On the Equivalence between Quadrature Rules and Random Features

Francis Bach

On the Equivalence between Quadrature Rules and Random Features

Francis Bach
INRIA - Sierra project-team
Département d'Informatique de l'Ecole Normale Supérieure
Paris, France
francis.bach@ens.fr

February 18, 2015

Abstract

We show that kernel-based quadrature rules for computing integrals are 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 known 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. The 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 elements, 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 complexity of these methods.

In this paper, we make the following contributions:

- We show in Section 3.2 that these two problems are equivalent; more precisely, kernel-based quadrature rules are a special case 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 known non-uniform distribution, while the lower bound if valid for any set of points.
- Applying our results to kernel quadrature, while our results is 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 feature expansions, we show 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]. 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 properties (see, e.g., Berlinet and Thomas-Agnan, 2004): (a) for any $x \in \mathcal{X}$, the function $k(\cdot, x) : y \mapsto k(y, x)$ is an element of $\mathcal{F}$, and (b) we have the *reproducing property*, that is, for all $f \in \mathcal{F}$ and $x \in \mathcal{X}$, $f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{F}}$. We always assume that $\mathcal{F}$ is infinite-dimensional.

Moreover, 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.

Integral operator. Let $\Sigma : L_2(d\rho) \to L_2(d\rho)$ be defined as $(\Sigma f)(x) = \displaystyle\int_{\mathcal{X}} f(x)k(x, y)d\rho(y)$. Since $\int_{\mathcal{X}} k(x, x)d\rho(x)$ is finite, Σ is self-adjoint, positive semi-definite and trace-class. Moreover, $\Sigma^{1/2}$ is a surjection from $L_2(d\rho)$ to $\mathcal{F}$; more precisely, for any $f \in \mathcal{F}$, there exists an unique $g \in (\operatorname{Ker}\Sigma)^{\perp} \subset L_2(d\rho)$ 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 $f \in \mathcal{F}$.

Mercer decomposition. From extensions of Mercer's theorem (König, 1986), there exists an orthonormal sequence $(e_m)_{m\geqslant 0}$ of $L_2(d\rho)$ and a summable non-increasing sequence of strictly positive eigenvalues $(\mu_m)_{m\geqslant 0}$ such that $\Sigma e_m = \mu_m e_m$. For simplicity we assume that there are no zero eigenvalues, i.e., $(e_m)_{m\geqslant 0}$ is an orthonormal *basis*, which implies that $\mathcal{F}$ is dense in $L_2(d\rho)$.

We have the decomposition $k(x, y) = \sum_{m\geqslant 0} \mu_m e_m(x)e_m(y)$. For each $m \geqslant 0$, the eigenfunction e_m is an element of $\mathcal{F}$ and $\|e_m\|_{\mathcal{F}} = \mu_m^{-1/2}$; moreover, $(\mu_m^{1/2}e_m)_{m\geqslant 0}$ is an orthonormal basis of $\mathcal{F}$. This justifies the view of $\mathcal{F}$ as the subspace of functions $f \in L_2(d\rho)$ such that $\|\Sigma^{-1/2}f\|^2_{L_2(d\rho)} = \sum_{m\geqslant 0} \mu_m^{-1}\langle f, e_m\rangle^2_{L_2(d\rho)}$ is finite.

Potential confusion with covariance operator. Note that the operator Σ is an operator on $L_2(d\rho)$. It should not ne confused with the (non-centered) covariance operator C, which is an autoadjoint operator on 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.

2.1 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 $d\mu$. We consider a function $\varphi : \mathcal{V} \times \mathcal{X} \to \mathbb{R}$ which

[1] For simplicity we identify functions and their equivalence classes for the equivalence relationship of being equal except for a zero-measure (for $d\rho$) subset of $\mathcal{X}$.

is square-integrable (for the measure $d\mu \otimes d\rho$), and assume that the kernel k may be written as, for all $x, y \in \mathcal{X}$:

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

In other words, for each $v \in \mathcal{V}$, we have a one-dimensional feature $\varphi(v,\cdot)$ which is a function from $\mathcal{X}$ to $\mathbb{R}$, and the kernel between x and y is simply the expectation of the dot-product for this one-dimensional feature for v following the probability distribution $d\mu$.

Such additional structure allows to give an explicit characterization of the RKHS $\mathcal{F}$ in terms of the features φ. Indeed, as shown by Bach (2014, App. A), a function $f \in L_2(d\rho)$ is in $\mathcal{F}$ if and only if it may be written as $\forall x \in \mathcal{X},\ f(x) = \int_{\mathcal{V}} g(v)\varphi(v,x)d\mu(v) = \langle g, \varphi(\cdot,x)\rangle_{L_2(d\mu)}$, for a certain function $g : \mathcal{V} \to \mathbb{R}$ such that $\|g\|^2_{L_2(d\mu)}$ is finite, with a norm $\|f\|^2_{\mathcal{F}}$ equal to the minimum (which is always attained) of $\|g\|^2_{L_2(d\mu)}$, over all possible decompositions of f.

2.2 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. From $k(x,y) = \sum_{m \geqslant 0} \frac{\mu_m}{\operatorname{tr}\Sigma}\Big[\big(\operatorname{tr}\Sigma\big)^{1/2} e_m(x)\Big] \cdot \Big[\big(\operatorname{tr}\Sigma\big)^{1/2} e_m(x)\Big]$, we obtain an expectation with $\mathcal{V} = \mathbb{N}$. In Section 3.2, we provide another generic decomposition with $\mathcal{V} = \mathcal{X}$.

