Journal of Machine Learning Research 26 (2025) 1-55 Submitted 11/23; Revised 4/25; Published 6/25

Efficient Numerical Integration in Reproducing Kernel Hilbert Spaces via Leverage Scores Sampling

Antoine Chatalic ANTOINE.CHATALIC@CNRS.FR
MaLGa Center - DIBRIS
Università di Genova
Via Dodecaneso 35, 16146 Genova, Italy

Nicolas Schreuder NICOLAS.SCHREUDER@CNRS.FR
MaLGa Center - DIBRIS
Università di Genova
Via Dodecaneso 35, 16146 Genova, Italy

Ernesto De Vito ERNESTO.DEVITO@UNIGE.IT
MaLGa Center - DIMA
Università di Genova
Via Dodecaneso 35, 16146 Genova, Italy

Lorenzo Rosasco LORENZO.ROSASCO@UNIGE.IT
MaLGa Center - DIBRIS, Università di Genova, Via Dodecaneso 35, 16146 Genova, Italy
Istituto Italiano di Tecnologia, Genoa, Italy
CBMM - Massachusets Institute of Technology, Cambridge, MA, USA

Editor: Florence d'Alché-Buc

Abstract

In this work we consider the problem of numerical integration, i.e., approximating integrals with respect to a target probability measure using only pointwise evaluations of the integrand. We focus on the setting in which the target distribution is only accessible through a set of n i.i.d. observations, and the integrand belongs to a reproducing kernel Hilbert space. We propose an efficient procedure which exploits a small i.i.d. random subset of $m < n$ samples drawn either uniformly or using approximate leverage scores from the initial observations. Our main result is an upper bound on the approximation error of this procedure for both sampling strategies. It yields sufficient conditions on the subsample size to recover the standard (optimal) $n^{-1/2}$ rate while reducing drastically the number of functions evaluations—and thus the overall computational cost. Moreover, we obtain rates with respect to the number m of evaluations of the integrand which adapt to its smoothness, and match known optimal rates for instance for Sobolev spaces. We illustrate our theoretical findings with numerical experiments on real datasets, which highlight the attractive efficiency-accuracy tradeoff of our method compared to existing randomized and greedy quadrature methods. We note that, the problem of numerical integration in RKHS amounts to designing a discrete approximation of the kernel mean embedding of the target distribution. As a consequence, direct applications of our results also include the efficient computation of maximum mean discrepancies between distributions and the design of efficient kernel-based tests.

Keywords: numerical integration, quadratures, kernel mean embeddings, maximum mean discrepancy, leverage scores, RKHS.

1. Introduction

Numerical integration is a key tool in applied mathematics and physics (Davis and Rabinowitz, 2007). It is particularly useful for approximating integrals that cannot be computed in closed form—for instance when the integrand depends on some data and does not have a simple analytical expression. It is used extensively in Bayesian inference (Gelman et al., 1995) as well as for the resolution of PDEs (Quarteroni and Valli, 2008), e.g. for the computation of the entries of the stiffness matrix used in finite elements methods, or in deep-learning-based approaches to estimate the loss function which is typically derived from a variational formulation of the problem (Rivera et al., 2022). Quadrature techniques are also commonly used in statistical physics for the computation of free energies, where one typically needs to integrate over large state spaces (Newman and Barkema, 1999).

Now we provide a formal definition of the problem. Let $(\mathcal{X}, \mathcal{B}, \rho)$ be a measurable space, and let $(\mathcal{H}, \|\cdot\|)$ be a normed vector space of functions defined over $\mathcal{X}$. We consider the problem of designing quadrature rules for functions in $\mathcal{H}$ with respect to a probability measure ρ. More precisely, we search for points $\tilde{X} := (\tilde{X}_1, \dots, \tilde{X}_m) \in \mathcal{X}^m$ (called the nodes or landmark points) and weights $w = [w_1, \dots, w_m]^T \in \mathbb{R}^m$ such that, for any function f in the unit ball of $\mathcal{H}$, the integral

$$\mathrm{I}(f) := \int f(x) \, \mathrm{d}\rho(x) \tag{1}$$

is well approximated by the quadrature rule defined by the weighted sum of pointwise evaluations

$$\mathrm{I}_{\tilde{X},w}(f) := \sum_{j=1}^{m} w_i f(\tilde{X}_i). \tag{2}$$

Importantly, the weights $(w_i)_{1 \le i \le m}$ can depend on the nodes $\tilde{X}$, but not on the integrand $f \in \mathcal{H}$. Moreover, we will consider the general setting in which the weights w are not required to be positive nor to sum to one, albeit some methods in the literature have been developed in order to satisfy such additional constraints, see for instance the work by Hayakawa et al. (2022). To quantify the performance of a given quadrature rule $\mathrm{I}_{\tilde{X},w}$, we define its approximation error as the worst-case error over the unit ball in $\mathcal{H}$,

$$\mathcal{E}(\mathcal{H}, \mathrm{I}_{\tilde{X},w}) := \sup_{f \in \mathcal{H} : \|f\| \le 1} \left| \mathrm{I}(f) - \mathrm{I}_{\tilde{X},w}(f) \right| . \tag{3}$$

We use the shorter notation $\mathcal{E}(\mathcal{H})$ when the quadrature rule $\mathrm{I}_{\tilde{X},w}$ is clear from the context.

Quadratures from empirical data We assume to have at our disposal a dataset of n i.i.d. samples $X = \{X_1, \dots, X_n\}$. A natural estimator of $\mathrm{I}(f)$ is the Monte-Carlo estimator

$$\hat{\mathrm{I}}(f) := \frac{1}{n} \sum_{i=1}^{n} f(X_i), \tag{4}$$

which estimates $\mathrm{I}(f)$ uniformly over the unit ball of $L^2(\rho)$ at the rate $O(1/\sqrt{n})$ with high probability. The complexity for computing this estimator grows linearly with the number n of samples in the dataset. We will be interested by applications in which one can easily

obtain i.i.d. samples from a target probability distribution, but pointwise evaluation of the integrand can be expensive. The need for approximating integrals of functions that are expensive to evaluate is a common problem which appears across many application fields, we refer for instance the reader to Oates et al. (2017) for an application to a computational cardiac model where each integrand evaluation takes about 100 CPU hours. It should be noted however that sampling from the target probability distribution may sometimes also be a major hurdle, in which case strategies such as Markov chain Monte Carlo (MCMC) or density estimation may be used (Oates et al., 2017; Delyon and Portier, 2016).

Objective Given a dataset X of n i.i.d. samples, our goal is to design a quadrature rule of the form (2) that (i) is computed using the knowledge of the n samples X and yet (ii) is supported on only $m < n$ nodes, while (iii) achieving the same finite-sample rate as the Monte-Carlo estimator $\hat{\mathrm{I}}$. We will show that these requirements are not incompatible, and that computational efficiency can be improved without sacrificing statistical accuracy.

We consider in particular the setting in which we first sample the nodes $(\tilde{X}_j)_{1\leq j\leq m}$ from the dataset X (so that the approximation bounds must hold with high probability on the draw of these points), and then set the weights $w = (w_j)_{1\leq j\leq m}$ deterministically.

1.1 Quadratures in Reproducing Kernel Hilbert Spaces

In this paper, we consider the setting in which $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) of functions over $\mathcal{X}$ with reproducing kernel κ (Aronszajn, 1950). Such spaces encompass many typical smoothness spaces considered in the learning literature. For instance, Sobolev spaces of high enough smoothness are RKHS, as reminded in the following example.

