note that when g is a constant function, it is sometimes required that the coefficients α are non-negative and sum to a fixed constant (so that constant functions are exactly integrated). We will not pursue this here as our theoretical results do not accommodate such constraints (see, e.g., Chen et al., 2010; Bach et al., 2012, and references therein).

Tolerance to noisy function values. In practice, independent (but not necessarily identically distributed) noise ε_i may be present with variance $\sigma^2(x_i)$. Then, the worst (with respect to $\|h\|_{\mathcal{F}} \leqslant 1$) expected (with respect to the noise) squared error is

$$\begin{aligned} &\inf_{\|h\|_{\mathcal{F}} \leqslant 1} \mathbb{E} \left| \sum_{i=1}^{n} \alpha_i (h(x_i) + \varepsilon_i) - \int_{\mathcal{X}} h(x) g(x) d\rho(x) \right|^2 \\ = \; &\left\| \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) - \int_{\mathcal{X}} k(\cdot, x) g(x) d\rho(x) \right\|_{\mathcal{F}}^2 + \sum_{i=1}^{n} \alpha_i^2 \sigma^2(x_i), \end{aligned}$$

and thus in order to be robust to noise, having a small weighted ℓ_2-norm for the coefficients $\alpha \in \mathbb{R}^n$ is important.

3.2 Reformulation as random features

For any $x \in \mathcal{X}$, the function $k(\cdot, x)$ is in $\mathcal{F}$, and since we have assumed that $\Sigma^{1/2}$ is an isometry from $L_2(d\rho)$ to $\mathcal{F}$, there exists a unique element, which we denote $\psi(\cdot, x)$, of $L_2(d\rho)$ such that $\Sigma^{1/2}\psi(\cdot, x) = k(\cdot, x)$. Given the Mercer decomposition $k(\cdot, x) = \sum_{m \geqslant 1} \mu_m e_m(x) e_m$, we have

the expansion $\psi(\cdot, x) = \sum_{m \geqslant 1} \mu_m^{1/2} e_m(x) e_m$ (with convergence in the L_2-norm for the measure $d\rho \otimes d\rho$; note that we do not assume that $\mu_m^{1/2}$ is summable), and thus we may consider ψ as a symmetric function. Note that ψ may not be easy to compute in many practical cases (except for some periodic kernels on $[0,1]$).

We thus have for $(x, y) \in \mathcal{X} \times \mathcal{X}$:

$$\begin{aligned} k(x,y) &= \langle k(\cdot,x), k(\cdot,y)\rangle_{\mathcal{F}} = \langle \Sigma^{1/2}\psi(\cdot,x), \Sigma^{1/2}\psi(\cdot,y)\rangle_{\mathcal{F}} = \langle \psi(\cdot,x), \psi(\cdot,y)\rangle_{L_2(d\rho)} \\ &\qquad \text{because of the isometry property of } \Sigma^{1/2}, \\ &= \int_{\mathcal{X}} \psi(v,x)\psi(v,y) d\rho(v). \end{aligned} \tag{8}$$

That is, the kernel k may always be written as an expectation. Moreover, we have the quadrature error in Eq. (7) equal to (again using the isometry $\Sigma^{1/2}$ from $L_2(d\rho)$ to $\mathcal{F}$):

$$\begin{aligned} \left\| \sum_{i=1}^n \alpha_i k(\cdot, x_i) - \int_{\mathcal{X}} k(\cdot,x) g(x) d\rho(x) \right\|_{\mathcal{F}} &= \left\| \sum_{i=1}^n \alpha_i \Sigma^{1/2}\psi(x_i,\cdot) - \int_{\mathcal{X}} \Sigma^{1/2}\psi(x,\cdot) g(x) d\rho(x) \right\|_{\mathcal{F}} \\ &= \left\| \sum_{i=1}^n \alpha_i \psi(x_i,\cdot) - \int_{\mathcal{X}} \psi(x,\cdot) g(x) d\rho(x) \right\|_{L_2(d\rho)}, \end{aligned}$$

which is exactly an instance of the approximation result in Eq. (6) with $\mathcal{V} = \mathcal{X}$ and $\varphi = \psi$, that is the random feature is indexed by the same set $\mathcal{X}$ as the kernel. Thus, the quadrature problem, that is finding points x_i and weights (α_i) to get the best possible error over all functions of the unit ball of $\mathcal{F}$, is a *subcase* of the random feature problem for a specific expansion. Note that this random decomposition in terms of ψ is always possible (although not in closed form in general).

Interpretation through square-roots of intergral operators. As shown in Section 2.2, random feature expansions correspond to square-roots of the integral operator $\Sigma : L_2(d\rho) \to L_2(d\rho)$ as $\Sigma = TT^*$. Among the many possible square roots, the quadrature case corresponds exactly to the positive self-adjoint square root $T = \Sigma^{1/2}$. In this situation, the basis $(f_m)_{m \geqslant 1}$ of the singular value decomposition of $T = \Sigma^{1/2}$ is equal to $(e_m)_{m \geqslant 1}$, recovering the expansion $\psi(x,y) = \sum_{m \geqslant 1} \mu_m^{1/2} e_m(x) e_m(y)$ which we have seen above.

Translation-invariant kernels on $[0,1]^d$ or $\mathcal{X} = \mathbb{R}^d$. In this important situation, we have two different expansions: the one based on Fourier features, where the random variable indexing the one-dimensional feature is a *frequency*, while for the one based on the square root ψ, the random variable is a *spatial variable* in $\mathcal{X}$. As we show in Section 4, our results are independent of the chosen expansions and thus apply to both. However, (a) when the goal is to do quadrature, we need to use ψ, and (b) in general, the decomposition based on Fourier features can be easily computed once samples are obtained, while for most kernels, $\psi(x,y)$ does not have any closed-form simple expression. In Section 6, we provide a simple example with $\mathcal{X} = [0,1]$ where the two decompositions are considered.

Goals. In order to be able to make the parallel with random feature approximations, we consider importance-weighted coefficients $\beta_i = \alpha_i q(x_i)^{1/2}$, and we thus aim at minimizing the approxima-

tion error

$$\Big\| \sum_{i=1}^{n} \beta_i q(x_i)^{-1/2} k(\cdot, x_i) - \int_{\mathcal{X}} k(\cdot, x) g(x) d\rho(x) \Big\|_{\mathcal{F}}.$$

We consider potential independent noise with variance $\sigma^2(x_i) \leqslant \tau^2 q(x_i)$ for all x_i, so that the tolerance to noise is characterized by the ℓ_2-norm $\|\beta\|_2$.

3.3 Relationship with column sampling

The problem of quadrature is related to the problem of column sampling. Given n observations $x_1, \dots, x_n \in \mathcal{X}$, the goal of column-sampling methods is to approximate the $n \times n$ matrix of pairwise kernel evalulations, the so-called *kernel matrix*, from a subset of its columns. It has appeared under many names: Nyström method (Williams and Seeger, 2001), sparse greedy approximations (Smola and Schölkopf, 2000), incomplete Cholesky decomposition (Fine and Scheinberg, 2001), Gram-Schmidt orthonormalization (Shawe-Taylor and Cristianini, 2004) or CUR matrix decompositions (Mahoney and Drineas, 2009).

While column sampling has typically been analyzed for a fixed kernel matrix, it has a natural extension which is related to quadrature problems: selecting n points $x_1, \dots, x_n$ from $\mathcal{X}$ such that the projection of any element of the RKHS $\mathcal{F}$ onto the subspace spanned by $k(\cdot, x_i)$, $i = 1, \dots, n$ is as small as possible. Natural functions from $\mathcal{F}$ are $k(\cdot, x)$, $x \in \mathcal{X}$, and thus the goal is to minimize, for such $x \in \mathcal{X}$,

$$\inf_{\alpha \in \mathbb{R}^n} \Big\| \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) - k(\cdot, x) \Big\|_{\mathcal{F}}^2$$

In the usual sampling approach, several points are considered for testing the projection error, and it is thus natural to consider the criterion averaged through the measure $d\rho$, that is:

$$\int_{\mathcal{X}} \inf_{\alpha \in \mathbb{R}^n} \Big\| \sum_{i=1}^{n} \alpha_i k(\cdot, x_i) - k(\cdot, x) \Big\|_{\mathcal{F}}^2 d\rho(x).$$

In fact, when $d\rho$ is supported on a finite set, this formulation is equivalent to minimizing the nuclear norm between the kernel matrix and its low-rank approximation. There are thus several differences and similarities between recent work on column sampling (Bach, 2013; El Alaoui and Mahoney, 2014) and the present paper on quadrature rules and random features:

- **Different error measures**: The column sampling approach corresponds to a function g in Eq. (7) which is a Dirac function at the point x, and is thus not in $L_2(d\rho)$. Thus the two frameworks are not equivalent.
- **Approximation vs. prediction**: The works by Bach (2013); El Alaoui and Mahoney (2014) aim at understanding when column sampling leads to no loss in predictive performance within a supervised learning framework, while the present paper looks at approximation properties, mostly regardless of any supervised learning problem, except in Section 4.5 for random features (but not for quadrature).
- **Lower bounds**: In Section 4.3, we provide explicit lower bounds of approximations, which are not available for column sampling.

– **Similar sampling issues**: In the two frameworks, points $x_1, \dots, x_n \in \mathcal{X}$ are sampled i.i.d. with a certain distribution q, and the best choice depends on the appropriate notion of leverage scores (Mahoney, 2011), while the standard uniform distribution leads to an inferior approximation result. Moreover, the proof techniques are similar and based on concentration inequalities for operators, here in Hilbert spaces rather in finite dimensions.

3.4 Related work on quadrature

Many methods have been designed for the computation of integrals of a function given evaluations at certain well-chosen points, in most cases when g is constant equal to one. We review some of these below.

Uni-dimensional integrals. When the underlying set $\mathcal{X}$ is a compact interval of the real line, several methods exists, such as the trapezoidal or Simpson's rules, which are based on interpolation between the sample points, and for which the error decays as $O(1/n^2)$ and $O(1/n^4)$ for functions with uniformly bounded second or fourth derivatives (Cruz-Uribe and Neugebauer, 2002).

Gaussian quadrature is another class of methods for one-dimensional integrals: it is based on a basis of orthogonal polynomials for $L_2(d\rho)$ where $d\rho$ is a probability measure supported in an interval, and their zeros (Hildebrand, 1987, Chap. 8). This leads to quadrature rules which are exact for polynomials of degree $2n-1$ but error bounds for non-polynomials rely on high-order derivatives, although the empirical performance on functions of a Sobolev space in our experiments is as good as optimal quadrature schemes (see Section 6); depending on the orthogonal polynomials, we get various quadrature rules, such as Gauss-Legendre quadrature for the Lebesgue measure on $[0, 1]$.

Quasi Monte-carlo methods employ a sequence of points with low discrepancy with uniform weights (Morokoff and Caflisch, 1994), leading to approximation errors of $O(1/n)$ for univariate functions with bounded variation, but typically with no adaptation to smoother functions.

Higher-dimensional integrals. All of the methods above may be generalized for products of intervals $[0, 1]^d$, typically with d small. For larger problems, Bayes-Hermite quadrature (O'Hagan, 1991) is essentially equivalent to the quadrature rules we study in this paper.

Some of the quadrature rules are constrained to have positive weights with unit sum (so that the positivity properties of integrals are preserved and constants are exactly inegrated). The quadrature rules we present do not satisfy these constraints. If these constraints are required, kernel herding (Chen et al., 2010; Bach et al., 2012) provides a novel way to select a sequence of points based on the conditional gradient algorithm, but with currently no convergence guarantees improving over $O(1/\sqrt{n})$ for infinite-dimensional spaces.

Theoretical results. The best possible error for a quadrature rule with n points has been well-studied in several settings; see Novak (1988) for a comprehensive review. For example, for $\mathcal{X} = [0, 1]$ and the space of Sobolev functions, which are RKHSs with eigenvalues of their integral operator decreasing as m^{-2s}, Novak (1988, Prop. 2 and 3, page 38) shows that the best possible quadrature rule for the uniform distribution and $g = 1$ leads to an error rate of n^{-s}, as well as for any squared-integrable function g. The proof of these results (both upper and lower bounds) relies

on detailed properties of Sobolev spaces. In this paper, we recover these results using only the decay of eigenvalues of the associated integral operator Σ, thus allowing straightforward extensions to many situations, like Sobolev spaces on manifolds such as hyperspheres (Hesse, 2006), where we also recover existing results (up to logarithmic terms).

Moreover, Novak (1988, page 17) shows that adaptive quadrature rules where points are selected sequentially with the knowledge of the function values at previous points cannot improve the worst-case guarantees. Our results do not recover this lower bound result for adaptivity.