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, that is, $k(x,y)$ may be written as $k(x,y) = \sum_{m=0}^{\infty} \mu_m \cos 2\pi m(x-y) = \sum_{m=0}^{\infty} \mu_m \big[\cos 2\pi mx \cos 2\pi my + \sin 2\pi mx \sin 2\pi my\big]$, with $\mu_m \geqslant 0$, which can be put trivially as an expectation with $\mathcal{V} = \mathbb{Z}$. These are the usual Fourier features (Rahimi and Recht, 2007). Moreover, the random Fourier features correspond to the Mercer decomposition above for the uniform distribution on $[0,1]$.

Among these, the sequence $\mu_0 = 1$ and $\mu_m = \frac{1}{m^{2s}}$ leads to $k(x,y) = 1 + \frac{(-1)^{s-1}(2\pi)^{2s}}{2(2s)!} B_{2s}(\{x-y\})$ where $\{x-y\}$ denotes the fractional part of $x-y$ (Wahba, 1990), and we recover the traditional Sobolev space with parameter $s \geqslant 1$ (Adams and Fournier, 2003). For $\mu_n = r^n$, we have $k(x,y) = \frac{1-r\cos 2\pi(x-y)}{1-2r\cos 2\pi(x-y)+r^2}$, with a geometric decay of eigenvalues.

Translation invariant kernels on $\mathbb{R}$. We consider $\mathcal{X} = \mathbb{R}$ and translation-invariant kernels $k(x,y)$ of the form $k(x,y) = t(x-y)$, where t is an integrable function from $\mathbb{R}$ 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 $\hat{t}(\omega) = \int_{\mathbb{R}} t(x) e^{-i\omega x} dx \in \mathbb{R}_+$, then $k(x,y) = \frac{1}{2\pi} \int_{\mathbb{R}} \hat{t}(\omega) e^{i\omega(x-y)} d\omega = \frac{1}{2\pi} \int_{\mathbb{R}} \hat{t}(\omega) \big[\cos \omega x \cos \omega y + \sin \omega x \sin \omega y \big] d\omega$, which is another form of random Fourier features with $\mathcal{V} = \mathbb{R}$ and distribution with density $\hat{t}(\omega)$ (this can be seen by splitting $\mathbb{R}$ into a positive part for the cosines and a negative part for the sines). For these kernels, the decay of eigenvalues has been well-studied by Widom (1963), who relate 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 on sub-Gaussian distributions, the decay of eigenvalues is geometric, and for kernels leading to Sobolev spaces, the decay is of the form m^{-2s}. See also examples by Birman and Solomyak (1977).

Extensions to $[0,1]^d$ and $\mathbb{R}^d$. In order to extend to $d > 1$, we may consider several extensions as described by Oates and Girolami (2015). We may first use the tensor product of the d individual Hilbert spaces, for which the kernel is simply the product of individual kernels. If each of the kernel leads to eigenvalues that are decaying as m^{-2s}, then the resulting eigenvalues decay as $(\log m)^{d-1} m^{-2s}$ (see Appendix A), and thus up to logarithmic terms at the same speed as $d = 1$. For Sobolev spaces of order s for each dimension, this corresponds to functions which have square-integrable partial derivatives with all *individual* orders less than s. This is to be contrasted with the usual multi-dimensional Sobolev space which is composed of functions which have square-integrable partial derivatives with orders with *sum* less than s. This last kernel leads to eigenvalues decaying as $m^{-2s/d}$, which is much slower (see Appendix A). Note that in terms of computation, there are extensions to avoid linear complexity in d (Le et al., 2013).

2.3 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 $d\mu$ and define the kernel approximation $\hat{k}(x,y) = \frac{1}{n} \sum_{i=1}^n \varphi(v_i, x) \varphi(v_i, y)$, 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 $1/\sqrt{n}$.

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.

Computation of error. Given the definition of the Hilbert space $\mathcal{F}$ in terms of φ, given $g \in L_2(d\mu)$ with $\|g\|_{L_2(d\mu)} \leqslant 1$, we aim at finding $\alpha \in \mathbb{R}^n$ such that $\hat{f} = \sum_{i=1}^n \alpha_i \varphi(v_i, \cdot)$ with norm $\|\hat{f}\|_{\hat{\mathcal{F}}}^2 \leqslant n\|\alpha\|_2^2 \leqslant 1$ as small as possible and with a small approximation error:

$$\|\hat{f} - f\|_{L_2(d\rho)} = \left\| \sum_{i=1}^n \alpha_i \varphi(v_i, \cdot) - \int_{\mathcal{V}} g(v)\varphi(v, \cdot) d\mu(v) \right\|_{L_2(d\rho)}. \tag{2}$$

Note that with $\alpha_i = \frac{1}{n} g(v_i)$ and v_i sampled from $d\mu$, then, $\mathbb{E}(\|\alpha\|_2^2) \leqslant \frac{1}{n}$ and $\mathbb{E}(\|f - \hat{f}\|_{L_2(d\rho)}^2) \leqslant \frac{1}{n} \sup_{v \in \mathcal{V}} \|\varphi(v, \cdot)\|_{L_2(d\rho)}^2$; our goal is to obtain an error rate with a better scaling in n, by (a) choosing a better distribution than $d\mu$ for the points $v_1, \ldots, v_n$ and (b) by finding the best possible weights $\alpha \in \mathbb{R}^n$ (that may depend on the function g but not on f).

Goals. We thus aim at sampling n points $v_1, \ldots, v_n \in \mathcal{V}$ from a distribution with density q with respect to $d\mu$. 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 to minimize $\left\| \sum_{i=1}^n \frac{\beta_i}{q(v_i)^{1/2}} \varphi(v_i, \cdot) - \int_{\mathcal{V}} g(v)\varphi(v, \cdot) d\mu(v) \right\|_{L_2(d\rho)}$, with $n\|\beta\|_2^2$ (which represents the norm of the approximation in $\hat{\mathcal{F}}$) as small as possible.