Example 1 (Sobolev)**.** *Let $s \in \mathbb{N}$. If $\mathcal{X} = \mathbb{R}^d$, denoting $\hat{f}$ the Fourier transform of f, the Sobolev space $\mathrm{H}^s(\mathbb{R}^d)$ is defined as*

$$\mathrm{H}^s(\mathbb{R}^d) := \left\{ f \in L^2(\mathbb{R}^d) \;\middle|\; \left(\int_{\mathcal{X}} (1+|\xi|^2)^s |\hat{f}(\xi)|^2 \,\mathrm{d}\xi\right)^{1/2} =: \|f\| < \infty \right\} .$$

When the smoothness parameter is high enough, namely $s > d/2$, $\mathrm{H}^s(\mathbb{R}^d)$ is an RKHS. For any non-empty domain $\Omega \subseteq \mathbb{R}^d$, we define $\mathrm{H}^s(\Omega)$ as the RKHS induced by the restriction of the reproducing kernel of $\mathrm{H}^s(\mathbb{R}^d)$ to $\Omega \times \Omega$.

When Ω has Lipschitz boundary, the above definition is known to be norm-equivalent to the alternative definition of Sobolev spaces involving weak derivatives (Wendland, 2004, Corollary 10.48). When $s > d/2$, it has been shown by Novak (1988, Section 1.3.12 Proposition 3) that the optimal rate for a deterministic quadrature rule of the form (2) on the unit hypercube is $\inf_{\tilde{X},w} \mathcal{E}(\mathrm{H}^s([0,1]^d), \mathrm{I}_{\tilde{X},w}) = \Theta(m^{-s/d})$, which suggests that the Monte-Carlo estimator (with $m = n$) might not be optimal in this setting as it gives the rate $m^{-1/2}$; see also Novak and Triebel (2006) for similar results on more general domains and the Lebesgue measure. The optimal rate can be reached in practice (see for instance the work by Briol et al. (2019) for quadrature rules based on Markov chain Monte-Carlo, by Santin et al. (2022) for greedy methods), and our goal is indeed to design quadrature rules that have this adaptivity to the smoothness of the considered RKHS in order to reduce the cost of numerical integration.

While the parameter s provides a direct control on the smoothness in the Sobolev setting, in this paper we develop a more generic analysis which depends on the decay of the spectrum of the integral operator associated to the reproducing kernel κ and the target distribution ρ.

Existing approaches Although we postpone to Section 3.3 the presentation of related works, we provide here a preliminary overview of the different approaches which have been proposed to design quadrature rules in reproducing kernel Hilbert spaces. Our method belongs to the family of random designs, obtained by sampling randomly and simultaneously the quadrature nodes; this includes i.i.d. uniform and importance sampling (Bach, 2017), as well as non-i.i.d. sampling strategies (Belhadji et al., 2019). Multiple greedy methods exist to iteratively select the nodes, typically by minimizing some notion of residual, or by filling the space as uniformly as possible (Briol et al., 2019). In the literature on core-sets, multiple algorithms have been proposed to compress the set of n samples down to m points using e.g. recursive halving approaches (Dwivedi and Mackey, 2022). Note that some of the methods can be declined in both deterministic and randomized variants, which makes it difficult to provide a clear classification of the literature.

Although formulated in the context of numerical integration in RKHS, our bounds can also be interpreted as approximation bounds for the computation of mean embeddings in reproducing kernel Hilbert spaces.

Remark 1 (Kernel Mean Embedding and Maximum Mean Discrepancy)**.** *Any quadrature rule in a RKHS can be interpreted as a way to approximate the so-called kernel mean embedding* $\mu := \int \kappa(x, \cdot)\, \mathrm{d}\rho(x) \in \mathcal{H}$ *of the probability distribution* ρ*. Indeed, when* $\mathcal{H}$ *is an RKHS it holds* $f(x) = \langle f, \kappa(x, \cdot)\rangle$ *for any* $f \in \mathcal{H}, x \in \mathcal{X}$*, and thus* $\mathrm{I}(f) = \langle f, \mu\rangle$*. This connection will be introduced and discussed in Section 4. It implies in particular that our work directly translates to algorithms and bounds for the efficient approximation of the maximum mean discrepancy, a standard metric between probability distributions in the context of kernel methods. The maximum mean discrepancy between two distributions indeed corresponds to the distance between their kernel mean embeddings.*

1.2 Summary of Contributions

This paper builds on the results by Chatalic et al. (2022b), that study kernel mean embeddings (see Remark 1) obtained by uniformly sampling the nodes. Our main contributions are the following:

- We introduce a quadrature rule whose nodes are randomly subsampled from the dataset X either uniformly or using leverage scores, and whose weights are optimally chosen by solving a least-square problem. This extends in particular the setting considered in (Chatalic et al., 2022b), which covers only uniform sampling.
- We provide high-probability bounds on the worst-case error of this quadrature rule, and obtain quantization rates (i.e. w.r.t. the number of nodes m) which are faster than the Monte-Carlo rate. For leverage score sampling, we obtain in particular asymptotic rates that match known optimal rates for Sobolev spaces (Novak, 1988).

- We show that our method adapts to the smoothness of the integrand by showing that faster rates can be derived for fractional subspaces of $\mathcal{H}$, i.e. assuming a source condition on the integrand (Engl et al., 2000).
- We compare empirically our method to other randomized and greedy approaches from the literature on real datasets, and show that our approach has a particularly interesting efficiency-accuracy tradeoff.

Layout The rest of the paper is organized as follows. We introduce our algorithm in Section 2. In Section 3, we summarize our main hypotheses and theoretical results, and put them in perspective by reviewing the state of the art. Leveraging tools from kernel methods, in Section 4 we detail how our bounds on the worst-case error are derived for both uniform and leverage scores sampling. We then compare experimentally our method with other quadrature approaches in Section 5. A table of notations is provided in Section A.

2. Two Algorithms Based on Subsampling

In this section, we describe the method we will analyze in the rest of the paper. It corresponds to a quadrature rule of the type (2) with randomly sampled nodes $(\tilde{X}_j)_{1\leq j\leq m}$ (Section 2.1) and weights w obtained by solving an unconstrained least-squares problem (Section 2.2).

2.1 Choice of the Nodes

We consider in the following two strategies for sampling the nodes $\tilde{X}$ from the empirical data X: uniform sampling and (ridge) leverage score sampling.

Uniform sampling The nodes $\tilde{X} = \{\tilde{X}_1, \dots, \tilde{X}_m\}$ are sampled uniformly from the set of all subsets of cardinality m of $\{X_1, \dots, X_n\}$. This is the most intuitive sampling strategy and arguably the easiest to implement. It will serve as a baseline against leverage scores sampling.

Approximate Ridge Leverage Score sampling (ARLS) Ridge leverage scores have been introduced by Alaoui and Mahoney (2015) in the setting of kernel ridge regression. They are related to the more general notion of statistical leverage score (Mahoney and Drineas, 2009). We now provide a formal definition.