Finally, Langberg and Schulman (2010) consider multiplicative errors in computing integrals and mainly focuses on different function spaces, such as ones used in clustering functionals. Although sampling quadrature points from a well-chosen density is common in the two approaches, the analysis tools are different. It would be interesting to see if some of these tools can be transferred to our RKHS setting.

From quadrature to function approximation and optimization. The problem of quadrature, uniformly over all functions $g \in L_2(d\rho)$ that define the integral, is in fact equivalent to the full approximation of a function h given values at n points, where the approximation error is characterized in L_2-norm. Indeed, given the observations $h(x_i)$, $i = 1, \dots, n$, we build $\sum_{i=1}^n \alpha_i h(x_i)$ as an approximation of $\int_{\mathcal{X}} g(x)h(x)d\rho(x)$. It turns out that the coefficients α_i are linear in g, that is, there exists $a_i \in L_2(d\rho)$ such that $\alpha_i = \langle a_i, g\rangle_{L_2(d\rho)}$. This implies that $\sum_{i=1}^n h(x_i)\langle a_i, g\rangle_{L_2(d\rho)}$ is an approximation of $\langle h, g\rangle_{L_2(d\rho)}$. Thus, the worst case error with respect to g in the unit ball of $L_2(d\rho)$ is $\big\| \sum_{i=1}^n h(x_i)a_i - h\big\|_{L_2(d\rho)}$, that is, we have an approximation result of h through observations of its values at certain points.

Novak (1988) considers the approximation problem in L_∞-norm and shows that for Sobolev spaces, going from L_2- to L_∞-norms incurs a loss of performance of $\sqrt{n}$. We recover partially these results in Section 5 from a more general perspective. When optimizing the points at which the function is evaluated (adaptively or not), the approximation problem is often referred to as experimental design (Cochran and Cox, 1957; Chaloner and Verdinelli, 1995).

Finally, a third problem is of interest (and outside of the scope of this paper), namely the problem of finding the minimum of a function given (potentially noisy) function evaluations. For noiseless problems, Novak (1988, page 26) shows that the approximation and optimization problems have the same worst-case guarantees (with no influence of adaptivity); this optimization problem has also been studied in the bandit setting (Srinivas et al., 2012) and in the framework of "Bayesian optimization" (see, e.g. Bull, 2011).

4. Theoretical Analysis

In this section, we provide approximation bounds for the random feature problem outlined in Section 2.4 (and thus the quadrature problem in Section 3). In Section 4.1, we provide generic upper bounds, which depend on the eigenvalues of the integral operator Σ and present matching lower bounds (up to logarithmic terms) in Section 4.3. The upper-bound depends on specific distributions of samples that we discuss in Section 4.2. We then consider consequences of these results on quadrature (Section 4.4) and random feature expansions (Section 4.5).

4.1 Upper bound

The following proposition (see proof in Appendix B.1) determines the minimal number of samples required for a given approximation accuracy:

Proposition 1 (Approximation of the unit ball of $\mathcal{F}$) *For $\lambda > 0$ and a distribution with positive density q with respect to $d\tau$, we consider*

$$d_{\max}(q,\lambda) = \sup_{v\in\mathcal{V}} \frac{1}{q(v)} \langle \varphi(v,\cdot), (\Sigma+\lambda I)^{-1}\varphi(v,\cdot)\rangle_{L_2(d\rho)}. \tag{9}$$

Let $v_1,\dots,v_n$ be sampled i.i.d. from the density q, then for any $\delta\in(0,1)$, if

$$n \geqslant 5 d_{\max}(q,\lambda)\log\frac{16 d_{\max}(q,\lambda)}{\delta},$$

with probability greater than $1-\delta$, we have $\frac{1}{n}\sum_{i=1}^n q(v_i)^{-1}\|\varphi(v_i,\cdot)\|^2_{L_2(d\rho)} \leqslant \frac{2\operatorname{tr}\Sigma}{\delta}$ and

$$\sup_{\|f\|_{\mathcal{F}}\leqslant 1} \inf_{\|\beta\|_2^2\leqslant \frac{4}{n}} \Big\| f - \sum_{i=1}^n \beta_i q(v_i)^{-1/2}\varphi(v_i,\cdot)\Big\|^2_{L_2(d\rho)} \leqslant 4\lambda.$$

We can interpret the proposition above as follows: given any squared error $4\lambda>0$ and a distribution with density q, the number n of samples from q needed so that the unit ball of $\mathcal{F}$ is approximated by the ball of radius 2 of $\hat{\mathcal{F}}$ is, up to logarithmic terms, at most a constant times $d_{\max}(q,\lambda)$, defined in Eq. (9). The result above is a statement for a fixed q and λ and this number of samples n depends on these.

We could also invert the relationship between λ and n, that is, answer the following question: given a fixed number n of samples, what is the approximation error λ? This requires inverting the function $\lambda\mapsto d_{\max}(q,\lambda)$. This will be done in Section 4.2 for a specific distribution q where the expression simplifies, together with specific examples from Section 2.3.

Finally, note that we also have a bound on $\frac{1}{n}\sum_{i=1}^n q(v_i)^{-1}\|\varphi(v_i,\cdot)\|^2_{L_2(d\rho)}$, which shows that our random functions are not too large on average (this constraint will be needed in the lower bound as well in Section 4.3).

Sketch of proof. The proof technique relies on computing an explicit candidate $\beta\in\mathbb{R}^n$ obtained from minimizing a regularized least-squares formulation

$$\inf_{\beta\in\mathbb{R}^n} \Big\| \sum_{i=1}^n \beta_i q(v_i)^{-1/2}\varphi(v_i,\cdot) - f\Big\|^2_{L_2(d\rho)} + n\lambda\|\beta\|_2^2.$$

It turns out that the final bound on the squared error is exactly proportional to the regularization parameter λ. As shown in Appendix B.1, this leads to an approximation $\hat{f}$ which is a linear function of f, as $\hat{f} = (\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma} f$, where $\hat{\Sigma}$ is a properly defined empirical integral operator and $\lambda>0$ is the regularization parameter. Then, Bernstein concentration inequalities for operators (Minsker, 2011) can be used in a way similar to the work of Bach (2013); El Alaoui and Mahoney (2014) on column sampling, to provide a bound on all desired quantities.

Result in expectation. In Section 4.5, we will need a result in expectation. As shown at the end of Appendix B.1, as soons as, $\lambda \leqslant (\operatorname{tr}\Sigma)/4$ and $n \geqslant 5 d_{\max}(\lambda) \log \dfrac{2(\operatorname{tr}\Sigma) d_{\max}(\lambda)}{\lambda}$, then

$$\mathbb{E}\left(\sup_{\|f\|_{\mathcal{F}} \leqslant 1} \inf_{\|\beta\|_2^2 \leqslant \frac{4}{n}} \left\| f - \sum_{i=1}^n \beta_i q(v_i)^{-1/2} \varphi(v_i, \cdot) \right\|_{L_2(d\rho)}^2 \right) \leqslant 8\lambda.$$

4.2 Optimized distribution

We may now consider a specific distribution that depends on the kernel and on λ, namely

$$q_\lambda^*(v) = \frac{\langle \varphi(v,\cdot), (\Sigma + \lambda I)^{-1} \varphi(v,\cdot) \rangle_{L_2(d\rho)}}{\int_{\mathcal{V}} \langle \varphi(v,\cdot), (\Sigma + \lambda I)^{-1} \varphi(v,\cdot) \rangle_{L_2(d\rho)} d\tau(v)} = \frac{\langle \varphi(v,\cdot), (\Sigma + \lambda I)^{-1} \varphi(v,\cdot) \rangle_{L_2(d\rho)}}{\operatorname{tr}\Sigma(\Sigma + \lambda I)^{-1}}, \quad (10)$$

for which $d_{\max}(q_\lambda^*, \lambda) = d(\lambda) = \operatorname{tr}\Sigma(\Sigma + \lambda I)^{-1}$. With this distribution, we thus need to have $n \geqslant 5 d(\lambda) \log \frac{16 d(\lambda)}{\delta}$ with $d(\lambda) = \operatorname{tr}\Sigma(\Sigma + \lambda I)^{-1}$ is the *degrees of freedom*, a traditional quantity in the analysis of least-squares regression (Hastie and Tibshirani, 1990; Caponnetto and De Vito, 2007), which is always smaller than $d_{\max}(1, \lambda)$ and can be upper-bounded explicitly for many examples, as we now explain. The computation of $d_{\max}(1, \lambda)$ in the operator setting (for which we may use $q = 1$), a quantity often referred to as the maximal *leverage score* (Mahoney, 2011), remains an open problem.

The quantity $d(\lambda)$ only depends on the integral operator Σ, that is, for all possible choices of square roots, i.e., all possible choices of feature expansions, the number of samples that our results guarantee is the same. This being said, some expansions may be more computationally practical than others, and when using the distribution with $q(v) = 1$, the bounds will be different.

Expression in terms of singular value decomposition. Given the singular value decomposition of φ in Eq. (3), we have, for any $v \in \mathcal{V}$, $\varphi(v,\cdot) = \sum_{m \geqslant 1} \mu_m^{1/2} f_m(v) e_m$ and thus

$$q_\lambda^*(v) \propto \langle \varphi(v,\cdot), (\Sigma + \lambda I)^{-1} \varphi(v,\cdot) \rangle_{L_2(d\rho)} = \sum_{m \geqslant 1} \frac{\mu_m}{\mu_m + \lambda} f_m(v)^2,$$

which provides an explicit expression for the density q_λ^*.

For a given squared error value λ, the optimized distribution q_λ^*, while leading to the degrees of freedom that will happen to be optimal in terms of approximation, has two main drawbacks:

- **Dependence on** λ: this implies that if we want a reduced error (i.e., a smaller λ), then the samples obtained from a higher λ, may not be reused to provably obtain the desired bound; in other words, the sampling is not *anytime*. For specific examples, e.g., quadrature with periodic kernels on $[0,1]$ with the uniform distribution, then $q = 1$ happens to be optimal for all λ, and thus, we may reuse samples for different values of the error.
- **Hard to compute in practice**: the optimal distribution depends on a *leverage score* $\langle \varphi(v,\cdot), (\Sigma + \lambda I)^{-1} \varphi(v,\cdot) \rangle_{L_2(d\rho)}$, which may be hard to use for several reasons; first, it requires access to the infinite-dimensional operator Σ, which may be difficult; moreover, even if it possible to invert $\Sigma + \lambda I$, the set $\mathcal{V}$ might be particularly large and impractical to sample from. At the end of Section 4.1, we propose a simple algorithm based on sampling.

Eigenvalues and degrees of freedom. In order to relate more directly to the eigenvalues of Σ, we notice that we may lower bound the degrees of freedom by a constant times the number $m^*(\lambda)$ of eigenvalues greater than λ:

$$d(\lambda) = \operatorname{tr}\Sigma(\Sigma+\lambda I)^{-1} = \sum_{m\geqslant 1}\frac{\mu_m}{\mu_m+\lambda} \geqslant \sum_{\mu_m\geqslant\lambda}\frac{\mu_m}{\mu_m+\lambda} \geqslant \frac{1}{2}\max(\{m,\ \mu_m\geqslant\lambda\}) = m^*(\lambda),$$

as defined in Section 2.1.

Moreover, we have the upper-bound:

$$d(\lambda) = \sum_{\mu_m\geqslant\lambda}\frac{\mu_m}{\mu_m+\lambda} + \sum_{\mu_m<\lambda}\frac{\mu_m}{\mu_m+\lambda} \leqslant \max(\{m,\ \mu_m\geqslant\lambda\}) + \frac{1}{\lambda}\sum_{\mu_m<\lambda}\mu_m.$$

We now make the assumption that there exists a $\gamma>0$ independent of j such that

$$\forall j\geqslant 1,\ \sum_{m=j}^{\infty}\mu_m \leqslant \gamma j\mu_j. \tag{11}$$

This assumption essentially states that the eigenvalues decay sufficiently homogeneously and is satisfied by $\mu_m\propto m^{-2\alpha}$ with $\gamma=(2\alpha-1)^{-1}$, $\mu_m\propto r^m$ with $\gamma=(1-r)^{-1}$ and similar bounds also hold for all examples in Section 2.3. It allows us to relate the degrees of freedom directly to eigenvalue decays.