3 Quadrature in RKHSs

Given a square-integrable (with respect to $d\rho$) function $g : \mathcal{X} \to \mathbb{R}$, we aim 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 $h(x_1), \ldots, h(x_n)$ of the function h at well-chosen points $x_1, \ldots, x_n \in \mathcal{X}$. 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.

3.1 Approximation of the mean element

Using the properties of RKHSs, the error is

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

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

$$\left\| \sum_{i=1}^n \alpha_i k(\cdot, x_i) - \int_{\mathcal{X}} k(\cdot, x)g(x)d\rho(x) \right\|_{\mathcal{F}}. \tag{3}$$

The goal of quadrature rules formulated in a RKHS is thus to find points $x_1, \ldots, x_n \in \mathcal{X}$ and weights $\alpha \in \mathbb{R}^n$ so that the quantity in Eq. (3) is as small as possible (Smola et al.,

2007). For $g = 1$, the function $\int_{\mathcal{X}} k(\cdot, x) d\rho(x)$ is 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 less than $\frac{1}{n} \sup_{x \in \mathcal{X}} k(x, x)$ (Smola et al., 2007). Note that when $g = 1$, Eq. (3) corresponds to 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 positive and sum to a fixed constant (so that constant functions are well integrated). We will not pursue this here as our theoretical results do not accommodate such constraints.

Tolerance to noisy function values. In practice, independent (but not necessarily identically distributed) noise 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 $\big\| \sum_{i=1}^n \alpha_i k(\cdot, x_i) - \int_{\mathcal{X}} k(\cdot, x) g(x) d\rho(x) \big\|_{\mathcal{F}}^2 + \sum_{i=1}^n \alpha_i^2 \sigma^2(x_i)$, 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 $x \in \mathcal{X}$, we denote by $\psi(\cdot, x)$ the unique element of $(\operatorname{Ker} \Sigma)^\perp \subset L_2(d\rho)$ such that $\Sigma^{1/2} \psi(\cdot, x) = k(\cdot, x)$. Given the Mercer decomposition $k(x, y) = \sum_{m \geqslant 0} \mu_m e_m(x) e_m(y)$, we have the expansion $\psi(x, y) = \sum_{m \geqslant 0} \mu_m^{1/2} e_m(x) e_m(y)$ (with convergence in the L_2-norm for the measure $d\rho \otimes d\rho$), and thus we may consider ψ as a symmetric function. Note that ψ many not be easy to compute in many practical cases (except for periodic kernels on $[0, 1]$).

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

$$k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{F}} = \langle \psi(\cdot, x), \psi(\cdot, y) \rangle_{L_2(d\rho)} = \int_{\mathcal{X}} \psi(v, x) \psi(v, y) d\rho(v).$$

That is, k may always be written as an expectation. Moreover, we have

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

which is exactly an instance of the approximation result in Eq. (2) with $\mathcal{V} = \mathcal{X}$ and $\varphi = \psi$. Thus, the quadrature problem is a subcase of the random feature problem for a specific expansion. Note that this random decomposition is always possible (although not in closed form in general).

Goals. In order to be able to make the parallel with random feature approximation, we consider importance-weighted coefficients $\beta_i = \alpha_i q(x_i)^{1/2}$, and we thus aim at minimizing the approximation 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 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 is as good as optimal quadrature schemes (see Appendix B); 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).

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.

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. 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 2.3 (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 4.2. We then consider consequences of these results on quadrature (Section 4.3) and random feature expansions (Section 4.4).

4.1 Upper bound

We have the following proposition (see proof in Appendix C.1):

Proposition 1 *For $\lambda > 0$, we denote by $d_{\max}(q,\lambda) = \sup_{v\in\mathcal{V}} \frac{1}{q(v)} \langle \varphi(v,\cdot), (\Sigma+\lambda I)^{-1}\varphi(v,\cdot)\rangle_{L_2(d\rho)}$. Let $v_1,\dots,v_n$ be sampled i.i.d. from the distribution with positive density q with respect to $d\mu$, then for any $\delta > 0$, if $n \geqslant 4 + 6 d_{\max}(q,\lambda)\log\frac{4 d_{\max}(q,\lambda)}{\delta}$, with probability greater than $1-\delta$, we have*

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

The proof technique relies on computing an explicit candidate $\beta \in \mathbb{R}^n$ obtained from minimizing a regularized least-squares formulation, leading to an approximation of f equal to $\hat{f} = (\hat{\Sigma}+\lambda I)^{-1}\hat{\Sigma} f$, where $\hat{\Sigma}$ is a properly defined empirical integral operator and $\lambda > 0$. 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.

Optimized distribution. We may now consider a specific distribution, namely

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

for which $d_{\max}(q,\lambda) = d(\lambda) = \operatorname{tr}\Sigma(\Sigma+\lambda I)^{-1}$. We thus need to have $n \geqslant 4 + 6 d(\lambda)\log\frac{4d(\lambda)}{\delta}$ 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), which is always smaller $d_{\max}(1,\lambda)$ 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.

Eigenvalues and degrees of freedom. In order to relate more directly to the eigenvalues of Σ, we notice that $d(\lambda) = \operatorname{tr}\Sigma(\Sigma+\lambda I)^{-1} = \sum_{m\geqslant 0} \frac{\mu_m}{\mu_m+\lambda} \geqslant \sum_{\mu_m\geqslant\lambda} \frac{\mu_m}{\mu_m+\lambda} \geqslant \frac{1}{2}\max(\{m,\ \mu_m \geqslant \lambda\})$. Moreover, $d(\lambda) = \sum_{\mu_m\geqslant\lambda} \frac{\mu_m}{\mu_m+\lambda} + \sum_{\mu_m<\lambda} \frac{\mu_m}{\mu_m+\lambda} \leqslant \max(\{m,\ \mu_m \geqslant \lambda\}) + \frac{1}{\lambda}\sum_{\mu_m<\lambda}\mu_m$.