Definition 2 (Ridge leverage scores)**.** *Given $n \geq 1$ data points $X_1, \dots, X_n$, let $K_n \in \mathbb{R}^{n\times n}$ denote the kernel matrix with entries $(K_n)_{i,j} = \kappa(X_i, X_j)$ for all $i, j \in [n]$. Let $\lambda > 0$. For any $i \in [n]$, the ridge leverage score of the datapoint X_i is defined as*

$$\ell_\lambda(i) \coloneqq \Big(K_n(K_n + \lambda n I)^{-1}\Big)_{ii}. \tag{5}$$

Such scores can be interpreted as a measure of the relative importance of each point in the dataset. They are directly related to Christoffel functions (Fanuel et al., 2022; Pauwels et al., 2018). The cost of exactly computing leverage scores quickly becomes prohibitive as the sample size grows due to the matrix inversion. Since the purpose of our approach is to reduce computational cost, we will rely on a approximate notion that has been studied in the literature.

Definition 3 (ARLS). *Let $\delta \in (0, 1]$, $\lambda_0 > 0$ and $z \in [1, \infty)$. A set $(\hat{\ell}_\lambda(i))_{i\in[n]}$ is said to be (z, λ_0, δ)-approximate ridge leverage scores (ARLS) of X if it satisfies with probability at least $1 - \delta$,*

$$\frac{1}{z}\ \ell_\lambda(i) \leq \hat{\ell}_\lambda(i) \leq z\ \ell_\lambda(i), \qquad \forall \lambda \geq \lambda_0, \forall i \in [n]. \tag{6}$$

Different algorithms have been proposed in the literature to obtain approximate ridge leverage scores. In this work we use BLESS (Rudi et al., 2018). It is based on a coarse-to-fine strategy with a computational cost of order $O(d_{\text{eff}}(\lambda)^2/\lambda)$, where $d_{\text{eff}}(\lambda)$ denotes the effective dimension defined in the next section. After computing the values $\hat{\ell}_\lambda(i)$, the landmarks $\tilde{X}$ are drawn with replacement from X proportionally to $\hat{\ell}_\lambda(i)$. We refer in the following to this method as ARLS sampling.

2.2 Choice of the Weights

Once the landmarks $\tilde{X}$ are selected, the weights are chosen as

$$w^* = \min_{w\in\mathbb{R}^m} \sup_{f\in\mathcal{H}:\|f\|\leq 1} |\hat{\mathrm{I}}(f) - \mathrm{I}_{\tilde{X},w}(f)|. \tag{7}$$

This problem is a least squares problem and our estimator can be computed using the closed form $w = \frac{1}{n} K_m^+ K_{mn} 1_n$ where A^+ denotes the Moore-Penrose pseudo-inverse of A, $K_m \in \mathbb{R}^{m\times m}$ and $K_{mn} \in \mathbb{R}^{m\times n}$ denote the kernel matrices with entries $(K_m)_{ij} = \kappa(\tilde{X}_i, \tilde{X}_j)$ for any $1 \leq i, j \leq m$ and $(K_{mn})_{ij} = \kappa(\tilde{X}_i, x_j)$ for any $1 \leq i \leq m$ and $1 \leq j \leq n$, and 1_n denotes a n-dimensional vector of ones. We refer the reader to Section C for a precise derivation of this expression.

We will see in Section 4 that the quadrature rule built using subsampling and optimal weights (7) is closely related to the so-called Nyström approximation. The latter is a standard way to approximate kernel matrices by rows/columns subsampling in the machine learning literature (Williams and Seeger, 2001), but actually takes its name from the work by Nyström (1930) to discretize linear integral equations, see also (Kress, 2014, Sec. 12.2). In this work, we thus use this designation in a broad sense: the subsampling procedure for the selection of nodes follows the literature on low-rank approximations of kernel matrices, however what we care about is the approximation of a linear operator, and thus the bounds we derive differ from what is usually done in the machine learning literature (see also Remark 12 in this regard).

Complexity The space complexity of the method (excluding the sampling phase) is $\Theta(m^2 + md)$ for storing K_m and the nodes. Note that K_{mn} does not need to be stored as $K_{mn}1_n$ can be computed sequentially in $\Theta(m)$ space. The time complexity (still excluding sampling) is $\Theta(nmc_\kappa + m^3)$ where c_κ corresponds to the cost of a kernel evaluation. The first term corresponds to the computation of $K_{mn}1_n$ while the second correspond to computing the pseudo-inverse of K_m (numerically stable algorithms can be used instead, but the complexity will be of this order regardless). When $\mathcal{X} \subseteq \mathbb{R}^d$, many standard kernel functions come with an evaluation cost which is of the order of the dimension, i.e. $c_\kappa = d$.

3. Main Results

We detail our technical assumptions in Section 3.1, and give an overview of our main results in Section 3.2. We then put our results in perspective by reviewing the state of the art in Section 3.3.

3.1 Assumptions

We recall that $(\mathcal{X}, \mathcal{B}, \rho)$ is a probability space, where ρ is the data probability distribution.

Assumption 4 (Independent and identically distributed samples)**.** *We have access to n data points $X_1, \dots, X_n$, drawn i.i.d. from the probability distribution ρ.*

The first assumption we make concerns the boundedness of the kernel.

Assumption 5 (Bounded kernel)**.** *$\mathcal{H}$ is a separable RKHS of functions on $\mathcal{X}$ with reproducing kernel κ. The canonical feature map $\phi : \mathcal{X} \to \mathcal{H}$, defined as $\phi(x) := \kappa(x, \cdot)$, is measurable for any $x \in \mathcal{X}$. There exists a positive constant $K < \infty$ such that $\sup_{x \in \mathcal{X}} \|\phi(x)\| \leq K$.*

Here and in the following, we denote $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ the RKHS inner-product and the associated norm. Assumption 5 is satisfied for feature maps derived from a large class of standard kernels such as, e.g., Gaussian and Laplacian kernels on the Euclidean space $\mathbb{R}^d$. It is also satisfied for polynomial kernels on a bounded domain $\mathcal{X}$.

We define the (uncentered) covariance operator of $\mathcal{H}$ for the target distribution ρ as

$$C = \int \phi(x) \otimes \phi(x) \, \mathrm{d}\rho(x) : \mathcal{H} \to \mathcal{H}.$$

where $(\phi(x) \otimes \phi(x))(f) := \langle f, \phi(x) \rangle \phi(x)$. Under Assumption 5 it holds $\mathrm{tr}(\phi(x) \otimes \phi(x)) = \|\phi(x)\|^2 \leq K^2$, and thus the operator C is a self-adjoint trace class operator on $\mathcal{H}$, which allows us to leverage tools from spectral theory. It is moreover a positive operator since $\phi(x) \otimes \phi(x)$ is positive for any x (cf. Section B).

We now define, for any $\lambda > 0$ the function

$$d_{\mathrm{eff}}(\lambda) := \mathbf{E}_{x \sim \rho} \|C_\lambda^{-1/2} \phi(x)\|^2 = \mathrm{tr}(C C_\lambda^{-1}), \tag{8}$$

where $C_\lambda := C + \lambda I$. Under Assumption 5, it always holds that $d_{\mathrm{eff}}(\lambda) \leq K^2/\lambda < \infty$ for any $\lambda > 0$. The quantity $d_{\mathrm{eff}}(\lambda)$ is known as the effective dimension, and is a measure of the interaction between the kernel (or feature map) and the data probability distribution. It is tightly linked to the notion of leverage scores and has been shown to constitute a proper measure of hardness for kernel ridge regression problems (Caponnetto and De Vito, 2007). It is a quantity of paramount importance in our analysis, and its decay w.r.t. λ essentially depends on the decay of the eigenvalues $(\sigma_i)_{i \in \mathbb{N}}$ of the covariance operator C, which characterizes the smoothness of the functions in $\mathcal{H}$. In this paper, we will assume that this decay is either polynomial or exponential, as formalized in the next two assumptions.