Indeed, this implies that $\frac{1}{\lambda}\sum_{\mu_m<\lambda}\mu_m \leqslant \gamma\max(\{m,\ \mu_m\geqslant\lambda\}) = m^*(\lambda)$ for all $\lambda\leqslant\mu_1$ (the largest eigenvalue) and thus

$$\frac{1}{2}m^*(\lambda) \leqslant d \leqslant \big[1+\gamma\big]m^*(\lambda).$$

We can now restate the approximation result of Prop. 1 from Section 4.1 with the optimized distribution (see proof in Appendix B.2):

Proposition 2 (Approximation of the unit ball of $\mathcal{F}$ for optimized distribution) *For $\lambda>0$ and the distribution with density q_λ^* defined in Eq. (10) with respect to $d\tau$, with degrees of freedom $d(\lambda)$. Let $v_1,\dots,v_n$ be sampled i.i.d. from the density q, defining the kernel (and its associated RKHS $\hat{\mathcal{F}}$) $\hat{k}(x,y)=\frac{1}{n}\sum_{i=1}^n\frac{1}{q(v_i)}\varphi(v_i,x)\varphi(v_i,y)$. Then, for any $\delta\in(0,1)$, with probability $1-\delta$, we have:*

$$\sup_{\|f\|_{\mathcal{F}}\leqslant 1}\ \inf_{\|\hat{f}\|_{\hat{\mathcal{F}}}\leqslant 2}\big\|f-\hat{f}\big\|^2_{L_2(d\rho)} \leqslant 4\lambda,$$

under any of the following conditions:

- *(a) if $n\geqslant 5\ d(\lambda)\log\big[16d(\lambda)/\delta\big]$,*
- *(b) if Eq. (11) is satisfied, and, by choosing $m\leqslant\frac{n}{5(1+\gamma)\log\frac{16n}{5\delta}}$, and $\lambda=\mu_m$.*

The statement (a) above, is a simple corollary of Prop. 1, and goes from level of error λ to minimum number n of samples. The statement (b) goes in the other direction, that is, from the number of samples n to the achieved approximation error. It depends on the eigenvalues μ_m of the integral

operator taken at $m = O(n/\log(n))$. For example, for polynomial decays of eigenvalues of the form $\mu_m = O(m^{-2s})$, we get (non squared) errors proportional to $(\log n)^s n^{-s}$ for n samples, while for geometric decays, we get geometric errors as a function of the number n of samples.

Note however that for the statement (b) to hold, we need to sample the points $v_1, \dots, v_n$ from the distribution $q^*_{\mu_m}$, that is, for different numbers of samples n, the distribution is unfortunately different (except in special cases). It would be interesting to study the properties of independent but *not identically distributed* samples $v_1, \dots, v_n$ and the possibility of achieving the same rate adaptively.

Corollary for Sobolev spaces. For the sake of concreteness, we consider the special case of $\mathcal{X} = \mathbb{R}^d$ and translation-invariant kernels. We assume that the distribution $d\rho$ is sub-Gaussian. Then for Sobolev spaces of order s, the eigenvalue decay is proportional to $m^{-2s/d}$. Thus, if we can sample from the optimized distribution, after n random features, we obtain an approximation of the unit ball of $\mathcal{F}$ with error $n^{-s/d}$, independently of the chosen expansion, the spatial one used for quadrature or the spectral one used in random Fourier features. For kernels in $\mathbb{R}^d$, these distributions are not readily computed in closed form and need to computed through a dedicated algorithm such as the one we present below.

The same approximation results holds for translation-invariant kernels on $[0,1]^d$; but when $d\rho$ is the uniform distribution, as shown in Section 4.4, the optimized distribution for the quadrature case is still the uniform distribution, for all values of λ, and can thus be computed.

Algorithm to estimate the optimized distribution. We now consider a simple algorithm for estimating the optimized distribution q^*_λ. It is based on using a large number N of points $v_1, \dots, v_N$ from $d\tau$, and replacing $d\tau$ by a potentially weighted empirical distribution $d\hat{\tau}$ associated with these N points. Therefore, we may use any set of points and weights, which leads to a distribution close to $d\tau$. In full generality, only random samples from $d\tau$ are readily available (with weights $1/N$), but for special cases, such as $\mathcal{V} = [0,1]$ or $\mathcal{V} = \mathbb{N}^*$, we may use deterministic representations. See examples in Section 6.

We thus assume that we have N pairs $(v_i, \eta_i) \in \mathcal{V} \times \mathbb{R}_+$, $i = 1, \dots, N$, such that $\sum_{i=1}^n \eta_i = 1$. Since $d\hat{\tau}$ has a finite support with at most N elements, we may identify $L_2(d\hat{\tau})$ and $\mathbb{R}^N$ (with its canonical dot-product), and the operator T goes now from $\mathbb{R}^N$ to $L_2(d\rho)$, with $Tg = \sum_{i=1}^N \eta_i^{1/2} g_i \varphi(v_i, \cdot) \in L_2(d\rho)$, with $T\delta_i = \eta_i^{1/2} \varphi(v_i, \cdot) \in L_2(d\rho)$, for δ_i the i-th element of the canonical basis of $\mathbb{R}^N$. Then, we have:

$$\begin{aligned}
\langle \varphi(v_i, \cdot), (\Sigma + \lambda I)^{-1} \varphi(v_i, \cdot) \rangle_{L_2(d\rho)} &= \eta_i^{-1} \langle T\delta_i, (TT^* + \lambda I)^{-1} T\delta_i \rangle_{L_2(d\rho)} \\
&= \eta_i^{-1} \langle T\delta_i, T(T^*T + \lambda I)^{-1} \delta_i \rangle_{L_2(d\rho)} \\
&= \eta_i^{-1} \big(T^*T(T^*T + \lambda I)^{-1} \big)_{ii}.
\end{aligned}$$

This implies that the density of the optimized distribution with respect to the uniform measure on $\{v_1, \dots, v_N\}$ is proportional to $\big(T^*T(T^*T + \lambda I)^{-1}\big)_{ii}$. We can then sample any number n of points from resampling from $\{v_1, \dots, v_N\}$ from the density above. The computational complexity is $O(N^3)$. A detailed analysis of the approximation properties of this algorithm is outside the scope of this paper.

We have $(T^*T)_{ij} = \eta_i^{1/2}\eta_j^{1/2}\int_{\mathcal{X}} \varphi(v_i,x)\varphi(v_j,x)d\rho(x)$. In some cases, it can be computed in closed form—such as for quadrature where this is equal to $\eta_i^{1/2}\eta_j^{1/2}k(v_i,v_j)$. In some others, it requires i.i.d. samples $x_1,\dots,x_M$ from $d\rho$, and the estimate: $\eta_i^{1/2}\eta_j^{1/2}M^{-1}\sum_{k=1}^M \varphi(v_i,x_k)\varphi(v_j,x_k)$.

4.3 Lower bound

In this section, we aim at providing lower-bounds on the number of samples required for a given accuracy. We have the following result (see proof in Appendix B.3):

Proposition 3 (Lower approximation bound) *For $\delta \in (0,1)$, if we have a family $\psi_1,\dots,\psi_n \in L_2(d\rho)$ such that*

$$\frac{1}{n}\sum_{i=1}^n \|\psi_i\|_{L_2(d\rho)}^2 \leqslant 2\operatorname{tr}\Sigma/\delta, \quad \textit{and} \quad \sup_{\|f\|_{\mathcal{F}}\leqslant 1}\ \inf_{\|\beta\|_2^2\leqslant \frac{4}{n}} \left\| f - \sum_{i=1}^n \beta_i\psi_i\right\|_{L_2(d\rho)}^2 \leqslant 4\lambda,$$

then $n \geqslant \dfrac{\max\{m,\ \mu_m \geqslant 144\lambda\}}{4\log\frac{10\operatorname{tr}\Sigma}{\lambda\delta}}$.

We can make the following observations:

- The proof technique not surprisingly borrows tools from minimax estimation over ellipsoids, namely the Varshamov-Gilbert's lemma.
- We obtain matching upper and lower bounds up to logarithmic terms, using only the decay of eigenvalues $(\mu_m)_{m\geqslant 1}$ of the integral operator Σ (of course, if sampling from the optimized distribution q_λ^* is possible). Indeed in that case, as shown in Prop. 2, we have shown that we need at most $10\ d(\lambda)\log\big[2d(\lambda)\big]$, where $d(\lambda)$ is the degrees of freedom, which is upper and lower bounded by a constant times $m^*(\lambda) = \max\{m,\ \mu_m \geqslant \lambda\}$.
- In order to obtain such a bound, we need to constrain both $\|\beta\|_2$ and the norms of the vectors ψ_i, which correspond to bounded features for the random feature interpretation and tolerance to noise for the quadrature interpretation. We choose our scaling to match the constraints we have in Prop. 1, for which the parameter δ ends up entering the lower bound logarithmically.

4.4 Quadrature

We may specialize the results above to the quadrature case, namely give a formulation where the features φ do not appear (or equivalently using ψ defined in Section 3.2). This is a special case where $\mathcal{V} = \mathcal{X}$ and $\varphi = \psi$. In terms of operators T in Section 2.2, this corresponds to $T = \Sigma^{1/2}$.

Optimized distribution. Following Section 4.1, we have an expression for the optimized distribution, both in terms of operators, as follows,

$$q_\lambda^*(x) \propto \langle \psi(x,\cdot), (\Sigma+\lambda I)^{-1}\psi(x,\cdot)\rangle_{L_2(d\rho)} = \langle \Sigma^{-1/2}k(x,\cdot), (\Sigma+\lambda I)^{-1}\Sigma^{-1/2}k(x,\cdot)\rangle_{L_2(d\rho)},$$

and in terms of eigenvalues and eigenvectors of k, that is,

$$q(x) \propto \langle k(\cdot,x), \Sigma^{-1/2}(\Sigma+\lambda I)^{-1}\Sigma^{-1/2}k(\cdot,x)\rangle_{L_2(d\rho)} = \sum_{m\geqslant 1}\frac{\mu_m}{\mu_m+\lambda}e_m(x)^2. \tag{12}$$

While this is uniform in some special cases (uniform distribution on $[0,1]$ and Sobolev kernels, as shown below), this is typically hard to compute and sample from. An algorithm for approximating it was presented at the end of Section 4.1.

A weakness of our result is that in general our optimized distribution $q_\lambda^*(x)$ depends on λ and thus on the number of samples. In some cases with symmetries (i.e., uniform distribution on $[0,1]$ or the hypersphere), q_λ^* happens to be constant for all λ. Note also that we have observed empirically that in some cases, q_λ^* converges to a certain distribution when λ tends to zero (see an example in Section 6).

Sobolev spaces. For Sobolev spaces with order s in $[0,1]^d$ or $\mathbb{R}^d$ (for which we assume $d < 2s$), the decay of eigenvalues is of the form $m^{-2s/d}$ and thus the error after n samples is $n^{-s/d}$ (up to logarithmic terms), which recovers the upper and lower bounds of Novak (1988, pages 37 and 38) (also up to logarithmic terms).

For the special case of Sobolev spaces on $[0,1]^d$ with $d\rho$ the uniform distribution, the optimized distribution in Eq. (12) is also the uniform distribution. Indeed, the eigenfunctions of the integral operator Σ are d-th order tensor products of the uni-dimensional Fourier basis (the constant and all pairs of sine/cosine at a given frequency), with the *same eigenvalue* for the 2^d possibilities of sines/cosines for a given multi-dimensional frequency $(m_1, \dots, m_d)$. Therefore, when summing all squared values of the eigenfunctions corresponding to $(m_1, \dots, m_d)$, we end up with the sum $\sum_{a \in \{0,1\}^d} \prod_{i=1}^d \cos^{2a_i}(2\pi m_i x_i) \sin^{2(1-a_i)}(2\pi m_i x_i)$, which ends up being constant equal to one (and thus independent of x) because $\cos^{2a_i}(2\pi m_i x_i) + \sin^{2a_i}(2\pi m_i x_i) = 1$.

Finally, we may consider Sobolev spaces on the hypersphere, with the kernels presented in Section 2.3. As shown by Bach (2014, Appendix D.3), the kernel $k(x,y) = \mathbb{E}(v^\top y)_+^s (v^\top y)_+^s$ for v uniform on the hypersphere, leads to a Sobolev space of order $t = s + \frac{d+1}{2}$, while the decay of eigenvalue of the integral operator was shown to be $m^{-1-1/d-2s/d}$ in Section 2.3. It is thus equal to $m^{-2t/d}$, and we recover the result from Hesse (2006).

Quadrature rule. We assume that points $x_1, \dots, x_n$ are sampled from the distribution with density q with respect to $d\rho$. The quadrature rule for a function $h \in \mathcal{F}$ is $\sum_{i=1}^n \frac{\beta_i h(x_i)}{q(x_i)^{1/2}}$. To compute β, we need to minimize with respect to β the error:

$$\left\| \sum_{i=1}^n \frac{\beta_i}{q(x_i)^{1/2}} k(\cdot, x_i) - \int_{\mathcal{X}} k(\cdot, x) g(x) d\rho(x) \right\|_{\mathcal{F}}^2 + n\lambda \|\beta\|_2^2,$$

which is the regularized worst case squared error in the estimation of the integral of h over $h \in \mathcal{F}$. The best error is obtained for $\lambda = 0$, but our guarantees are valid for $\lambda > 0$, with an explicit control over the norm $\|\beta\|_2^2$, which is important for robustness to noise.