We now make the assumption that there exists $\gamma > 0$ such that

$$\forall j \geqslant 1,\quad \sum_{m=j}^{\infty} \mu_m \leqslant \gamma j \mu_j. \tag{4}$$

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 also for all examples in Section 2.2. It allows us to relate the degrees of freedom directly to eigenvalue decays.

Indeed, this implies that $\frac{1}{\lambda}\sum_{\mu_m<\lambda}\mu_m \leqslant \gamma\max(\{m,\ \mu_m\geqslant\lambda\})$ for all $\lambda < \mu_0$ and thus

$$\tfrac{1}{2}\max(\{m,\ \mu_m \geqslant \lambda\}) \leqslant d \leqslant \big[1+\gamma\big]\max(\{m,\ \mu_m \geqslant \lambda\}).$$

From Prop. 1, we thus need to have, up to logarithmic terms, $n \geqslant \max(\{m,\ \mu_m \geqslant \lambda\})$ random samples. For example, for polynomial decays of eigenvalues of the form $\mu_m = O(m^{-2s})$, we get errors proportional to n^{-s} for n samples, while for geometric decays, we get geometric errors.

4.2 Lower bound

We have the following lower bound (see proof in Appendix C.3):

Proposition 2 *For $\delta \in (0,1)$, if we have a family $\psi_1, \dots, \psi_n \in L_2(d\rho)$ such that*

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

then $n \geqslant \dfrac{\max(\{m,\ \mu_m \geqslant 144\lambda\})}{4\log\frac{2\operatorname{tr}\Sigma}{\lambda\delta}}$.

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 (μ_m) of the integral operator Σ.
- In order to obtain such a bound, we need to constrain either $\|\beta\|_2$ or the norms of the vectors ψ_i, which corresponds 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, in particular the bound shown in Eq. (7) in the proof of Prop. 1 in Appendix C.1, for which the parameter δ ends up entering the lower bound logarithmically.

4.3 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). All detailed computations are given in Appendix C.2.

If the points $x_1, \dots, x_n$ are sampled from the distribution with density q with respect to $d\rho$, then the quadrature rule becomes:

$$\sum_{i=1}^n \frac{\beta_i h(x_i)}{q(x_i)^{1/2}} = \big\langle g, \Sigma^{1/2}\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}\Sigma^{-1/2}h\big\rangle_{L_2(d\rho)},$$

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).

Finally, the density q may be expressed as $q(x) = \sum_{m \geqslant 0} \frac{\mu_m}{\mu_m + \lambda} e_m(x)^2$. A weakness of our result is that our optimized distribution $q(x) = q_\lambda(x)$ 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 typically q_λ converges to a certain distribution when λ tends to zero (see an example in Appendix B).

Recovering lower and upper bounds. For Sobolev spaces with parameters α in $[0,1]^d$ (for which we assume $d < 2\alpha$), the decay of eigenvalues is of the form $m^{-2\alpha/d}$ and thus the error after n samples is $n^{-\alpha/d}$, which recovers the upper and lower bounds of Novak (1988, pages 37 and 38).

Algorithms. The quadrature weights α may be obtained by minimizing $\| \sum_{i=1}^n \alpha_i k(\cdot, x_i) - \int_{\mathcal{X}} g(x) k(\cdot, x) d\rho(x) \|_{\mathcal{F}}^2$, that is, minimizing $\frac{1}{2}\alpha^\top K \alpha - \mu^\top \alpha$, where K is the kernel matrix and $\mu \in \mathbb{R}^n$ the vector such that $\mu_i = \int_{\mathcal{X}} g(x) k(x_i, x) d\rho(x)$. Given these, 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), we need to invert a linear system. 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 $O(n^3)$.

4.4 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 $\ell(y, \cdot)$. 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 with marginal distribution $d\rho$. 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):

$$\mathbb{E}\big[L(\hat{f})\big] \leqslant \inf_{\|f\|_{\mathcal{F}} \leqslant F} L(f) + 2\mathbb{E}\Big[\sup_{\|f\|_{\mathcal{F}} \leqslant 1} |L(f) - \hat{L}(f)|\Big] \leqslant \inf_{\|f\|_{\mathcal{F}} \leqslant F} L(f) + \frac{4FGR}{\sqrt{m}}. \qquad (5)$$

We now consider learning by sampling n features, leading to a function parameterized by $\beta \in \mathbb{R}^n$, that is $\hat{g}_\beta = \sum_{i=1}^n \beta_i q(v_i)^{-1/2} \varphi(v_i, \cdot) \in L_2(d\rho)$. We assume that n is large enough to have an expected squared error 8λ as in the end of the proof of Prop. 1 (Appendix C.1), that is, $n \geqslant 4 + 6d(\lambda)\log\frac{R^2 d(\lambda)}{\lambda}$ (if we consider the optimized distribution q). We consider a minimizer $\hat{\beta}$ of $\hat{L}(\hat{g}_\beta)$ over $\|\beta\|_2 \leqslant 2F/\sqrt{n}$. We obtain the following upper-bound on