Assumption 6 (Polynomial Decay)**.** *There exist $\gamma \in]0, 1]$ and $a_\gamma > 0$ such that $\sigma_i \leq a_\gamma i^{-1/\gamma}$.*

Given that C is trace class, Assumption 6 always holds at least for $\gamma = 1$, however we are obviously interested in settings $\gamma < 1$ where better rates can be derived. We stress that assuming a polynomial decay of the spectrum of C is equivalent to assuming a polynomial decay of the effective dimension $d_{\text{eff}}(\lambda)$, see for instance Fischer and Steinwart (2020, Lemma 11).

Assumption 7 (Exponential Decay). *There exists $\beta > 0$ and $a_\beta > 0$ such that $\sigma_i \leq a_\beta e^{-\beta i}$.*

This assumption implies a bound on the effective dimension which is logarithmic in $1/\lambda$, as recalled in Section F.1.

The spectral decay of the covariance operator has been studied by Widom (1963, 1964), and it is known that Sobolev spaces on bounded domains correspond to a polynomial decay.

Remark 8 (Sobolev Decay). *For a bounded domain $\Omega \subseteq \mathbb{R}^d$, the Sobolev space $\mathrm{H}^s(\Omega)$ from Example 1 satisfies the polynomial decay assumption with $\gamma = d/(2s) < 1$ as shown by Widom (1963).*

In dimension $d = 1$, the Gaussian kernel associated to a gaussian density or a density supported on a compact domain yields an exponential decay, see Rasmussen and Williams (2006, Section 4.3.1) and Hayakawa et al. (2022, Section B.3). These results can be generalized to higher dimensions for product measures (Bach, 2017, Appendix A), yielding a decay of the form $\sigma_i \leq a_\beta e^{-\beta i^{1/d}}$. The same decay has been derived for the Gaussian kernel and a density with Gaussian tails in Harchaoui et al. (2008, Lemma 27) by combining the result of Widom (1964) with a perturbation argument.

3.2 Main Rates

We now provide an informal version of Theorem 19, which provides rates for our quadrature rule based on leverage scores sampling. We provide in Section 4 additional variants of this result for uniform sampling (yielding weaker rates) as well as for smoother fractional subspaces of $\mathcal{H}$ (yielding faster rates).

Theorem 9 (Main result, informal). *Let assumptions 4 and 5 hold. Let the nodes $\tilde{X}_1, \dots, \tilde{X}_m$ be drawn according to approximate leverage scores (6) from the dataset $\{X_1 \dots, X_n\}$, and w be the optimal weights (7). For n large enough it holds:*

- *under Assumption 6 (polynomial decay), choosing $m = \Omega(n^\gamma \log(n)^{1-\gamma})$, with high probability*

$$\mathcal{E}(\mathcal{H}, \mathrm{I}_{\tilde{X},w}) = O\left(\frac{\log(m)^{1/(2\gamma)}}{m^{1/(2\gamma)}}\right) = O\left(\frac{\log(n)^{1/2}}{n^{1/2}}\right);$$

- *under Assumption 7 (exponential decay), choosing $m = \Omega(\log(n)^2)$, with high probability*

$$\mathcal{E}(\mathcal{H}, \mathrm{I}_{\tilde{X},w}) = O\left(\frac{m^{1/4}}{\exp(\sqrt{m}/c)}\right) = O\left(\frac{\log(n)^{1/2}}{n^{1/2}}\right).$$

  for some constant c which does not depend on the dimension.

Our analysis uses some tools developed by Rudi et al. (2015) in the context of kernel ridge regression, as well as ideas developed by Chatalic et al. (2022a,b) for the approximation of kernel mean embeddings. Note that contrarily to methods which try to fill the domain as uniformly as possible, our analysis is not restricted to a bounded domain, and the constants in the $O(\cdot)$ and $\Omega(\cdot)$ notations of Theorem 9 do not depend on the dimension.

According to Remark 8, the polynomial decay hypothesis covers as a special case the Sobolev setting taking $\gamma = d/(2s)$.

Corollary 10 (Sobolev space)**.** *Under the hypotheses of Theorem 9, for* $\mathcal{H} = \mathrm{H}^s(\mathcal{X})$ *it holds*

$$\mathcal{E}(\mathrm{H}^s(\mathcal{X}), \mathrm{I}_{\tilde{X},w}) = O\left(\frac{\log(m)^{s/d}}{m^{s/d}}\right).$$

Optimality and smoothness adaptivity We stress that in all our results, the number of nodes m is directly chosen as a function of the number n of samples, and thus all rates can be interpreted w.r.t. to both variables. On one side, it is known from a statistical perspective that the minimax estimation rate when building the quadrature from n i.i.d. samples is $O(n^{-1/2})$ for continuous translation-invariant kernels on $\mathbb{R}^d$ and discrete measure, or measures with infinitely differentiable densities (e.g., Gaussian) (Tolstikhin et al., 2017); a similar rate has also been obtained in a non-iid setting (Chérief-Abdellatif and Alquier, 2022). Since we do not make extra assumptions on the probability distribution ρ in this paper, the bound in Theorem 9 is thus optimal (up to the log term) in the sense that no other estimator could reach a better rate with respect to n under the same assumptions. Quantization rates, on the other side, correspond to rates with respect to the number m of nodes on which the quadrature is supported, and lower bounds are known to be faster than $O(m^{-1/2})$ in this case, such as shown in Corollary 10 where we obtain a rate which adapts to the smoothness of the underlying space. This result turns out to match (up to log terms) the optimal rate in this setting (Novak, 1988, 1.3.12 Proposition 3).

For instance, Theorem 9 shows that in the case of polynomial decay our estimator achieves the quantization rate $O(m^{-1/(2\gamma)})$ (up to log terms) provided that one has access to $n = \Omega(m^{1/\gamma})$ i.i.d. samples in the first place. Alternatively, we recover the rate $O(n^{-1/2})$ (up to logarithmic terms) at the reduced cost of manipulating an estimator built using only $m = \Omega(n^{\gamma})$ samples. In the following, we will formulate the rates in this first manner (i.e. as a function of m) and always compare estimators built using the same number of samples: although the complexity of the algorithms used to pick these m points and weights may differ, the complexity of afterwards evaluating the quadrature rule for a new function is directly driven by m.