Given the values of $\int_{\mathcal{X}} k(x_i, x) g(x) d\rho(x) = z_i$, for $i = 1, \dots, n$, which can be computed in closed form for several triplet $(k, g, d\rho)$ (see, e.g., Smola et al., 2007; Oates and Girolami, 2015), then the problem above is equivalent to minimizing with respect to β:

$$\sum_{i=1}^n \sum_{j=1}^n \frac{\beta_i \beta_j}{q(x_i)^{1/2} q(x_j)^{1/2}} k(x_i, x_j) - \sum_{i=1}^n \frac{\beta_i}{q(x_i)^{1/2}} z_i + n\lambda \|\beta\|_2^2,$$

which leads to a $n \times n$ linear system with running time complexity $O(n^3)$. Note that when adding points sequentially (in particular for kernels for which the distribution q_λ^* is independent of λ, such as Sobolev spaces on $[0,1]$), one may update the solution so that after n steps, the overall complexity is $O(n^3)$.

Approximation of functions in $\mathcal{F}$. With the quadrature weights β estimated above and the quadrature rule $\sum_{i=1}^n \frac{\beta_i h(x_i)}{q(x_i)^{1/2}}$ for the estimation of $\int_{\mathcal{X}} g(x) f(x) d\rho(x)$, we may derive an expression which is explicitly linear in g. Following the proof of Prop. 1 in Appendix B.1, we have, when specialized to the quadrature case:

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n \frac{1}{q(v_i)} \psi(x_i,\cdot) \otimes_{L_2(d\rho)} \psi(x_i,\cdot) = \Sigma^{-1/2}\left(\frac{1}{n}\sum_{i=1}^n \frac{1}{q(v_i)} k(x_i,\cdot) \otimes_{L_2(d\rho)} k(x_i,\cdot)\right)\Sigma^{-1/2},$$

Moreover, we have $\beta_i = \frac{1}{nq(x_i)^{1/2}} \langle k(\cdot,x_i), \Sigma^{-1/2}(\hat{\Sigma}+\lambda I)^{-1}\Sigma^{1/2} g\rangle_{L_2(d\rho)}$ from Eq. (15) in Appendix B.1, and the quadrature rule becomes:

$$\begin{aligned}
\sum_{i=1}^n \frac{\beta_i h(x_i)}{q(x_i)^{1/2}} &= \sum_{i=1}^n \frac{\beta_i}{q(x_i)^{1/2}} \langle h, \Sigma^{-1} k(\cdot,x_i)\rangle_{L_2(d\rho)} \\
&= \left\langle h, \frac{1}{n}\sum_{i=1}^n \Sigma^{-1}\frac{1}{q(x_i)}\big[k(x_i,\cdot)\otimes_{L_2(d\rho)} k(x_i,\cdot)\big]\Sigma^{-1/2}(\hat{\Sigma}+\lambda I)^{-1}\Sigma^{1/2} g\right\rangle_{L_2(d\rho)} \\
&= \big\langle h, \Sigma^{-1/2}\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}\Sigma^{1/2} g\big\rangle_{L_2(d\rho)} = \big\langle g, \Sigma^{1/2}\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}\Sigma^{-1/2} h\big\rangle_{L_2(d\rho)},
\end{aligned}$$

which can be put in the form $\langle \hat{h}, g\rangle_{L_2(d\rho)}$ with the approximation $\hat{h} = \Sigma^{1/2}\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}\Sigma^{-1/2} h$ of the function $h \in \mathcal{F}$. Having a bound for all functions g such that $\|g\|_{L_2(d\rho)} \leqslant 1$ is equivalent to having a bound on $\|h - \hat{h}\|_{L_2(d\rho)}$. In Section 5, we consider extensions, where we consider other norms than the L_2-norm for characterizing the approximation error $\hat{h} - h$. Moreover, we consider cases where h belongs to a strict subspace of $\mathcal{F}$ (with improved results).

4.5 Learning with random features

We consider supervised learning with m i.i.d. samples from a distribution on inputs/outputs (x,y), and a uniformly G-Lipschitz-continuous loss function $\ell(y,\cdot)$, which includes logistic regression and the support vector machine. We consider the empirical risk $\hat{L}(f) = \frac{1}{m}\sum_{i=1}^m \ell(y_i, f(x_i))$ and the expected risk $L(f) = \mathbb{E}\ell(y, f(x))$, with x having the marginal distribution $d\rho$ that we consider in earlier sections. We assume that $\mathbb{E}k(x,x) = \operatorname{tr}\Sigma = R^2$. We have the usual generalization bound for the minimizer $\hat{f}$ of $\hat{L}(f)$ with respect to $\|f\|_{\mathcal{F}} \leqslant F$, based on Rademacher complexity (see, e.g., Shalev-Shwartz and Ben-David, 2014):

$$\mathbb{E}\big[L(\hat{f})\big] \leqslant \inf_{\|f\|_{\mathcal{F}}\leqslant F} L(f) + 2\mathbb{E}\Big[\sup_{\|f\|_{\mathcal{F}}\leqslant F} |L(f) - \hat{L}(f)|\Big] \leqslant \inf_{\|f\|_{\mathcal{F}}\leqslant F} L(f) + \frac{4FGR}{\sqrt{m}}. \tag{13}$$

We now consider learning by sampling n features from the optimized distribution from Section 4.2, leading to a function parameterized by $\beta \in \mathbb{R}^n$, that is $\hat{g}_\beta = \sum_{i=1}^n \beta_i q(v_i)^{-1/2}\varphi(v_i,\cdot) \in L_2(d\rho)$.

Applying results from Section 4.1, we assume that $\lambda \leqslant R^2/4$ and $n \geqslant 5d(\lambda)\log\frac{2(\operatorname{tr}\Sigma)d(\lambda)}{\lambda}$, where $d(\lambda)$ is equal to the degrees of freedom associated with the kernel k and distribution $d\rho$. Thus, the expected squared error for approximating the unit-ball of $\mathcal{F}$ by the ball of radius 2 of the approximation $\hat{\mathcal{F}}$ obtained from the approximated kernel is less than 8λ.

If we consider the estimator $\hat{\beta}$ obtained by minimizing the empirical risk of $\hat{g}_\beta$ subject to $\|\beta\|_2 \leqslant 2F/\sqrt{n}$. We have the following decomposition of the error for any $\gamma \in \mathbb{R}^n$ such that $\|\gamma\|_2 \leqslant 2F/\sqrt{n}$ and $f \in \mathcal{F}$ such that $\|f\|_{\mathcal{F}} \leqslant F$:

$$\begin{aligned}
L(\hat{g}_{\hat{\beta}}) &= L(\hat{g}_{\hat{\beta}}) - \hat{L}(\hat{g}_{\hat{\beta}}) + \hat{L}(\hat{g}_{\hat{\beta}}) - \hat{L}(\hat{g}_\gamma) + \hat{L}(\hat{g}_\gamma) - L(\hat{g}_\gamma) + L(\hat{g}_\gamma) - L(f) + L(f) \\
&\leqslant 2\Big[\sup_{\|\beta'\|_{\mathcal{F}} \leqslant 2F/\sqrt{n}} |L(\hat{g}_{\beta'}) - L(\hat{g}_{\beta'})|\Big] + \big[L(\hat{g}_\gamma) - L(f)\big] + L(f) \\
&\leqslant 2\Big[\sup_{\|\beta'\|_{\mathcal{F}} \leqslant 2F/\sqrt{n}} |L(\hat{g}_{\beta'}) - L(\hat{g}_{\beta'})|\Big] + \sup_{\|f'\|_{\mathcal{F}} \leqslant F} \inf_{\|\gamma\|_2 \leqslant 2F/\sqrt{n}} \big[L(\hat{g}_\gamma) - L(f')\big] + \inf_{\|f\|_{\mathcal{F}} \leqslant F} L(f).
\end{aligned}$$

We now take expectation with respect to the data and the random features. Following standard results for Rademacher complexities of ℓ_2-balls (Bartlett and Mendelson, 2003, Lemma 22), the first term is less than

$$\frac{4FG}{m\sqrt{n}}\mathbb{E}\Big(\sum_{i=1}^m \sum_{j=1}^n \frac{\varphi(v_i, x_j)^2}{q(v_i)}\Big)^{1/2} \leqslant \frac{4FG}{m\sqrt{n}}(nm \operatorname{tr}\Sigma)^{1/2} = \frac{4FGR}{\sqrt{m}}.$$

Because of the G-Lipschitz-continuity of the loss, we have $L(\hat{g}_\gamma) - L(f') \leqslant G\|\hat{g}_\gamma) - f'\|_{L_2(d\rho)}$, and thus the second term is less than $\sqrt{8\lambda}GF \leqslant 3GF\sqrt{\lambda}$. Overall, we obtain

$$\mathbb{E}\big[L(\hat{g}_{\hat{\beta}})\big] \leqslant \inf_{\|f\|_{\mathcal{F}} \leqslant F} L(f) + 3GF\sqrt{\lambda} + \frac{4FGR}{\sqrt{m}}.$$

If we consider $\lambda = R^2/m$ in order to lose only a constant factor compared to Eq. (13), we have the constraint $n \geqslant 5d(R^2/m)\log\big[2md(R^2/m)\big]$.

We may now look at several situations. In the worst case, where the decay of eigenvalue is not fast, i.e., very close to $1/i$, then we may only use the bound $d(\lambda) = \operatorname{tr}\Sigma(\Sigma+\lambda I)^{-1} \leqslant \lambda^{-1}\operatorname{tr}\Sigma = R^2/\lambda$, and thus a sufficient condition $n \geqslant 10m\log 2m$, and we obtain the same result as Rahimi and Recht (2009).

However, when we have eigenvalue decays as $R^2 i^{-2s}$, we get (up to constants), following the same computation as Section 4.2, $d(\lambda) \leqslant (R^2/\lambda)^{1/(2s)}$, and thus $n \geqslant m^{1/(2s)}\log m$, which is a significant improvement (regardless of the value of F). Moreover, if the decay is geometric as r^i, then we get $d(\lambda) \leqslant \log(R^2/\lambda)$, and thus $n \geqslant (\log m)^2$, which is even more significant.

5. Quadrature-related Extensions

In Section 4.4, we have built an approximation $\hat{h} = \Sigma^{1/2}\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}\Sigma^{-1/2}h$ of a function $h \in \mathcal{F}$, which is based on n function evaluations $h(x_1), \dots, h(x_n)$. We have presented in Section 4.4 a convergence rate for the L_2-norm $\|\hat{h} - h\|_{L_2(d\rho)}$ for functions h with less than unit $\mathcal{F}$-norm $\|h\|_{\mathcal{F}} \leqslant 1$. Up to logarithmic terms, if using the optimal distribution for sampling $x_1, \dots, x_n$, then we get a squared error of μ_n where μ_n is the n-th largest eigenvalue of the integral operator Σ.

Robustness to noise. We have seen that if the noise in the function evaluations $h(x_i)$ has a variance less than $q(x_i)\tau^2$, then the error $\|h - \hat{h}\|^2_{L_2(d\rho)}$ has an additional term $\tau^2\|\beta\|_2^2 \leqslant \frac{4\tau^2}{n}$. Hence, the amount of noise has to be less than $n\mu_n$ in order to incur no loss in performance (a bound which decreases with n).