$\mathbb{E}\big[L(\hat{g}_{\hat{\beta}})\big]$ (for simplicity in expectation with respect to the data and the random features) for the learned function $\hat{g}_{\hat{\beta}}$ based on random features:

$$\mathbb{E}\big[L(\hat{g}_{\hat{\beta}})\big] \leqslant \inf_{\|f\|_{\mathcal{F}} \leqslant F} L(f) + \mathbb{E} \sup_{\|f\|_{\mathcal{F}} \leqslant F} \inf_{\|\beta\|_2 \leqslant 2F/\sqrt{n}} |L(f) - L(g_\beta)| + 2\mathbb{E} \sup_{\|\beta\|_{\mathcal{F}} \leqslant 2F/\sqrt{n}} |L(\hat{g}_\beta) - \hat{L}(\hat{g}_\beta)|.$$

Because of the G-Lipschitz-continuity of the loss, the second term is less than $GF\sqrt{8\lambda}$. Following standard results for Rademacher complexities of ℓ_2-balls (Bartlett and Mendelson, 2003, Lemma 22), the third term is less than $\frac{4FG}{m\sqrt{n}}\mathbb{E}\big(\sum_{i=1}^m \sum_{j=1}^n \frac{\varphi(v_i,x_j)^2}{q(v_i)}\big)^{1/2} \leqslant \frac{4FG}{m\sqrt{n}}(nm \operatorname{tr}\Sigma)^{1/2} = \frac{4FGR}{\sqrt{m}}$. Overall, we obtain

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

The bound thus requires that we have $\lambda = R^2/m$ and thus $n \geqslant 4 + 6d(R^2/m)\log\big[md(R^2/m)\big]$ in order to lose only a constant factor compared to Eq. (5).

In the worst case, we have $d(\lambda) \leqslant \lambda^{-1}\operatorname{tr}\Sigma = R^2/\lambda$, and thus $n \geqslant 4 + 6m\log m$, and we lose a logarithmic factor compared to Rahimi and Recht (2009). However, when we have eigenvalue decays as $R^2 i^{-2s}$, we get (up to constants) $d(\lambda) \leqslant (R^2/\lambda)^{1/(2s)}$, 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 4.3, 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 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 $\|h\|_{\mathcal{F}} \leqslant 1$.

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).

Robustness to functions not in the RKHS. If the function h happens to be smoother than elements of the RKHS $\mathcal{F}$, that is, if $\|\Sigma^{-s}h\|_{L_2(d\rho)} \leqslant 1$, where $s \geqslant 1/2$, then we have the error, w.h.p.,

$$\begin{aligned}
\|\hat{h} - h\|_{L_2(d\rho)} &= \lambda\big\|\Sigma^{1/2}(\hat{\Sigma} + \lambda I)^{-1}\Sigma^{-1/2+s}\Sigma^{-s}h\big\|_{L_2(d\rho)} \\
&\leqslant \lambda\big\|\Sigma^{1/2}(\hat{\Sigma} + \lambda I)^{-1/2}\big\|_{\mathrm{op}}\big\|(\hat{\Sigma} + \lambda I)^{-1/2}\Sigma^{-1/2+s}\big\|_{\mathrm{op}}\|\Sigma^{-s}h\|_{L_2(d\rho)} \\
&\leqslant \lambda \cdot 2 \cdot \lambda^{s-1}\big\|(\hat{\Sigma} + \lambda I)^{1/2-s}\Sigma^{-1/2+s}\big\|_{\mathrm{op}} \leqslant 4\lambda^s.
\end{aligned}$$

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. 1 and 2, 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 Appendix B), 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).

Other 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 $r \in [0, 1/2]$. We have:

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

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 $t > 1$,

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

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

6 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.

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). 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.

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.

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.

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.

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

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.

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.

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

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.

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.

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.

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.

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.

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.

A Eigenvalues for tensor products

Given a kernel k on the set $\mathcal{X}$, with eigenvalues $(\mu_m)_{m \geqslant 0}$ with respect to the measure $d\rho$, there are mainly three ways to define a kernel K on $X = (x_1, \dots, x_d) \in \mathcal{X}^d$ with the product probability distribution $(d\rho)^{\otimes d}$:

- **Sum**: $K(X,Y) = \sum_{j=1}^d k(x_j, y_j)$. Its eigenvalues are μ_m, $m \geqslant 0$, each with multiplicity d. Thus, if $\mu_m \propto m^{-2s}$, we have a number of eigenvalues of K greater than λ equivalent to $d\lambda^{-1/(2s)}$.
- **Product**: $K(X,Y) = \prod_{j=1}^d k(x_j, y_j)$. Its eigenvalues are $\mu_{m_1} \times \cdots \times \mu_{m_d}$, $m_1, \dots, m_d \geqslant 0$. Thus, if $\mu_m \propto m^{-2s}$, we have a number of eigenvalues of K greater than λ equivalent to the number of multi-indices such that $m_1 \times \cdots \times m_d$ less than $\lambda^{-1/(2s)}$, which can easily be upper bounded by $\lambda^{-1/(2s)}(\log \lambda^{-1})^{d-1}$.
- **Harmonic means of eigenvalues**. If $\sum_{m \geqslant 0} \mu_m^{1/d}$ is finite, we may define

$$K(X,Y) = \sum_{m_1, \dots, m_d \geqslant 0} \frac{d}{\sum_{j=1}^d \mu_{m_j}^{-1}} \prod_{j=1}^d e_{m_j}(x_j) e_{m_j}(y_j),$$

  with eigenvalues $\frac{d}{\sum_{j=1}^d \mu_{m_j}^{-1}} \leqslant \big(\prod_{j=1}^d \mu_{m_j}\big)^{1/d}$ (and hence summable). For $\mathcal{F}$ a Sobolev space in 1 dimension, we obtain the regular Sobolev space in d dimensions. That is, for $2s > d$, if $\mu_m \propto m^{-2s}$, we have a number of eigenvalues of K greater than λ equivalent to the number of multi-indices such that $\|m\|_{2s} \leqslant (\lambda/d)^{-1/(2s)}$, which can easily be upper bounded by a constant times $\lambda^{-d/(2s)}$.