3.3 Related Work

Numerical approximation of integrals is a very broad topic and has a long history. Our focus here is on worst-case quadrature methods in reproducing kernel Hilbert spaces when one targets a non-uniform probability distribution known via i.i.d. samples. We provide below an overview of existing methods in the literature with a focus on available rates and associated computational complexities. One can roughly categorize these methods in a few categories: random designs (where the m nodes are sampled, either independently or

Method	Weights	Time complexity	Guarantees
Random selection of the nodes			
Monte-Carlo (Uniform)	Uniform	$O(m)$	$O(m^{-1/2})$ (Novak, 1988, 2.1.3)
MCMC targeting ρ (Briol et al., 2019)	Optimized	Not found	$\mathcal{E}(\mathrm{H}^s([0,1]^d), \mathrm{I}_{\tilde{X},w}) = O(n^{-s/d+\epsilon})$ for any $\epsilon > 0$
Projection DPP (Belhadji et al., 2019) (Requires eigendecomposition of C)	Optimized	Rejection sampling + $O(m^3)$	$(\mathbf{E}\,\mathcal{E}^2)^{1/2} \lesssim r_{m+1}^{1/2}$ (Belhadji, 2021, Theorem 4)
Ermakov-Zolotukhin (Belhadji, 2021)	Non-optimal	Rejection sampling + $O(m^3)$	$(\mathbf{E}\,\mathcal{E}^2)^{1/2} \lesssim r_{m+1}^{1/2}$ (Belhadji, 2021, Theorem 3)
Continuous volume sampling (Belhadji et al., 2020)	Optimized	$O(m^5)$ for MCMC mixing guarantees	$(\mathbf{E}\,\mathcal{E}^2)^{1/2} \lesssim \sigma_{m+1}^{1/2}$
(True) Leverage scores sampling (Bach, 2017)	By regularized LS	✗ No algorithm	$\mathcal{E} \leq 4\lambda$ provided $m \gtrsim d_{\mathrm{eff}}(\lambda)\log(\lambda^{-1})$
This work, Corollary 16 (Uniform)	Optimized	$\Theta(m^3 + nmd)$	Under Assumption 6: $\mathcal{E} \lesssim m^{-(1-\gamma/2)}\log(m)$ in particular $\mathcal{E}(\mathrm{H}^s(\mathcal{X})) = O\left(m^{-(1-d/4s)}\log(m)\right)$ Under Assumption 7 $\mathcal{E} = O(m^{-1}\log(m))$
This work, Theorem 19 ((A)RLS)	Optimized	$\Theta(m^3 + nmd + n^{1+2\gamma})$ $\Theta(m^3 + nmd + \log(n)^2 n)$	Under Assumption 6: $\mathcal{E} = O(m^{-1/(2\gamma)}\log(m))$ in particular $\mathcal{E}(\mathrm{H}^s(\mathcal{X})) \lesssim m^{-s/d}\,\mathrm{polylog}(m)$ Under Assumption 7: $\mathcal{E} = O(m^{1/4}\exp(-\sqrt{m}/\sqrt{cst}))$
Greedy methods focusing on the residual			
f/P-greedy on $\mathcal{X}$ (Müller, 2009) / SBQ (Huszár and Duvenaud, 2012)	Optimized	$O(m^3) + m$ nonconvex subproblems $O(dn + m^2)$/objective evaluation	$\mathcal{E} = O(m^{-1/2})$ ($\mathcal{X}$ bounded) (Santin et al., 2022, Theorem 5.1)
f/P-greedy on X	Optimized	$O(n^2 + nm(d+m))$	✗ Not found.
Herding (Chen et al., 2010)	Uniform	m non-convex subproblems with $O(nd)$/objective evaluation	$\mathcal{E} = O(m^{-1})$ in finite dimension (Chen et al., 2010) $\mathcal{E} = O(m^{-1/2})$ otherwise
Frank-Wolfe (FW) with line search	In the simplex		Exponential in finite dimension (Bach et al., 2012) $\mathcal{E} = O(m^{-1/2})$ otherwise
Fully-corrective FW (Jaggi, 2013) / Continuous OMP / f-greedy	Optimized		Exponential in finite dimension (Bach et al., 2012) $\mathcal{E} = O(m^{-1/2})$ otherwise
OMP (a.k.a. f-greedy) on X	Optimized	$O(n^2 + nm(d+m))$	$\mathcal{E} = O(m^{-1/2})$ (DeVore and Temlyakov, 1996)
Continuous OMP w/ global steps + Nyström or RF approximation (Chatalic et al., 2022a; Keriven et al., 2017)	Optimized	For RF: $O(nmd\log(d))$ $+ m$ non-convex subproblems $O(m^2 d\log(d))$/objective eval.	✗ Not found.
Other approaches			
Recombination Mercer (Hayakawa et al., 2022) (Requires eigendecomposition of C)	Opt. in simplex	$O(nm^2 + m^3)$ in average	$(\mathbf{E}_X\,\mathcal{E}^2)^{1/2} \lesssim r_m^{1/2} + n^{-1/2}$ (Hayakawa et al., 2022, Cor. 2)
Recombination Nyström (Hayakawa et al., 2022)	Opt. in simplex	$O(nm^2 + m^3\log(n/m))$	Under Assumption 7: $\mathbf{E}[\mathcal{E}] = O(r_{m+1}^{1/2} + \mathrm{polylog}(m)/m + n^{-1/2})$ (Hayakawa et al., 2023, Th. 6 + Rem. 1)
Thinning (Dwivedi and Mackey, 2022, 2021)	Uniform	$O(n^2 c_\kappa)$	For $m = \sqrt{n}$, subexponential/compactly supported distrib.: analytic kernel: $\mathcal{E} \lesssim \mathrm{polylog}(m)m^{-1}$ Matérn kernel: $\mathcal{E} \lesssim \mathrm{polylog}(m)m^{-(1-d/\lfloor s \rfloor)}$
Thinning (Shetty et al., 2022)	Uniform	$O(n\log(n)^3)$	
Space-filling methods			
P-greedy	Optimized	m non-convex subproblems, $O(m^2 + md)$ / objective evaluation	$\mathcal{E}(\mathrm{H}^s(\mathcal{X})) = O(m^{-s/d})$ ($\mathcal{X}$ bounded w/ cone condition, TI kernel) (Santin et al., 2022, Th. 3.2/Rem. 4.1)
P-greedy on X	Optimized	$O(nm(d+m))$	✗ Not found.

Table 1: Summary of main quadrature methods. We denote $\mathcal{E} := \mathcal{E}(\mathcal{H}, \mathrm{I}_{\tilde{X},w})$ for conciseness for a generic RKHS $\mathcal{H}$, and $r_m = \sum_{j \geq m} \sigma_j$. Complexities are given assuming that the kernel evaluation costs $O(d)$. For greedy algorithms, complexities are intended w.r.t. the empirical problem, so that the nonconvex subproblems have complexities depending on n. SBQ = sequential Bayesian quadrature; OMP = orthogonal matching pursuit; RF = random features; DPP = determinantal point process; TI = translation-invariant. Note that under Assumption 6, for $\gamma < 1$ it holds $r_m \leq \frac{\gamma\, a_\gamma}{1-\gamma}(m-1)^{1-1/\gamma}$, and under Assumption 7 it holds $r_m \leq \frac{a_\beta}{1-e^{-\beta}} e^{-\beta m}$.

jointly), coreset methods (which reduce, often recursively, the initial set of n samples while maintaining some key properties), methods which try to fill the space, and greedy methods which pick the nodes iteratively. Table 1 provides a summary of the different approaches. We refer the reader to Novak and Wozniakowski (2010) for a broader coverage of the topic, and in particular of existing lower bounds in terms of information complexity.

Random designs Our method belongs to the family of random designs, in the sense that the locations of the nodes are randomly drawn - in our case subsampled among the i.i.d. samples X, but this could be relaxed. The simplest way to produce a random design is the Monte-Carlo method, which achieves a $O(m^{-1/2})$ rate (Novak, 1988, 2.1.3). This rate is optimal in many settings when having access to m i.i.d. samples, e.g. for translation-invariant kernels and discrete measures or measures with infinitely differentiable densities (Tolstikhin et al., 2017), however we consider here quadrature rules built starting from $n > m$ i.i.d. samples and that can thus have better rates with respect to m.