Adaptivity to smoother functions. We assume that the function h happens to be smoother than what is sufficient to be an element of the RKHS $\mathcal{F}$, that is, if $\|\Sigma^{-s}h\|_{L_2(d\rho)} \leqslant 1$, where $s \geqslant 1/2$. The case $s = 1/2$ corresponds to being in the RKHS. In the proof of Prop. 1 in Appendix B.1, we have seen that with high-probability we have:

$$(\hat{\Sigma} + \lambda I)^{-1} \preccurlyeq 4(\Sigma + \lambda I)^{-1}. \tag{14}$$

We now see that we can bound the error $\|\hat{h} - h\|_{L_2(d\rho)}$ as follows:

$$\begin{aligned}
\|\hat{h} - h\|_{L_2(d\rho)} &= \|\Sigma^{1/2}\hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-1}\Sigma^{-1/2}h - h\|_{L_2(d\rho)} \\
&= \lambda\big\|\Sigma^{1/2}(\hat{\Sigma} + \lambda I)^{-1}\Sigma^{-1/2+s}\Sigma^{-s}h\big\|_{L_2(d\rho)} \\
&\leqslant \lambda\big\|\Sigma^{1/2}(\hat{\Sigma} + \lambda I)^{-1/2}\big\|_{\rm op}\big\|(\hat{\Sigma} + \lambda I)^{-1/2}\Sigma^{-1/2+s}\big\|_{\rm op}\|\Sigma^{-s}h\|_{L_2(d\rho)}.
\end{aligned}$$

We may now bound each term. The first one $\big\|\Sigma^{1/2}(\hat{\Sigma} + \lambda I)^{-1/2}\big\|_{\rm op}$ is less than 2, because of Eq. (14). The second one $\big\|(\hat{\Sigma}+\lambda I)^{-1/2}\Sigma^{-1/2+s}\big\|_{\rm op}$ is equal to $\big\|(\hat{\Sigma}+\lambda I)^{s-1}(\hat{\Sigma}+\lambda I)^{1/2-s}\Sigma^{-1/2+s}\big\|_{\rm op}$, and thus less than $\big\|(\hat{\Sigma} + \lambda I)^{s-1}\|_{\rm op} \cdot \big\|(\hat{\Sigma} + \lambda I)^{1/2-s}\Sigma^{-1/2+s}\big\|_{\rm op} \leqslant 2\lambda^{s-1}$. Overall we obtain

$$\|\hat{h} - h\|_{L_2(d\rho)} \leqslant 4\lambda^s.$$

The norm $h \mapsto \|\Sigma^{-s}h\|_{L_2(d\rho)}$ is an RKHS norm with kernel $\sum_{m\geqslant 0}\mu_m^{2s}e_m(x)e_m(y)$, with corresponding eigenvalues equal to $(\mu_m)^{2s}$. From Prop. 2 and 3, the optimal number of quadrature points to reach a squared error less than ε is proportional to the number $\max(\{m,\ \mu_m^{2s} \geqslant \varepsilon\})$, while using the quadrature points from $s = 1/2$, leads to a number $\max(\{m,\ \mu_m \geqslant \varepsilon^{1/(2s)}\})$, which is equal. Thus if the RKHS used to compute the quadrature weights is a bit too large (but not too large, see experiments in Section 6), then we still get the optimal rate. Note that this robustness is only shown for the regularized estimation of the quadrature coefficients (in our simulations, the non-regularized ones also exhibit the same behavior).

Approximation with stronger norms. We may consider characterizing the difference $\hat{h} - h$ with different norms than $\|\cdot\|_{L_2(d\rho)}$, in particular norms $\|\Sigma^{-r}(\hat{h} - h)\|_{L_2(d\rho)}$, with $r \in [0, 1/2]$. For $r = 0$, this is our results in L_2-norm, while for $r = 1/2$, this is the RKHS norms. We have, using the same manipulations than above:

$$\begin{aligned}
\|\Sigma^{-r}(\hat{h} - h)\|_{L_2(d\rho)} &= \lambda\big\|\Sigma^{1/2-r}(\hat{\Sigma} + \lambda I)^{-1}\Sigma^{-1/2}h\big\|_{L_2(d\rho)} \\
&\leqslant \lambda^{1/2-r}\big\|\Sigma^{1/2-r}(\hat{\Sigma} + \lambda I)^{r-1/2}\big\|_{\rm op}\|\Sigma^{-1/2}h\|_{L_2(d\rho)} \leqslant 2\lambda^{1/2-r}.
\end{aligned}$$

When $r = 1/2$, we get a result in the RKHS norm, but with no decay to zero; the RKHS norm $\|\cdot\|_{\mathcal{F}}$ would allow a control in L_∞-norm, but as noticed by Steinwart et al. (2009); Mendelson and Neeman (2010), such a control may be obtained in practice with r much smaller. For example, when the

eigenfunctions e_m are uniformly bounded in L_∞-norm by a constant C (as is the case for periodic kernels in $[0,1]$ with the uniform distribution), then, for any $x \in \mathcal{X}$, we have for $t > 1$,

$$f(x)^2 = \sum_{m=1}^{\infty} (m+1)^t \langle f, e_m \rangle_{L_2(d\rho)}^2 e_m(x)^2 (m+1)^{-t} \leqslant \sum_{m=0}^{\infty} (m+1)^t \langle f, e_m \rangle_{L_2(d\rho)}^2 \frac{C^2}{t-1}.$$

If for simplicity, we assume that $\mu_m = (m+1)^{-2s}$ (like for Sobolev spaces), we have $\|\Sigma^{-r} f\|_{L_2(d\rho)}^2 = \sum_{m=1}^{\infty} \mu_m^{-2r} \langle f, e_m \rangle_{L_2(d\rho)}^2 = \sum_{m=1}^{\infty} (m+1)^t \langle f, e_m \rangle_{L_2(d\rho)}^2$ with $r = t/4s$. If $\lambda \leqslant O(n^{-2s})$ (as suggested by Prop. 1), then we obtain a squared L_∞-error less than $\frac{1}{t-1} \lambda^{1-2r} = O\big(\frac{1}{t-1} n^{-2s(1-t/2s)}\big) = O\big(\frac{n^t}{t-1} n^{-2s}\big)$. With $t = 1 + \frac{1}{\log n}$, we get $O\big(\frac{n \log n}{n^{-2s}}\big)$, and thus a degradation compared to the squared L_2-loss of n (plus additional logarithmic terms), which corresponds to the (non-improvable) result of Novak (1988, page 36).

6. Simulations

In this section, we consider simple illustrative quadrature experiments[4] with $\mathcal{X} = [0,1]$ and kernels $k(x,y) = 1 + \sum_{m=1}^{\infty} \frac{1}{m^{2s}} \cos 2\pi m(x-y)$, with various values of s and distributions $d\rho$ which are Beta random variable with the two parameters equal to $a = b$, hence symmetric around $1/2$.

Uniform distribution. For $b = 1$, we have the uniform distribution on $[0,1]$ for which the cosine/sine basis is orthonormal, and the optimized distribution q_λ^* is also uniform. Moreover, we have $\int_0^1 k(x,y) d\rho(x) = 1$. We report results comparing different Sobolev spaces for testing functions to integrate (parameterized by s) and learning quadrature weights (parameterized by t) in Figure 1, where we compute errors averaged over 1000 draws. We did not use regularization to compute quadrature weights α. We can make the following observations:

- The exponents in the convergence rates for $s = t$ (matching RKHSs for learning quadrature weights and testing functions) are close to $2s$ as expected.
- When the functions to integrate are less smooth than the ones used for learning quadrature weights (that is $t > s$), then the quadrature performance does not necessarily decay with the number of samples.
- On the contrary, when $s > t$, then we have convergence and the rate is potentially worse than the optimal one (attained for $s = t$), and equal when $t \geqslant s/2$, as shown in Section 5.

In Figure 2, we compare several quadrature rules on $[0,1]$, namely Simpson's rule with uniformly spread points, Gauss-Legendre quadrature and the Sobol sequence with uniform weights. For $s = 1$, as expected, all squared errors decay as n^{-2} with a worse constant for our kernel-based rule, while for $s = 2$ (smoother test functions), the Sobol sequence is not adaptive, while all others are adaptive and get convergence rates around n^{-4}.

4. Matlab code for all 5 figures may be downloaded from http://www.di.ens.fr/~fbach/quadrature.html.

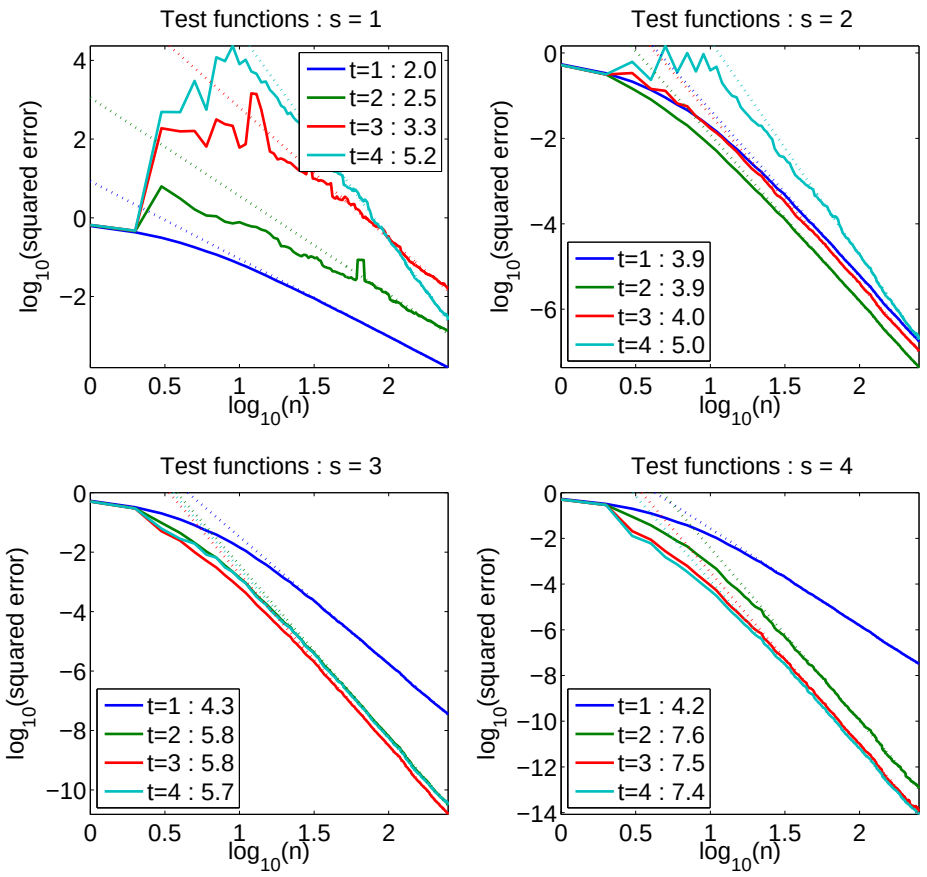

Figure 1: Quadrature for functions in a Sobolev space with parameter s (four possible values) for the uniform distribution on $[0, 1]$, with quadrature rules obtained from different Sobolev spaces with parameters t (same four possible values). We compute affine fits in log-log-space (in dotted) to estimate convergence rates of the form C/n^u and report the value of u. Best seen in color.

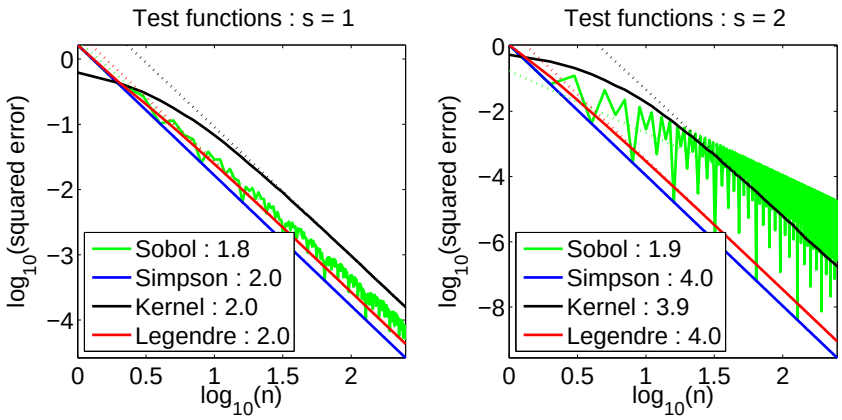

Figure 2: Quadrature for functions in a Sobolev space with parameters $s = 1$ (left) and $s = 2$ (right), for the uniform distribution on $[0, 1]$, with various quadrature rules. We compute affine fits in log-log-space (in dotted) to estimate convergence rates of the form C/n^u and report the value of u. Best seen in color.

Non-uniform distribution. We consider the case $a = b = 1/2$, which is the distribution $d\rho$ with density $\pi^{-1}x^{-1/2}(1-x)^{-1/2}$ with respect to the Lebesgue measure, and with cumulative distribution function $F(x) = \pi^{-1}\arccos(1-2x)$. We may use an approximation of $d\tau$ with N unweighted points $F^{-1}(k/N) = \big(1 - \cos\frac{k\pi}{N}\big)/2$, for $k \in \{1, \dots, N\}$ and the algorithms from the end of Section 4.2. We consider the Sobolev kernel with $s = 1$.

In Figure 3, we plot all densities q_λ^* as a function of λ. When λ is large, we unsuprisingly obtain the uniform density, while, more surprisingly, when λ tends to zero, the density tends to a density, which happens here to be proportional to $x^{1/4}(1-x)^{1/4}$ (leading to a Beta distribution with parameters $a = b = .25$).