B Simulations

In this section, we consider simple illustrative quadrature experiments 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 $a = b$.

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 weights α. We can make the following observations:

- The exponents in the convergence rates for $s = t$ (matching RKHSs) are close to $2s$ as expected.
- When the functions to integrate are less smooth than the ones used for learning (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 $t \geqslant s/2$.

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 $s = 1$, all squared errors decay as n^{-2} with a worse constant for our kernel-based rule (which is the only one with such a general applicability), while for $s = 2$, the Sobol sequence is not adaptive.

Non-uniform distribution. In order to compute the weights and the error from points $x_1, \ldots, x_n$, we need to compute for all $y \in [0, 1]$, $\int_0^1 k(x, y)p(x)dx$. In order to compute the optimal density $q(x)$ from Eq. (8), we may take a set of $y_1, \ldots, y_m$ uniformly spread in $[0, 1]$, for m large, and compute

$$\begin{aligned} \langle k(\cdot, x), \Sigma^{-1} k(\cdot, y_i)\rangle_{L_2(d\rho)} &= k(x, y_i) \\ \langle k(\cdot, y_i), k(\cdot, y_j)\rangle_{L_2(d\rho)} &= \int_0^1 k(x, y_j)k(x, y_i)d\rho(x), \end{aligned}$$

to estimate coordinates of the approximation of the operator Σ^{-1} on the large independent set $\{k(\cdot, y_1), \ldots, k(\cdot, y_m)\}$.

All of these are computed by Gauss-Chebyshef quadrature with 10000 points. In Figure 3, we compare several distributions $q(x)$. The distribution $a = b = .25$ happens to be the limit of q_λ as λ tends to zero. All distributions achieve the optimal rate $O(n^{-2})$ except the one with $a = b = 2$, illustrating the fact that wrong distributions may have an adverse impact.

C Proofs

C.1 Proof of Prop. 1

Any $f \in \mathcal{F}$ with $\mathcal{F}$-norm less than one, may be represented as $f = \int_{\mathcal{V}} g(v)\varphi(v, \cdot)d\mu(v)$, for a certain $g \in L_2(d\mu)$ with $L_2(d\mu)$-norm less than one. We do not solve the problem

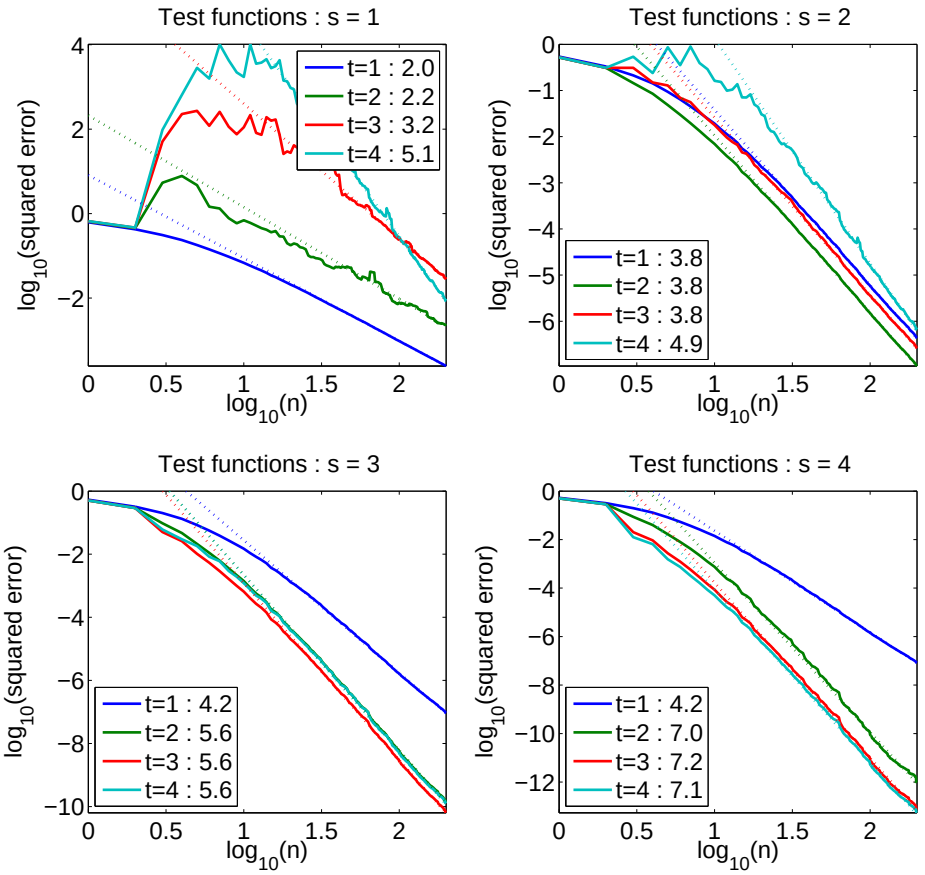

Figure 1: Quadrature for functions in a Sobolev space with parameter s (four possible values) for the uniform distribution on $[0, 1]$, with quadrature rules obtained from different Sobolev spaces with parameters t (same four possible values). We compute affine fits in log-log-space to estimate convergence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