Our method is closely related to the work of Bach (2017), who considers i.i.d. sampling of the nodes according to (continuous) leverage scores and slightly different weights. For a particular choice of the random features, the bound in Bach (2017, Proposition 1) translates to a bound on the worst-case error. However, in general the method cannot be implemented as it involves multiple quantities that cannot be computed.

Briol et al. (2017) have also introduced a heuristic distribution with heavy tails as well as a sequential Monte-Carlo procedure to sample from it, and reported empirically better stability.

Joint sampling of the nodes has been considered, for instance using determinantal point processes (Belhadji, 2021), which is also related to the Ermakov-Zolotukhin quadrature rule (Belhadji, 2021). Defining $r_m = \sum_{i \geq m} \sigma_i$, theoretical convergence rates of order $\mathbf{E}[\mathcal{E}(\mathcal{H})^2] = O(r_{m+1})$ have been proven for both methods. Belhadji et al. (2020) also considered continuous volume sampling, which consists in jointly sampling the nodes following a probability density $\det(K_m)$ with respect to the base measure $\rho^{\otimes m}$. This method yields a faster theoretical rate $\mathbf{E}[\mathcal{E}(\mathcal{H})^2] = O(\sigma_{m+1})$. Empirically, DPP sampling has also been reported to converge at this faster rate.

Random sampling from data streams, i.e. in one pass over the data without knowing beforehand the size n of the dataset, has been investigated by Paige et al. (2016); no convergence rates have however been reported in this setting. Note that our quadrature rule can be interpreted as a kind of Nyström approximation (Williams and Seeger, 2001), and many other sampling rules have been studied in this context (Fanuel et al., 2022; Kumar et al., 2012).

Space-filling methods In the setting where $\mathcal{X}$ is a compact set, multiple methods have been proposed to fill the space with more regularity than what a Monte-Carlo sample would typically produce. Such methods have been studied for decades in the literature on model-free design of experiments, see for instance Garud et al. (2017). Quasi Monte-Carlo (QMC) methods is a well-known way to generate low-discrepancy sequences, but is usually restricted to very particular domain and distributions - such as the uniform distribution on the hypercube, or the Gaussian distribution on the sphere. Dick and Pillichshammer (2010, Theorem 15.21) for instance derived rates that are arbitrary close to the optimal one for

QMC and Sobolev spaces of dominating mixed smoothness on $[0,1]^d$, see also Briol et al. (2019). We refer the reader to Dick et al. (2022) for a broader coverage of these methods.

In the context of kernel interpolation, Fekete points are defined as the nodes $\tilde{X}$ maximizing $\det(K_m)$, by analogy with polynomial interpolation in 1d where one is interested in the points maximizing the determinant of the Vandermonde matrix (Bos et al., 2010). Maximizing directly $\det(K_m)$ is most often untractable or expensive, but kernel approximations can naturally be used (Karvonen et al., 2021). Note that this objective is related to the density used in the continuous volume sampling method mentioned above (Belhadji et al., 2020), however there is here no dependence in the probability measure ρ (or ρ is assumed to be uniform).

Greedy maximization of $\det(K_m)$ as been introduced as the P-greedy method in the kernel interpolation literature (De Marchi et al., 2005, Section 4) (cf. Section H.2 for more details), and used in multiple contexts (Chen et al., 2018; Carratino et al., 2021).

Other randomized methods Recently, Hayakawa et al. (2022, 2023) used recombination algorithms to compute a discrete measure ρ_m supported on m points such that for a set of m test functions $(\varphi_i)_{1\leq i\leq m}$ it holds exactly $\int \varphi_i \,\mathrm{d}\rho_m = \int \varphi_i \,\mathrm{d}\hat{\rho}_n$. The test functions are built either using the Mercer decomposition or using a Nyström approximation with truncation, and both randomized and deterministic algorithms are known to compute the reduction from $\hat{\rho}_n$ to ρ_m. Assuming an exponential decay of the covariance's spectrum, the authors obtain a bound on the expected worst case error in $\mathbf{E}[\mathcal{E}(\mathcal{H})] = O(\sqrt{r_{m+1}} + \mathrm{polylog}(m)/m + n^{-1/2})$ (Hayakawa et al., 2022, Theorem 6, Remark 1).

Quadrature rules which are supported on a subset of the initial n samples (as we do) can also be interpreted as (weighted) coresets. For instance, the simple greedy algorithm of Karnin and Liberty (2019, Section 3.1) covers the case of kernel density estimation as a special case, however it only induces a $O(n^{-1/2})$ rate. Thinning methods have been proposed to build a coreset of size $m = \sqrt{n}$ by recursively reducing by half the initial dataset. The initial $O(n^2)$ complexity of kernel thinning (Dwivedi and Mackey, 2021) has been reduced to $O(n\log(n)^3)$ by Shetty et al. (2022), and the error of the coreset has been studied under various hypotheses but goes down to $O(\mathrm{polylog}(m)/m)$ for e.g. a Gaussian kernel with a sub-exponential data distribution (Dwivedi and Mackey, 2022).

Greedy methods An alternative to random design (where the m nodes are sampled, either i.i.d. or jointly) and coreset methods (which often recursively reduce the initial set of n samples), is to iteratively select the nodes by minimizing some notion of residual.

Kernel herding (Chen et al., 2010) falls in this category, and has originally been introduced with uniform quadrature weights. It can be interpreted as a particular case of the Frank-Wolfe algorithm (Bach et al., 2012) and has been extended in multiple directions (Jaggi, 2013; Lacoste-Julien et al., 2015; Briol et al., 2015). These algorithms are also known to be closely related to matching pursuit and its variants (Locatello et al., 2017). Fast rates in $O(1/m)$ and even exponential rates have been obtained for such methods, but depend on geometric quantities that cannot always be controlled easily and thus essentially cover finite-dimensional spaces. Khanna et al. (2021) derived rates that hold in infinite dimension, but rely on the hypothesis that the target distribution is sparse. Tsuji et al. (2022) introduced blended pairwise conditional gradients as a variant of Frank-Wolfe more

amenable to analysis in the infinite-dimensional setting, however theoretical rates remain of order $O(m^{-1/2})$.

In order to limit the impact of local minimas, global optimization steps can be added after each selection of a new node. This leads to the compressive clustering algorithm, which additionally relies on random features or Nyström approximations of the kernel (Keriven et al., 2017; Chatalic et al., 2022a) and is closely related to the sliding Frank-Wolfe algorithm (Denoyelle et al., 2019). Although theoretical guarantees in this context rather focus on the recovery of sparse measures, the considered objective function corresponds to a tractable approximation of the quadrature worst-case error and the algorithms proposed in this context are thus highly relevant for our goal.

Interestingly, greedy minimization of the quadrature worst-case error $\inf_w \mathcal{E}(\mathcal{H}, \mathrm{I}_{\tilde{X},w})$ actually does not lead to orthogonal matching pursuit, but to the f/P-greedy method from the kernel interpolation literature (Müller, 2009), which is also known as sequential Bayesian quadrature (Huszár and Duvenaud, 2012). Rates of order $O(m^{-1/2})$ have been obtained both for f-greedy and f/P-greedy methods (Santin et al., 2022, Corollary 20), however faster rates are typically observed in practice.