We may also consider the same kernel but with the Fourier expansion on $\mathbb{N}$. This is done by representing $d\tau \propto \delta_0 + \sum_{k\in\mathbb{Z}^*} \frac{1}{k^2}\delta_k$ by truncating to all $|k| \leqslant K$, with $K = 50$, which is a weighted representation. We plot in Figure 4 the optimal density over the set of integers, both with respect to the input density (which decays as $1/n^2$) and the counting measure. When λ is large, we recover the input density, while when λ tends to zero, q_λ^* tends to be uniform (and thus, does not converge to a finite measure).

7. Conclusion

In this paper, we have shown that kernel-based quadrature rules are a special case of random feature expansions for positive definite kernels, and derived upper and lower bounds on approximations, that match up to logarithmic terms. For quadrature, this leads to widely applicable results while for random features this allows a significantly improved guarantee within a supervised learning framework.

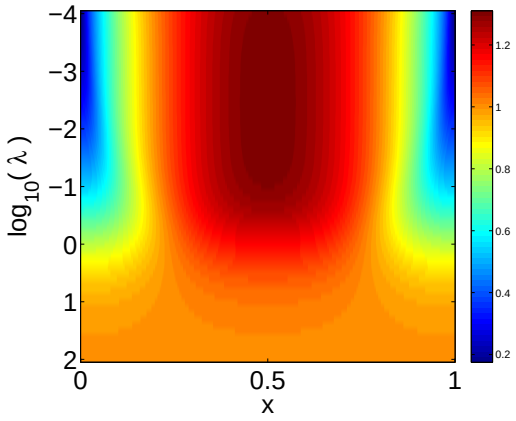

Figure 3: Optimal log-densities $q^*_\lambda(x)$ (with respect to the input distribution) for several values of λ, for the expansion used for quadrature. Best seen in color.

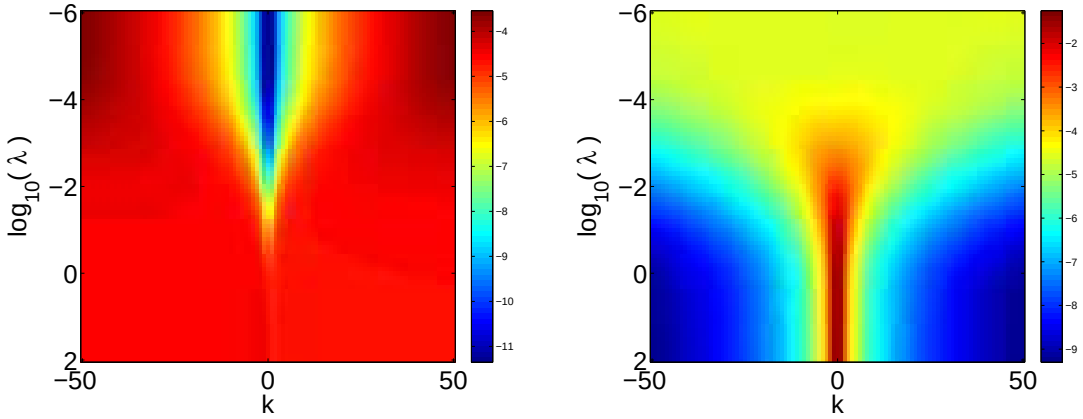

Figure 4: Optimal densities $q^*_\lambda(k)$ for several values of λ, for Fourier feature expansions. Left: with respect to the input distribution (which itself has distribution proportional to $1/k^2$ with respect to the counting measure); right: with respect to the counting measure. Best seen in color.

The present work could be extended in a variety of ways, for example towards bandit optimization rather than quadrature (Srinivas et al., 2012), the use of quasi-random sampling within our framework in the spirit of Yang et al. (2014); Oates and Girolami (2015), a similar analysis for kernel herding (Chen et al., 2010; Bach et al., 2012), an extension to fast rates for non-parametric least-squares regression (Hsu et al., 2014) but with an improved computational complexity, and a study of the consequences of our improved approximation result for online learning and stochastic approximation, in the spirit of Dai et al. (2014); Dieuleveut and Bach (2014).

Acknowledgements

This work was partially supported by the MSR-Inria Joint Centre and a grant by the European Research Council (SIERRA project 239993). Comments of the reviewers were greatly appreciated and helped improve the presentation significantly. The author would like to thank the STVI for the opportunity of writing a single-handed paper.

References

R. A. Adams and J. F. Fournier. *Sobolev Spaces*, volume 140. Academic Press, 2003.

F. Bach. Sharp analysis of low-rank kernel matrix approximations. In *Proceedings of the International Conference on Learning Theory (COLT)*, 2013.

F. Bach. Breaking the curse of dimensionality with convex neural networks. Technical Report 01098505, HAL, 2014.

F. Bach, S. Lacoste-Julien, and G. Obozinski. On the equivalence between herding and conditional gradient algorithms. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

C. R. Baker. Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289, 1973.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2003.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, volume 3. Springer, 2004.

R. Bhatia. *Positive definite matrices*. Princeton University Press, 2009.

M. Sh. Birman and M. Z. Solomyak. Estimates of singular numbers of integral operators. *Russian Mathematical Surveys*, 32(1):15–89, 1977.

A. D. Bull. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 12:2879–2904, 2011.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.*, 7(3):331–368, 2007.

K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3): 273–304, 1995.

Y. Chen, M. Welling, and A. Smola. Super-samples from kernel herding. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2010.

Y. Cho and L. K. Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

W. G. Cochran and G. M. Cox. *Experimental designs*. John Wiley & Sons, 1957.

D. Cruz-Uribe and C. J. Neugebauer. Sharp error bounds for the trapezoidal rule and Simpson's rule. *Journal of Inequalities in Pure and Applied Mathematics*, 3(4), 2002.

B. Dai, B. Xie, N. He, Y. Liang, A. Raj, M.-F. Balcan, and L. Song. Scalable kernel methods via doubly stochastic gradients. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

A. Dieuleveut and F. Bach. Non-parametric stochastic approximation with large step sizes. Technical Report 1408.0361, ArXiv, 2014.

A. El Alaoui and M. W. Mahoney. Fast randomized kernel methods with statistical guarantees. Technical Report 1411.0306, arXiv, 2014.

S. Fine and K. Scheinberg. Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, 2:243–264, 2001.

E. M. Furrer and D. W. Nychka. A framework to understand the asymptotic properties of kriging and splines. *Journal of the Korean Statistical Society*, 36(1):57–76, 2007.

A. Gelman. *Bayesian Data Analysis*. CRC Press, 2004.

Z. Harchaoui, F. Bach, and E. Moulines. Testing for homogeneity with kernel Fisher discriminant analysis. Technical Report 00270806, HAL, April 2008.

T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman & Hall, 1990.

K. Hesse. A lower bound for the worst-case cubature error on spheres of arbitrary dimension. *Numerische Mathematik*, 103(3):413–433, 2006.

F. B. Hildebrand. *Introduction to Numerical Analysis*. Courier Dover Publications, 1987.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 2012.

D. Hsu, S. M. Kakade, and T. Zhang. Tail inequalities for sums of random matrices that depend on the intrinsic dimension. *Electronic Communications in Probability*, 17(14):1–13, 2012.

D. Hsu, S. M. Kakade, and T. Zhang. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14(3):569–600, 2014.

G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1):489–501, 2006.

F. Huszár and D. Duvenaud. Optimally-weighted herding is Bayesian quadrature. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2012.

T. Kato. *Perturbation theory for linear operators*. Springer Science & Business Media, 1995.

H. König. Eigenvalues of compact operators with applications to integral operators. *Linear Algebra and its Applications*, 84:111–122, 1986.

M. Langberg and L. J. Schulman. Universal epsilon-approximators for integrals. In *Proceedings of ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2010.

Q. Le, T. Sarlós, and A. Smola. Fastfood: approximating kernel expansions in log-linear time. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2013.

M. W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011.

M. W. Mahoney and P. Drineas. CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences*, 106(3):697–702, 2009.

P. Massart. *Concentration Inequalities and Model Selection: Ecole d'été de Probabilités de Saint-Flour 23*. Springer, 2003.

S. Mendelson and J. Neeman. Regularization in kernel learning. *The Annals of Statistics*, 38(1): 526–565, 2010.

S. Minsker. On some extensions of Bernstein's inequality for self-adjoint operators. Technical Report 1112.5448, arXiv, 2011.

W. J. Morokoff and R. E. Caflisch. Quasi-random sequences and their discrepancies. *SIAM Journal on Scientific Computing*, 15(6):1251–1279, 1994.

R. M. Neal. *Bayesian Learning for Neural Networks*. PhD thesis, University of Toronto, 1995.

E. Novak. *Deterministic and Stochastic Error Bounds in Numerical Analysis*. Springer-Verlag, 1988.

C. J. Oates and M. Girolami. Variance reduction for quasi-Monte-Carlo. Technical Report 1501.03379, arXiv, 2015.

H. Ogawa. An operator pseudo-inversion lemma. *SIAM Journal on Applied Mathematics*, 48(6): 1527–1531, 1988.

A. O'Hagan. Bayes-Hermite quadrature. *Journal of statistical planning and inference*, 29(3):245–260, 1991.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

A. Rahimi and B. Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

C. E. Rasmussen and Z. Ghahramani. Bayesian Monte Carlo. In *Advances in Neural Information Processing Systems (NIPS)*, 2003.

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer New York, 2005.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

B. Simon. *Trace ideals and their applications*, volume 35. Cambridge University Press, 1979.

S. Smale and F. Cucker. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49, 2001.

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

A. J. Smola and B. Schölkopf. Sparse greedy matrix approximation for machine learning. In *Proc. ICML*, 2000.

A. J. Smola, Z. L. Ovari, and R. C. Williamson. Regularization with dot-product kernels. *Advances in Neural Information Processing Systems (NIPS)*, 2001.

N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, 2012.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

I. Steinwart, D. R. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In *Proceedings of the International Conference on Learning Theory (COLT)*, 2009.

G. Wahba. *Spline Models for observational data*. SIAM, 1990.

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

C. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *Adv. NIPS*, 2001.

J. Yang, V. Sindhwani, H. Avron, and M. Mahoney. Quasi-Monte Carlo feature maps for shift-invariant kernels. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014.

L. Zwald, G. Blanchard, P. Massart, and R. Vert. Kernel projection machine: a new tool for pattern recognition. In *Advances in Neural Information Processing Systems (NIPS)*, 2004.

Appendix A. Kernels on product spaces

In this appendix, we consider sets $\mathcal{X}$ which are products of several simple sets $\mathcal{X}_1, \dots, \mathcal{X}_d$, with known kernels $k_1, \dots, k_d$, each with RKHS $\mathcal{F}_1, \dots, \mathcal{F}_d$. We also assume that we have d measures $d\rho_1, \dots, d\rho_d$, leading to sequences of eigenvalues $(\mu_{jm_j})_{m_j \geqslant 1}$ and eigenfunctions $(e_{jm_j})_{m_j \geqslant 1}$.

Our aim is to define a kernel k on $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ with the product measure $d\rho = d\rho_1 \cdots d\rho_d$. For illustration purposes, we consider decays of the form $\mu_m \propto m^{-2s}$ for the d kernels, that will be useful for Sobolev spaces. We also consider the case where $\mu_m \propto \exp(-\rho m)$. For some combinations, eigenvalue decay is the most natural, in others, the number of eigenvalues $m^*(\lambda)$ greater than a given $\lambda > 0$ is more natural.

A.1 Sum of kernels: $k(x,y) = \sum_{j=1}^d k_j(x_j, y_j)$

In this situation, the RKHS for k is isomorphic to $\mathcal{F}_1 \times \cdots \times \mathcal{F}_d$, composed of functions g such that there exists $f_1, \dots, f_d$ in $\mathcal{F}_1, \dots, \mathcal{F}_d$ such that $g(x) = \sum_{j=1}^d f_j(x_j)$, that is we obtain separable functions, which are sometimes used in the context of generalized additive models (Hastie and Tibshirani, 1990). The corresponding integral operator is then block-diagonal with j-th block equal to the integral operator for k_j and $d\rho_j$. This implies that that its eigenvalues are the concatenation of all sequences $(\mu_{jm_j})_{m_j \geqslant 0}$. Thus the function $m^*(\lambda)$ is the sum of functions $m_1^*(\lambda) + \cdots + m_d^*(\lambda)$.

In terms of norms of functions, we have a norm equal to $\|g\|_{\mathcal{F}}^2 = \sum_{j=1}^d \|f_j\|_{\mathcal{F}_j}^2$.

In the particular case where $\mu_{jm_j} \propto m_j^{-2s}$ for all j, or equivalently, a number of eigenvalues of k_j greater than λ proportional to $\lambda^{-1/(2s)}$, we have a number of eigenvalues of k greater than λ equivalent to $d\lambda^{-1/(2s)}$, that is a decay for the eigenvalues proportional to $(m/d)^{-2s}$. Similarly, when the decay is exponential as $\exp(-\rho m)$, we get a decay of $\exp(-\rho m/d)$.