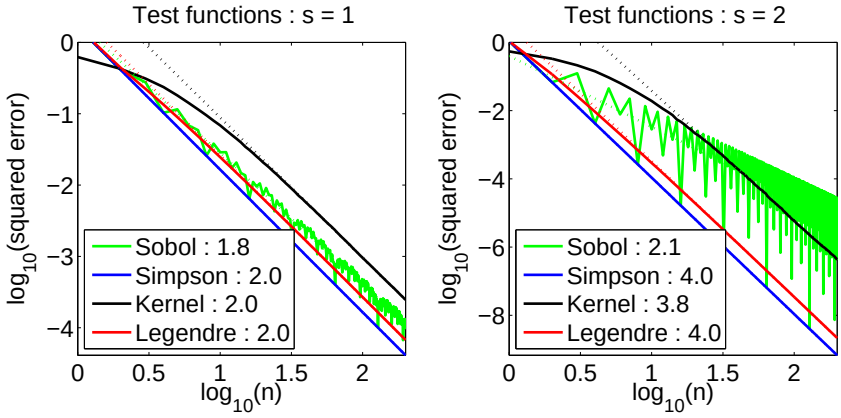

Figure 2: Quadrature for functions in a Sobolev space with parameters $s = 1$ (left) and $s = 2$ (right), for the uniform distribution on $[0, 1]$, with various quadrature rules. We compute affine fits in log-log-space to estimate convergence rates of the form C/n^u and report the value of u. Best seen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

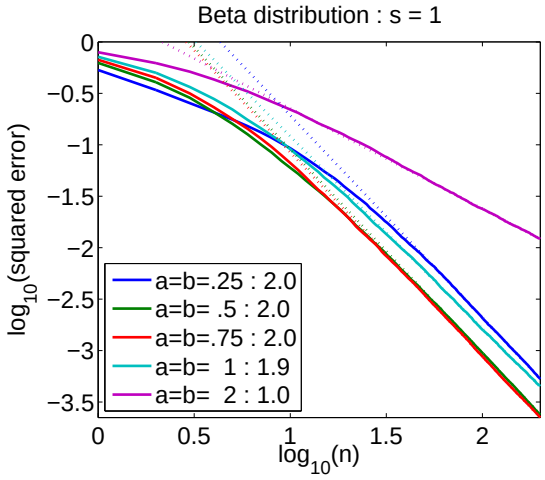

Figure 3: Quadrature for functions in a Sobolev space with parameter $s = 1$ for the Beta distribution on $[0, 1]$ with parameters $a = b = 1/2$, with quadrature rules obtained from sampling from different Beta distributions with equal parameters $a = b$. We compute affine fits in log-log-space to estimate convergence rates of the form C/n^u and report the value of u. Best seen in color.

in β exactly, but use a properly chosen Lagrange multiplier λ and consider the following minimization problem:

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

By introducing the notation Φ which is a matrix with infinitely many rows and i-th column equal to $q(v_i)^{-1/2}\varphi(v_i, \cdot) \in L_2(d\rho)$, we 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:

$$\beta = (\Phi^\top \Phi + n\lambda I)^{-1} \Phi^\top h = \frac{1}{n} \Phi^\top (\frac{1}{n} \Phi\Phi^\top + \lambda I)^{-1} f. \tag{6}$$

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

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

that is, for $a, b \in L_2(d\rho)$, $\langle a, \hat{\Sigma} b \rangle_{L_2(d\rho)} = \displaystyle\sum_{i=1}^{n} \frac{\langle a, \varphi(v_i, \cdot)\rangle_{L_2(d\rho)} \langle b, \varphi(v_i, \cdot)\rangle_{L_2(d\rho)}}{q(v_i)}$.

This operator $\hat{\Sigma}$ allows us to make more formal the heuristic computations above.

The value of $\|f - \Phi\beta\|^2_{L_2(d\rho)}$ is equal to $\lambda^2 \big\langle f, (\hat{\Sigma} + \lambda I)^{-2} f \big\rangle_{L_2(d\rho)} \leqslant \lambda \big\langle f, (\hat{\Sigma} + \lambda I)^{-1} f \big\rangle_{L_2(d\rho)}$.

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

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

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

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

Thus $f \otimes_{L_2(d\rho)} f \preccurlyeq \Sigma$ (with the classical partial order between self-adjoint operators), 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 $\|\beta\|_2^2$ and the approximation error. We have, following a similar argument than the one of Bach (2013); El Alaoui and Mahoney (2014) for column sampling:

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

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

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

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 $X_i = \frac{1}{n}(\Sigma+\lambda I)^{-1}\Sigma - \frac{1}{n}\frac{1}{q(v_i)}\big[(\Sigma+\lambda I)^{-1/2}\varphi(v_i,\cdot)\big] \otimes_{L_2(d\rho)} \big[(\Sigma+\lambda I)^{-1/2}\varphi(v_i,\cdot)\big]$, so that our goal is to provide an upperbound on the probability that $\|\sum_{i=1}^n X_i\|_{\rm op} > t$. We denote by $d = \operatorname{tr}\Sigma(\Sigma+\lambda I)^{-1} = \int_{\mathcal{V}} \frac{\langle \varphi(v,\cdot),(\Sigma+\lambda I)^{-1}\varphi(v,\cdot)\rangle_{L_2(d\rho)}}{q(v)} q(v) d\mu(v) \leqslant d_{\max}$. We have $\mathbb{E}X_i = 0$, $\|X_i\|_{\rm op} \leqslant \frac{d_{\max}}{n}$ and

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}(X_i^2) &\preccurlyeq \frac{1}{n}\Bigg(\mathbb{E}\Bigg[\frac{\langle \varphi(v,\cdot),(\Sigma+\lambda I)^{-1}\varphi(v,\cdot)\rangle_{L_2(d\rho)}}{q(v)^2} \\ &\qquad \cdot\big[(\Sigma+\lambda I)^{-1/2}\varphi(v,\cdot)\big] \otimes \big[(\Sigma+\lambda I)^{-1/2}\varphi(v,\cdot)\big]\Bigg] \preccurlyeq \frac{d_{\max}}{n}(\Sigma+\lambda I)^{-1}\Sigma,\end{aligned}$$

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:

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

We now consider $t = \frac{3}{4}$, $\delta \in (0,1)$, and $n \geqslant 4 + 6d_{\max}\log\frac{4d_{\max}}{\delta} \geqslant 10 d_{\max}$. This implies that

$$\exp\Bigg(-\frac{nt^2}{2d_{\max}(1+3t)}\Bigg) \leqslant \frac{\delta}{4d_{\max}} \leqslant \frac{\delta}{4d}, \text{ and } \Bigg(1+\frac{6d_{\max}^2}{n^2\log^2(1+t)}\Bigg) \leqslant \Bigg(1+\frac{6}{100\log^2(7/4)}\Bigg) \leqslant 2.$$

Thus the probability is less than δ. 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.
- 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. 2).