Bayesian Quadratures In the Bayesian literature, one is typically interested in computing not only the integral $\mathrm{I}(f)$, but also a probability distribution encoding the belief in this estimation. To achieve this goal, a prior distribution over the integrand f is assumed. When this prior is chosen to be a Gaussian process whose covariance function is a kernel κ, the maximum a posteriori estimator corresponds to the optimally-weighted quadrature rule in the RKHS associated to κ. Moreover, the variance of $\mathrm{I}(g)$ when g follows the posterior distribution corresponds exactly to the worst-case error of the optimally-weighted quadrature supported on the m nodes, $\mathrm{Var}[\mathrm{I}(g)] = \inf_{w\in\mathbb{R}^m} \mathcal{E}(\mathcal{H}, \mathrm{I}_{\tilde{X},w})$, see e.g. (Huszár and Duvenaud, 2012, Section 3.2). This gives another interpretation to our target objective, and justifies for instance that the sequential Bayesian quadrature is equivalent to the greedy minimization of the worst-case error (cf. section H.2).

In this Bayesian context, Briol et al. (2019) derived optimal convergence rates for MCMC sampling in Sobolev spaces on $[0,1]^d$ using bounds based of the fill distance, and Quasi Monte-Carlo sampling in Sobolev spaces of dominating mixed smoothness using a result from Dick and Pillichshammer (2010).

Rates for adaptive bayesian quadrature methods, for which the choice of the nodes is allowed to depend on the integrand, have also been studied assuming that the target function can be modeled as a transformation of a Gaussian process (Kanagawa and Hennig, 2019).

Other contributions A few other methods exist beyond the main families of algorithms presented above, such as particle methods which start directly from a pool of m nodes whose locations are jointly updated by gradient descent (Arbel et al., 2019), but no rates have been reported in this setting. Muandet et al. (2014) introduced shrinkage estimators and showed that they perform better than Monte-Carlo approaches under mild assumptions on the probability distribution of interest. Such shrinkage strategies are complementary to our approach, in the sense that they can be combined with any existing estimator. In another context, Kanagawa et al. (2020) proposed a theoretical analysis in the misspecified setting

(i.e. when the integrand in (1) does not belong to the RKHS used to design the quadrature rule), and showed that adaptivity to the smoothness of the integrand can still be achieved.

Summary Overall, our approach has the merit of achieving optimal rates while being efficiently implementable, which complements nicely the state of the art. For instance, greedy methods obtain very good empirical results, but the observed rates are not matched by existing theoretical guarantees. Other existing random designs do not always yield optimal rates, and are often costly to implement, when not intractable. Methods trying to fill uniformly the domain are restricted by definition to bounded domains, and perform poorly in practice (see Section 5) despite optimal rates being known in some settings (Santin et al., 2022); this can likely be explained by high multiplicative constants, and the fact that such methods do not adapt to the target distribution.

4. Theoretical Analysis

We show in Section 4.1 that the problem of designing quadrature rules can be recast as the approximation of the so-called kernel mean embedding, and then provide bounds on the worst-case error for uniform sampling (Section 4.2), ARLS sampling (Section 4.3), as well as improved rates for ARLS sampling under an additional smoothness condition (Section 4.4).

4.1 RKHS Quadratures and Kernel Mean Embeddings

When considering $\mathcal{H}$ to be a reproducing kernel Hilbert space associated to a kernel κ satisfying Assumption 5, the quadrature error is connected to the approximation of the so-called kernel mean embedding of the considered probability measure ρ,

$$\mu := \mu(\rho) := \int \phi(x) \, \mathrm{d}\rho(x). \tag{9}$$

Indeed ϕ is integrable with respect to any probability distribution over $\mathcal{X}$ under Assumption 5, and thus the kernel mean embedding (9) is well defined, interpreting the integral as a Bochner integral (Diestel and Uhl, 1977, Chapter 2). It should be noted that for any y the linear functional $h \mapsto \langle \phi(y), h \rangle$ is bounded under Assumption 5, and thus closed (Kreyszig, 1989, 4.13.5 (a)). Hence by Hille's theorem (Diestel and Uhl, 1977, Theorem 6) it holds $\langle \phi(y), \int \phi(x) \, \mathrm{d}\rho(x) \rangle = \int \langle \phi(y), \phi(x) \rangle \, \mathrm{d}\rho(x) = \int \kappa(y, x) \, \mathrm{d}\rho(x)$. Moreover, the operator $I_\rho : f \mapsto \int f(x) \, \mathrm{d}\rho(x)$ is a continuous linear functional under Assumption 5 given that $|I_\rho f| \leq \int |\langle f, \phi(x) \rangle| \, \mathrm{d}\rho(x) \leq K \|f\|$, and thus admits a Riesz representation m_ρ, i.e. $I_\rho(f) = \langle f, m_\rho \rangle$ holds for any $f \in \mathcal{H}$ (Reed and Simon, 1981). Considering $f = \phi(y)$ we get $m_\rho(y) = \int \kappa(y, x) \, \mathrm{d}\rho(x)$ for any y, i.e., the kernel mean embedding is also the Riesz representant of I_ρ.

Initially introduced by Smola et al. (2007), kernel mean embeddings (KME) conveniently allow to represent a probability distribution via a mean vector in a Hilbert spaces (Muandet et al., 2017). They have found applications in various areas such as anomaly detection (Zou et al., 2014), approximate Bayesian computation (Park et al., 2016), domain adaptation (Zhang et al., 2013), imitation learning (Kim and Park, 2018), nonparametric inference in graphical models (Song et al., 2013), functional data analysis (Hayati et al., 2020), discriminative learning for probability measures (Muandet et al., 2012) and differential privacy (Balog et al., 2018; Chatalic et al., 2021).

In the following lemma, we show how the error of a quadrature rule can be related to the error of a kernel mean embedding estimation problem. This result is common knowledge, but included for completeness.

Lemma 11. *For any set of points $\tilde{X} = (\tilde{X}_i)_{1\leq j\leq m}$ and any weights $(w_i)_{1\leq i\leq m}$, it holds*

$$\mathcal{E}(\mathcal{H}, \mathrm{I}_{\tilde{X},w}) = \left\| \mu - \sum_{j=1}^{m} w_j \phi(\tilde{X}_j) \right\| .$$

Proof For any $h \in \mathcal{H}$ such that $\|h\| \leq 1$, it holds that

$$\begin{aligned}
\left| \int h(x)\,\mathrm{d}\rho(x) - \sum_{j=1}^{m} w_j h(\tilde{X}_j) \right| &\overset{(i)}{=} \left| \int \langle h, \phi(x)\rangle\,\mathrm{d}\rho(x) - \sum_{j=1}^{m} w_i \langle h, \phi(\tilde{X}_j)\rangle \right| \\
&\overset{(ii)}{=} \left| \left\langle h, \int \phi(x)\,\mathrm{d}\rho(x) - \sum_{j=1}^{m} w_j \phi(\tilde{X}_j) \right\rangle \right| \\
&\overset{(iii)}{\leq} \left\| \mu - \sum_{j=1}^{m} w_j \phi(\tilde{X}_j) \right\| ,
\end{aligned}$$

where we used the reproducing property of the RKHS $\mathcal{H}$ for (i) and the Cauchy-Schwarz inequality for (iii). Equality (ii) follows from Hille's theorem (Diestel and Uhl, 1977, Theorem 6) applied to the linear functional $f \mapsto \langle h, f\rangle$, which is bounded given that $\|h\| \leq 1$ and thus closed (Kreyszig, 1989, 4.13.5 (a)).