A.2 Product of kernels: $k(x,y) = \prod_{j=1}^d k_j(x_j, y_j)$

In this situation, the RKHS for K is exactly the tensor product of $\mathcal{F}_1, \dots, \mathcal{F}_d$, i.e., the span of all functions $\prod_{j=1}^d f_j(x_j)$, for $f_1, \dots, f_d$ in $\mathcal{F}_1, \dots, \mathcal{F}_d$ (Berlinet and Thomas-Agnan, 2004). Moreover, the integral operator for k is a tensor product of the d integral operator for $k_1, \dots, k_d$. This implies that its eigenvalues are $\mu_{1m_1} \times \cdots \times \mu_{dm_d}$, $m_1, \dots, m_d \geqslant 0$. In terms of norms of functions defined on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_d$, this thus corresponds to

$$\sum_{m_1, \dots, m_d \geqslant 0} \left(\prod_{j=1}^d \mu_{jm_j} \right)^{-1} \left\langle f, \prod_{j=1}^d e_{jm_j}(x_j) \right\rangle_{L_2(d\rho^{\otimes d})}^2 .$$

Special cases. In the particular case where $\mu_{jm_j} \propto m_j^{-2s}$ for all j, we have a number of eigenvalues of k greater than λ equivalent to the number of multi-indices such that $m_1 \times \cdots \times m_d$ is less than $\lambda^{-1/(2s)}$. By counting first the index m_1, this can be upper-bounded by the sum of $\frac{\lambda^{-1/(2s)}}{m_2 \cdots m_d}$ over all indices $m_2, \dots, m_d$ less than $\lambda^{-1/(2s)}$, which is less than $\lambda^{-1/(2s)} \big(\sum_{m=1}^{\lambda^{-1/(2s)}} \frac{1}{m} \big)^{d-1} = O\Big(\lambda^{-1/(2s)} \big(s \log \frac{1}{\lambda} \big)^{d-1} \Big)$. This results in a decay of eigenvalues bounded by $(\log m)^{2s(d-1)} m^{-2s}$ (this can be obtained by inverting approximately the function of λ).

When the decay is exponential as $\exp(-\rho\lambda)$, then we get that $m^*(\lambda)$ is the number of multi-indices $(m_1, \ldots, m_d)$ such that their sum is less than $c = \frac{\log \frac{1}{\lambda}}{\rho}$; when c is large, this is equivalent to c^d times the volume of the d-dimensional simplex, and thus less than $\frac{c^d}{d!} = \big(\frac{\log \frac{1}{\lambda}}{\rho}\big)^d \frac{1}{d!}$. This leads to a decay of eigenvalues as $\exp(-\rho d!^{1/d} m^{1/d})$ or, by using Stirling formula, less than $\exp(-\rho d m^{1/d})$.

Appendix B. Proofs

B.1 Proof of Prop. 1

As shown in Section 2.2, any $f \in \mathcal{F}$ with $\mathcal{F}$-norm less than one may be represented as $f = \int_{\mathcal{V}} g(v)\varphi(v,\cdot)d\tau(v)$, for a certain $g \in L_2(d\tau)$ with $L_2(d\tau)$-norm less than one. We do not solve the problem in β exactly, but use a properly chosen Lagrange multiplier λ and consider the following minimization problem:

$$\Big\| \sum_{i=1}^n \beta_i q(v_i)^{-1/2} \varphi(v_i,\cdot) - \int_{\mathcal{X}} \varphi(v,\cdot) g(v) d\tau(v) \Big\|^2_{L_2(d\rho)} + n\lambda \|\beta\|_2^2.$$

We consider the operator $\Phi : \mathbb{R}^n \to L_2(d\rho)$ such that

$$\Phi\beta = \sum_{i=1}^n \beta_i q(v_i)^{-1/2} \varphi(v_i,\cdot).$$

We then need to minimize the familiar least-squares problem:

$$\big\| f - \Phi\beta \big\|^2_{L_2(d\rho)} + n\lambda \|\beta\|_2^2,$$

with solution from the usual normal equations and the matrix inversion lemma for operators (Ogawa, 1988):

$$\beta = (\Phi^*\Phi + n\lambda I)^{-1}\Phi^* f = \frac{1}{n}\Phi^*(\frac{1}{n}\Phi\Phi^* + \lambda I)^{-1} f. \tag{15}$$

We consider the empirical integral operator $\hat{\Sigma} : L_2(d\rho) \to L_2(d\rho)$, defined as

$$\hat{\Sigma} = \frac{1}{n}\Phi\Phi^* = \frac{1}{n}\sum_{i=1}^n \frac{1}{q(v_i)} \varphi(v_i,\cdot) \otimes_{L_2(d\rho)} \varphi(v_i,\cdot),$$

that is, for $a, b \in L_2(d\rho)$, $\langle a, \hat{\Sigma} b\rangle_{L_2(d\rho)} = \displaystyle\sum_{i=1}^n \frac{\langle a, \varphi(v_i,\cdot)\rangle_{L_2(d\rho)} \langle b, \varphi(v_i,\cdot)\rangle_{L_2(d\rho)}}{q(v_i)}$. By construction, and following the end of Section 2.2, we have $\mathbb{E}\hat{\Sigma} = \Sigma$.

The value of $\|f - \Phi\beta\|^2_{L_2(d\rho)}$ is equal to

$$\begin{aligned} \|f - \Phi\beta\|^2_{L_2(d\rho)} &= \|f - \hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-1} f\|^2_{L_2(d\rho)} = \|\lambda(\hat{\Sigma} + \lambda I)^{-1} f\|^2_{L_2(d\rho)} \\ &= \lambda^2 \big\langle f, (\hat{\Sigma} + \lambda I)^{-2} f \big\rangle_{L_2(d\rho)} \leqslant \lambda \big\langle f, (\hat{\Sigma} + \lambda I)^{-1} f \big\rangle_{L_2(d\rho)}, \end{aligned} \tag{16}$$

because $(\hat{\Sigma}+\lambda I)^{-2} \preccurlyeq \lambda^{-1}(\hat{\Sigma}+\lambda I)^{-1}$ (with the classical partial order between self-adjoint operators).

Finally, we have, with $\beta = \frac{1}{n}\Phi^*(\hat{\Sigma}+\lambda I)^{-1}f$:

$$n\|\beta\|_2^2 = \big\langle (\hat{\Sigma}+\lambda I)^{-1}f, \hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}f\big\rangle_{L_2(d\rho)} \leqslant \big\langle f, (\hat{\Sigma}+\lambda I)^{-1}f\big\rangle_{L_2(d\rho)}, \tag{17}$$

using $(\hat{\Sigma}+\lambda I)^{-2}\hat{\Sigma} \preccurlyeq (\hat{\Sigma}+\lambda I)^{-1}$.

By construction, we have $\mathbb{E}(\hat{\Sigma}) = \Sigma$. Moreover, we have, by Cauchy-Schwarz inequality:

$$\begin{aligned}
\langle a, (f \otimes_{L_2(d\rho)} f)a\rangle_{L_2(d\rho)} &= \bigg(\int_{\mathcal{X}} a(x)f(x)d\rho(x)\bigg)^2 = \bigg(\int_{\mathcal{X}}\int_{\mathcal{V}} a(x)g(v)\varphi(v,x)d\tau(v)d\rho(x)\bigg)^2 \\
&\leqslant \bigg(\int_{\mathcal{V}} g(v)^2 d\tau(v)\bigg)\int_{\mathcal{V}}\bigg(\int_{\mathcal{X}} a(x)\varphi(v,x)d\rho(x)\bigg)^2 d\tau(v) \\
&= \|g\|^2_{L_2(d\rho)}\langle a, \Sigma a\rangle_{L_2(d\rho)} \leqslant \langle a, \Sigma a\rangle_{L_2(d\rho)}.
\end{aligned}$$

Thus $f \otimes_{L_2(d\rho)} f \preccurlyeq \Sigma$, and we may thus define $\langle f, \Sigma^{-1}f\rangle_{L_2(d\rho)}$, which is less than one.

Overall we aim to study $\langle f, (\hat{\Sigma}+\lambda I)^{-1}f\rangle_{L_2(d\rho)}$, for $\langle f, \Sigma^{-1}f\rangle_{L_2(d\rho)} \leqslant 1$, to control both the norm $\|\beta\|_2^2$ in Eq. (17) and the approximation error $\|f - \Phi\beta\|^2_{L_2(d\rho)}$ in Eq. (16). We have, following a similar argument than the one of Bach (2013); El Alaoui and Mahoney (2014) for column sampling, i.e., by a formulation using $\Sigma - \hat{\Sigma}$ in terms of operators in an appropriate way:

$$\begin{aligned}
&\langle f, (\hat{\Sigma}+\lambda I)^{-1}f\rangle_{L_2(d\rho)} \\
= \;& \langle f, (\Sigma+\lambda I+\hat{\Sigma}-\Sigma)^{-1}f\rangle_{L_2(d\rho)} \\
= \;& \big\langle (\Sigma+\lambda I)^{-1/2}f, \big[I + (\Sigma+\lambda I)^{-1/2}(\hat{\Sigma}-\Sigma)(\Sigma+\lambda I)^{-1/2}\big]^{-1}(\Sigma+\lambda I)^{-1/2}f\big\rangle_{L_2(d\rho)}.
\end{aligned}$$

Thus, if $(\Sigma+\lambda I)^{-1/2}(\hat{\Sigma}-\Sigma)(\Sigma+\lambda I)^{-1/2} \succcurlyeq -tI$, with $t \in (0,1)$, we have

$$\begin{aligned}
\langle f, (\hat{\Sigma}+\lambda I)^{-1}f\rangle_{L_2(d\rho)} &\leqslant \langle (\Sigma+\lambda I)^{-1/2}f, (1-t)^{-1}(\Sigma+\lambda I)^{-1/2}f\rangle_{L_2(d\rho)} \\
&= (1-t)^{-1}\langle f, (\Sigma+\lambda I)^{-1}f\rangle_{L_2(d\rho)} \\
&\leqslant (1-t)^{-1}\langle f, \Sigma^{-1}f\rangle_{L_2(d\rho)} \leqslant (1-t)^{-1}.
\end{aligned}$$

Moreover, we have shown $(\hat{\Sigma}+\lambda I)^{-1} \preccurlyeq \frac{1}{1-t}(\Sigma+\lambda I)^{-1}$.

Thus, the performance depends on having $(\Sigma+\lambda I)^{-1/2}(\Sigma-\hat{\Sigma})(\Sigma+\lambda I)^{-1/2} \preccurlyeq tI$.

We consider the self-adjoint operators X_i, for $i = 1, \dots, n$, which are independent and identically distributed:

$$X_i = \frac{1}{n}(\Sigma+\lambda I)^{-1}\Sigma - \frac{1}{n}\frac{1}{q(v_i)}\big[(\Sigma+\lambda I)^{-1/2}\varphi(v_i,\cdot)\big] \otimes_{L_2(d\rho)} \big[(\Sigma+\lambda I)^{-1/2}\varphi(v_i,\cdot)\big],$$

so that our goal is to provide an upperbound on the probability that $\|\sum_{i=1}^n X_i\|_{\rm op} > t$, where $\|\cdot\|_{\rm op}$ is the operator norm (largest singular values). We use the notation

$$d = \operatorname{tr}\Sigma(\Sigma+\lambda I)^{-1} = \int_{\mathcal{V}} \frac{\langle \varphi(v,\cdot), (\Sigma+\lambda I)^{-1}\varphi(v,\cdot)\rangle_{L_2(d\rho)}}{q(v)} q(v)d\tau(v) \leqslant d_{\max}.$$