– 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 $n \geqslant 4 + 6d_{\max}(\lambda)\log\frac{(\operatorname{tr}\Sigma)d_{\max}(\lambda)}{\lambda}$. We use this result in Section 4.4.

– We have $\mathbb{E}\operatorname{tr}\hat{\Sigma} = \operatorname{tr}\Sigma = \int_{\mathcal{X}} k(x,x)d\rho(x)$, and thus, by Markov's inequality, with probability $1-\delta$,

$$\operatorname{tr}\hat{\Sigma} \leqslant \frac{1}{\delta}\operatorname{tr}\Sigma. \tag{7}$$

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

C.2 Application to quadrature

In this section, we specialize the results from the section above to the quadrature subcase, namely we give a formulation where the features φ do not appear (or equivalently using ψ defined in Section 3.2).

We assume that points $x_1, \dots, x_n$ are sampled from the distribution with density q with respect to $d\rho$. We aim to write $f \in \mathcal{F}$ as $f = \int_{\mathcal{X}} \psi(x,\cdot)g(x)d\rho(x) = \Sigma^{-1/2}\int_{\mathcal{X}} k(x,\cdot)g(x)d\rho(x) = \Sigma^{1/2}g$. Moreover, we have

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

and we have

$$\langle f, \Sigma^{1/2}\hat{\Sigma}\Sigma^{1/2}g\rangle_{L_2(d\rho)} = \sum_{i=1}^n \frac{1}{q(x_i)}\left(\int_{\mathcal{X}} k(x_i,y)f(y)d\rho(y)\right)\left(\int_{\mathcal{X}} k(x_i,y)g(y)d\rho(y)\right).$$

We have from Eq. (6) $\beta_i = \frac{1}{nq(x_i)^{1/2}}\langle k(\cdot,x_i), \Sigma^{-1/2}(\hat{\Sigma}+\lambda I)^{-1}\Sigma^{1/2}g\rangle_{L_2(d\rho)}$, and the quadrature rule becomes:

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

which can be put in the form $\langle \hat{h}, g\rangle_{L_2(d\rho)}$ with the approximation $\hat{h} = \Sigma^{1/2}\hat{\Sigma}(\hat{\Sigma}+\lambda I)^{-1}\Sigma^{-1/2}h$.

Finally, the density q may be expressed as

$$q(x) \propto \langle k(\cdot,x), \Sigma^{-1/2}(\Sigma+\lambda I)^{-1}\Sigma^{-1/2}k(\cdot,x)\rangle_{L_2(d\rho)} = \sum_{i\in I}\frac{\mu_i}{\mu_i+\lambda}e_i(x)^2.$$

An alternative expression that can be used to compute q in practise is

$$q(x) \propto \langle k(\cdot,x), (I+\lambda\Sigma^{-1})^{-1}k(\cdot,x)\rangle_{L_2(d\rho)}. \tag{8}$$

C.3 Proof of Prop. 2

We first use the Varshamov-Gilbert's lemma (see, e.g., Massart, 2003, Lemma 4.7). That is, for an integer s, there exists a family $(\theta_j)_{j\in J}$ of at most $|J| \geqslant e^{s/8}$ 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 $f(\theta) = \frac{\sqrt{\mu_s}}{\sqrt{s}} \sum_{i=1}^s \theta_i e_i$, where (e_i, μ_i), $i = 1, \dots, s$ are the eigenvector/eigenvalue pairs associated with the s largest eigenvalues of Σ. We indeed have

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

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

Thus, if $\sqrt{4\lambda} \leqslant \sqrt{\frac{\mu_s}{4}}/3$, 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 any $j \neq j' \in J$, we have $\big\| \sum_{i=1}^n (\beta_j - \beta_{j'})_i \psi_i \big\|_{L_2(d\rho)} \geqslant \sqrt{\frac{\mu_s}{4}}/3$. This implies $n(4\delta^{-1} \operatorname{tr}\Sigma)\|\beta_j - \beta_{j'}\|_2^2 \geqslant (\beta_j - \beta_{j'})^\top \Psi^\top \Psi (\beta_j - \beta_{j'}) \geqslant \frac{\mu_s}{36}$, that is $\|\beta_j - \beta_{j'}\|_2 \geqslant \sqrt{\frac{\delta \mu_s}{144 n \operatorname{tr}\Sigma}} = \Delta$. Thus, $e^{s/8}$ is less than the packing number of the sphere 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 $\Delta/r \leqslant 1/24$, we have

$$\frac{s}{8} \leqslant n \left(\frac{1}{2} \log \frac{36 \operatorname{tr}\Sigma}{\delta \mu_s} + \log(2 + 1/24) \right).$$

This implies $n \geqslant \frac{s}{4 \log \frac{\operatorname{tr}\Sigma}{\delta \mu_s} + 21}$. Given that we have to choose $\mu_s \geqslant 144\lambda$ for the result to hold, this implies the desired result.