We have

$$\begin{aligned}
\mathbb{E}X_i &= 0, \text{ by construction of } X_i,\\
X_i &\preccurlyeq \frac{1}{n}(\Sigma+\lambda I)^{-1}\Sigma \preccurlyeq \frac{1}{n}\operatorname{tr}\left[(\Sigma+\lambda I)^{-1}\Sigma\right] I \preccurlyeq \frac{d_{\max}}{n} I,\\
X_i &\succcurlyeq -\frac{1}{n}\frac{1}{q(v_i)}\left[(\Sigma+\lambda I)^{-1/2}\varphi(v_i,\cdot)\right] \otimes_{L_2(d\rho)} \left[(\Sigma+\lambda I)^{-1/2}\varphi(v_i,\cdot)\right]\\
&\succcurlyeq -\frac{1}{n}\frac{1}{q(v_i)}\left\|(\Sigma+\lambda I)^{-1/2}\varphi(v_i,\cdot)\right\|^2_{L_2(d\rho)} I \succcurlyeq -\frac{d_{\max}}{n} I,\\
\|X_i\|_{\mathrm{op}} &\leqslant \frac{d_{\max}}{n} \text{ as a consequence of the two previous inequalities,}\\
\mathbb{E}(X_i^2) &= \mathbb{E}\left[\frac{1}{n}\frac{1}{q(v_i)}\left[(\Sigma+\lambda I)^{-1/2}\varphi(v_i,\cdot)\right] \otimes_{L_2(d\rho)} \left[(\Sigma+\lambda I)^{-1/2}\varphi(v_i,\cdot)\right]\right]^2 - \left[\frac{1}{n}(\Sigma+\lambda I)^{-1}\Sigma\right]^2\\
&\preccurlyeq \mathbb{E}\left[\frac{1}{n}\frac{1}{q(v_i)}\left[(\Sigma+\lambda I)^{-1/2}\varphi(v_i,\cdot)\right] \otimes_{L_2(d\rho)} \left[(\Sigma+\lambda I)^{-1/2}\varphi(v_i,\cdot)\right]\right]^2\\
&= \mathbb{E}\frac{\langle\varphi(v_i,\cdot),(\Sigma+\lambda I)^{-1}\varphi(v_i,\cdot)\rangle_{L_2(d\rho)}}{n^2 q(v_i)^2}\left[(\Sigma+\lambda I)^{-1/2}\varphi(v_i,\cdot)\right] \otimes_{L_2(d\rho)} \left[(\Sigma+\lambda I)^{-1/2}\varphi(v_i,\cdot)\right]\\
&\preccurlyeq \frac{d_{\max}}{n^2}\mathbb{E}\left(\left[\frac{1}{q(v_i)}(\Sigma+\lambda I)^{-1/2}\varphi(v_i,\cdot)\right] \otimes_{L_2(d\rho)} \left[(\Sigma+\lambda I)^{-1/2}\varphi(v_i,\cdot)\right]\right) = \frac{d_{\max}}{n^2}\Sigma(\Sigma+\lambda I)^{-1},\\
\sum_{i=1}^n \mathbb{E}(X_i^2) &\preccurlyeq \frac{d_{\max}}{n}(\Sigma+\lambda I)^{-1}\Sigma,
\end{aligned}$$

with a maximal eigenvalue less than $\frac{d_{\max}}{n}$ and a trace less than $\frac{d_{\max}}{n}\operatorname{tr}\Sigma(\Sigma+\lambda I)^{-1} = \frac{d\, d_{\max}}{n}$.

Following Hsu et al. (2014), we use a matrix Bernstein inequality which is independent of the underlying dimension (which is here infinite). We consider the bound of Minsker (2011, Theorem 2.1), which improves on the earlier result of Hsu et al. (2012, Theorem 4), that is:

$$\mathbb{P}\left(\left\|\sum_{i=1}^n X_i\right\|_{\mathrm{op}} > t\right) \leqslant 2d\left(1+\frac{6}{t^2\log^2(1+nt/d_{\max})}\right)\exp\left(-\frac{t^2/2}{d_{\max}/n(1+t/3)}\right)$$

We now consider $t = \frac{3}{4}$, $\delta \in (0,1)$, and $n \geqslant B d_{\max}\log\frac{Cd_{\max}}{\delta}$, with appropriate constants $B, C > 0$. This implies that

$$\exp\left(-\frac{t^2/2}{d_{\max}/n(1+t/3)}\right) \leqslant \exp\left(-\frac{(3/4)^2/2}{5/4}B\log\frac{Cd_{\max}}{\delta}\right) \leqslant \left(\frac{\delta}{Cd_{\max}}\right)^{\frac{(3/4)^2 B/2}{5/4}} \leqslant \left(\frac{\delta}{Cd}\right)^{\frac{(3/4)^2 B/2}{5/4}},$$

and, if $d_{\max} \geqslant D$, using $n \geqslant B d_{\max}\log CD$,

$$1+\frac{6}{t^2\log^2(1+nt/d_{\max})} \leqslant 1+\frac{6\cdot 16/9}{\log^2\left(1+(3B/4)\log(CD)\right)},$$

while if $d_{\max} \leqslant D$ and $n \geqslant 1$,

$$1+\frac{6}{t^2\log^2(1+nt/d_{\max})} \leqslant 1+\frac{6\cdot 16/9}{\log^2\left(1+(3/4D)\right)}.$$

In order to get a bound, we need $\frac{(3/4)^2 B/2}{5/4} \geqslant 1$, and we can take $B = 5$. If we take $C = 8$, then in order to have $1 + \frac{6}{t^2 \log^2(1+nt/d_{\max})} \leqslant 4$, we can take $D = 3/8$. Thus the probability is less than δ.

Finally, in order to get the extra bound on $\frac{1}{n}\sum_{i=1}^n q(v_i)^{-1}\|\varphi(v_i,\cdot)\|^2_{L_2(d\rho)}$, we consider $\mathbb{E}\operatorname{tr}\hat{\Sigma} = \operatorname{tr}\Sigma = \int_{\mathcal{X}} k(x,x)d\rho(x)$, and thus, by Markov's inequality, with probability $1-\delta$,

$$\frac{1}{n}\sum_{i=1}^n q(v_i)^{-1}\|\varphi(v_i,\cdot)\|^2_{L_2(d\rho)} = \operatorname{tr}\hat{\Sigma} \leqslant \frac{1}{\delta}\operatorname{tr}\Sigma. \tag{18}$$

By taking $\delta/2$ instead of δ in the control of $\|\sum_{i=1}^n X_i\|_{\mathrm{op}} > t$ and in the Markov inequality above, we have a control over $\|\beta\|_2^2$, $\operatorname{tr}\hat{\Sigma}$ and the approximation error, which leads to the desired result in Prop 1. This will be useful for the lower bound of Prop. 3.

We can make the following extra observations regarding the proof:

- It may be possible to derive a similar result with a thresholding of eigenvalues in the spirit of Zwald et al. (2004), but this would require Bernstein-type concentration inequalities for the projections on principal subspaces.

- We have seen that with high-probability, we have $(\hat{\Sigma}+\lambda I)^{-1} \preccurlyeq 4(\Sigma+\lambda I)^{-1}$. Note that $A \preccurlyeq B$ does not imply in $A^2 \preccurlyeq B^2$ (Bhatia, 2009, page 9) and that in general we do not have $(\hat{\Sigma}+\lambda I)^{-2} \preccurlyeq C(\Sigma+\lambda I)^{-2}$ for any constant C (which would allow an improvement in the error by replacing λ by λ^2, and violate the lower bound of Prop. 3).

- We may also obtain a result in expectation, by using $\delta = 4\lambda/\operatorname{tr}\Sigma$ (which is assumed to be less than 1), leading to a squared error with expectation less than 8λ as soon as $n \geqslant 5 d_{\max}(\lambda)\log\frac{2(\operatorname{tr}\Sigma)d_{\max}(\lambda)}{\lambda}$. Indeed, we can use the bound 4λ with probability $1-\delta$ and $\|f\|^2_{L_2(d\rho)} \leqslant \operatorname{tr}\Sigma$ with probability δ, leading to a bound of $4\lambda(1-\delta)+\delta\operatorname{tr}\Sigma \leqslant 8\lambda$. We use this result in Section 4.5.

B.2 Proof of Prop. 2

We start from the bound above, with the constraint $n \geqslant 5d(\lambda)\log\frac{16d(\lambda)}{\delta}$. Statement (a) is a simple reformulation of Prop. 1. For statement (b), if we assume $m \leqslant \frac{n}{5(1+\gamma)\log\frac{16n}{5\delta}}$, and $\lambda = \mu_m$, then we have $d(\lambda) \leqslant (1+\gamma)m$, which implies $n \geqslant 5d(\lambda)\log\frac{16d(\lambda)}{\delta}$, and (b) is a consequence of (a).

B.3 Proof of Prop. 3

We first use the Varshamov-Gilbert's lemma (see, e.g., Massart, 2003, Lemma 4.7). That is, for any integer s, there exists a family $(\theta_j)_{j\in J}$ of at least $|J| \geqslant e^{s/8}$ distinct elements of $\{0,1\}^s$, such that for $j \neq j' \in J$, $\|\theta_j - \theta_{j'}\|_2^2 \geqslant \frac{s}{4}$.

For each $\theta \in \{0,1\}^s$, we define an element of $\mathcal{F}$ with norm less than one, as $f(\theta) = \frac{\sqrt{\mu_s}}{\sqrt{s}}\sum_{i=1}^s \theta_i e_i \in \mathcal{F}$, where (e_i,μ_i), $i=1,\dots,s$ are the eigenvector/eigenvalue pairs associated with the s largest

eigenvalues of Σ. We have, since $\mu_i \geqslant \mu_s$ for $i \in \{1, \dots, s\}$ and $\|\theta\|_2^2 \leqslant 1$:

$$\|f(\theta)\|_{\mathcal{F}}^2 = \frac{\mu_s}{s} \sum_{i=1}^{s} \theta_i^2 \mu_i^{-1} \leqslant \frac{\mu_s}{s} \sum_{i=1}^{s} \theta_i^2 \mu_s^{-1} \leqslant \frac{1}{s} \sum_{i=1}^{s} \theta_i^2 \leqslant 1.$$

Moreover, for any $j \neq j' \in J$, we have $\|f(\theta_j) - f(\theta_{j'})\|_{L_2(d\rho)}^2 = \frac{\mu_s}{s}\|\theta_j - \theta_{j'}\|_2^2 \geqslant \frac{\mu_s}{4}$.

We now assume that s is selected so that $\sqrt{4\lambda} \leqslant \sqrt{\frac{\mu_s}{4}}/3$. By applying the existence results to all functions f_j, $j \in J$, then there exists a family $(\beta_j)_{j \in J}$ of elements of $\mathbb{R}^n$, with squared ℓ_2-norm less than $\frac{4}{n}$, and for which, for all j,

$$\Big\| f_j - \sum_{i=1}^{n} (\beta_j)_i \psi_i \Big\|_{L_2(d\rho)} \leqslant \sqrt{4\lambda}.$$

This leads to, for any $j \neq j' \in J$,

$$\begin{aligned}\Big\| \sum_{i=1}^{n} (\beta_j - \beta_{j'})_i \psi_i \Big\|_{L_2(d\rho)} &\geqslant \big\| f_j - f_{j'} \big\|_{L_2(d\rho)} - \Big\| \sum_{i=1}^{n} (\beta_j)_i \psi_i - f_j \Big\|_{L_2(d\rho)} - \Big\| \sum_{i=1}^{n} (\beta_{j'})_i \psi_i - f_{j'} \Big\|_{L_2(d\rho)} \\ &\geqslant \sqrt{\mu_s/4} - 2\sqrt{\frac{\mu_s}{4}}/3 = \sqrt{\frac{\mu_s}{4}}/3.\end{aligned}$$

Moreover, we have the bound

$$\Big\| \sum_{i=1}^{n} (\beta_j - \beta_{j'})_i \psi_i \Big\|_{L_2(d\rho)}^2 \leqslant \Big(\sum_{i=1}^{n} (\beta_j - \beta_{j'})_i^2 \Big) \sum_{i=1}^{n} \|\psi_i\|_{L_2(d\rho)}^2 \leqslant \|\beta_j - \beta_{j'}\|_2^2 \cdot n(2\delta^{-1} \operatorname{tr} \Sigma).$$

Combining the last two inequalities, we get $\|\beta_j - \beta_{j'}\|_2 \geqslant \sqrt{\frac{\delta \mu_s}{72 n \operatorname{tr} \Sigma}} = \Delta$. Thus, $e^{s/8}$ is less than the Δ-packing number of the ball of radius $r = 2/\sqrt{n}$, which is itself less than $(r/\Delta)^n (2 + \Delta/r)^n$ (see, e.g., Massart, 2003, Lemma 4.14). Since $\Delta/r = \sqrt{\frac{\delta \mu_s}{4 \cdot 72 \operatorname{tr} \Sigma}} \leqslant \frac{1}{12\sqrt{2}}$, we have

$$\frac{s}{8} \leqslant n \left(\frac{1}{2} \log \frac{4 \cdot 72 \operatorname{tr} \Sigma}{\delta \mu_s} + \log(2 + \frac{1}{12\sqrt{2}}) \right).$$

This implies $n \geqslant \frac{s}{4 \log \frac{\operatorname{tr} \Sigma}{\delta \mu_s} + 29}$. Given that we have to choose $\mu_s \geqslant 144\lambda$ for the result to hold, this implies the desired result, since $4 \log(1440) \geqslant 29$.