The proof is concluded by observing that

$$h = \left\| \mu - \sum_{j=1}^{m} w_j \phi(\tilde{X}_j) \right\|^{-1} \left(\mu - \sum_{j=1}^{m} w_j \phi(\tilde{X}_j) \right) ,$$

is on the unit sphere in $\mathcal{H}$ and gives the equality. ■

Discrete estimators Denoting $\hat{\rho}_n = \frac{1}{n}\sum_{1\leq i\leq n} \delta(X_i)$ the empirical distribution of X, where $\delta(\cdot)$ denotes the Dirac delta function, one can define

$$\hat{\mu}_n := \mu(\hat{\rho}_n) = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i). \tag{10}$$

By Lemma 11, the error of the empirical estimator (4) is $\mathcal{E}(\mathcal{H}, \hat{\mathrm{I}}) = \|\hat{\mu}_n - \mu\|$. This quantity decreases at the rate $O(1/\sqrt{n})$ for any ρ as a consequence of Bernstein inequality in Hilbert spaces (Yurinsky, 1995, Th. 3.3.4). More generally, any quadrature rule $\mathrm{I}_{\tilde{X},w}$ can be associated to a sparse estimator of the kernel mean embedding

$$\tilde{\mu}_m := \sum_{j=1}^{m} w_j \phi(\tilde{X}_j) \tag{11}$$

and the discrete approximation (2) can be computed as $\mathrm{I}_{\tilde{X},w}(f) = \langle \tilde{\mu}_m, f\rangle$ for any $f \in \mathcal{H}$.

A randomized Nyström estimator Our quadrature rule, obtained by sampling the landmarks $\tilde{X}$ from the data X and choosing the weights according to (7), has a simple expression in terms of kernel mean embeddings. Let

$$\mathcal{H}_m := \operatorname{span}\Big\{\phi(\tilde{X}_1), \dots, \phi(\tilde{X}_m)\Big\} \subseteq \mathcal{H}$$

be the finite dimensional subspace spanned by the features of the landmarks, and P_m the orthogonal projection on this subspace, one can easily check (see Section C) that

$$\tilde{\mu}_m := P_m \hat{\mu}_n. \tag{12}$$

One can in particular think of $\tilde{\mu}_m$ as an interpolator of $\hat{\mu}_n$ at the location of the nodes, given that for any $j \in \{1, \dots, m\}$, as $\phi(\tilde{X}_j) \in \operatorname{ran}(P_m)$ it holds $\tilde{\mu}_m(\tilde{X}_j) = \langle P_m \hat{\mu}_n, \phi(\tilde{X}_j)\rangle = \langle \hat{\mu}_n, \phi(\tilde{X}_j)\rangle = \hat{\mu}_n(\tilde{X}_j)$.

As a consequence of (12) and Lemma 11, our main goal from a theoretical perspective is to bound the quantity

$$\mathcal{E}(\mathcal{H}, \mathrm{I}_{\tilde{X},w}) = \|\mu - P_m \hat{\mu}_n\|$$

both for uniform and ARLS sampling.

Remark 12 (Kernel matrix)**.** *It can easily be checked that*

$$\|\hat{\mu}_n - \tilde{\mu}_m\|^2 = \|P_m^{\perp} \hat{\mu}_n\|^2 \leq \tfrac{1}{n} \|K_n - \tilde{K}_n\|_{op}$$

where K_n and $\tilde{K}_n$ respectively denote the $n \times n$ kernel matrices of the data X with and without Nyström approximation. Hence, existing results on the Nyström approximation of the kernel matrix in operator norm induce bounds on the worst-case quadrature error, using the error decomposition $\mathcal{E}(\mathcal{H}, \mathrm{I}_{\tilde{X},w}) \leq \|\mu - \hat{\mu}_n\| + \|\hat{\mu}_n - P_m \hat{\mu}_n\|$. Such bounds would however be sub-optimal, and we thus rely for our analysis on a different decomposition.

Remark 13 (Power function)**.** *In another context, Hayakawa et al. (2023) obtained quadrature guarantees by studying the integral w.r.t. the probability distribution ρ of the quantity $\|P_m^{\perp} \phi(x)\|$, which is known in the kernel interpolation literature as the power function and has been well studied (Wendland, 2004). This still differs from our analysis, which rather relies on bounds on $\|P_m^{\perp}(C + \lambda I)^{1/2}\|$.*

Remark 14 (Maximum Mean Discrepancy)**.** *Mean embeddings naturally induce a semi-metric on the space of probability distributions $\mathcal{P}(\mathcal{X})$ known as the maximum mean discrepancy (Smola et al., 2007). It is defined, for any two probability distributions ρ_1 and ρ_2, as*

$$\mathrm{MMD}(\rho_1, \rho_2) := \|\mu(\rho_1) - \mu(\rho_2)\| \ .$$

It satisfies all the properties of a metric except, in general, the definiteness, depending on whether the mean embedding $\rho \mapsto \mu(\rho)$ is injective or not (we refer the interested reader to Sriperumbudur et al. (2010) for more details). Such metrics have found applications in many contexts such as, to cite a few, two-sample testing (Gretton et al., 2012; Borgwardt et al., 2006), neural networks optimization (Borgwardt et al., 2006), generative models (Li et al., 2017; Sutherland et al., 2017). Given their wide applicability, maximum mean

discrepancies are also an important motivation for better approximating mean embeddings. An interesting property of the MMD is that it is an integral probability metric (Müller, 1997), a class of metrics which uses test functions to compare distributions. More precisely, we have

$$\mathrm{MMD}(\rho_1, \rho_2) = \sup_{f \in \mathcal{H}: \|f\| \leq 1} |\mathbb{E}_{X_1 \sim \rho_1} f(X_1) - \mathbb{E}_{X_2 \sim \rho_2} f(X_2)|$$

where $\mathcal{H}$ denotes the reproducing kernel Hilbert space associated to the chosen kernel. These two representations of the MMD allow to leverage the wide set of tools from both kernel methods and integral probability metric theories (see Sriperumbudur et al. (2012, 2009) for examples of the latter). Although we focus on the problem of designing quadratures, it should be noted that the algorithms and bounds discussed in this paper directly translate to results on the MMD, see for instance the discussion in Chatalic et al. (2022b, Section 5).

4.2 Rates for Uniform Sampling

We now state our general result for uniform sampling. We then specialize it using additional knowledge on the spectral properties of the covariance operator. This result was initially presented in Chatalic et al. (2022b). We restate it for completeness and for comparison with ARLS sampling. In the following, we denote $\mathcal{L}(\mathcal{H})$ the set of bounded linear operators from $\mathcal{H}$ to itself, and $\|\cdot\|_{\mathcal{L}(\mathcal{H})}$ the operator norm on $\mathcal{L}(\mathcal{H})$.

Theorem 15. *Let Assumptions 4 and 5 hold. Let $12 \leq m \leq n$ and let $\delta \in (0, 1)$. When the m sub-samples $\tilde{X}_1, \ldots, \tilde{X}_m$ are drawn uniformly without replacement from the dataset $\{X_1 \ldots, X_n\}$ and w is chosen as in (7), it holds with probability at least $1 - \delta$ that*

$$\mathcal{E}(\mathcal{H}, \mathrm{I}_{\tilde{X},w}) \leq \frac{c_1}{\sqrt{n}} + \frac{c_2}{m} + \frac{c_3 \sqrt{\log(m/\delta)}}{m} \sqrt{d_{\mathrm{eff}}\left(\frac{12K^2 \log(m/\delta)}{m}\right)}, \tag{13}$$

provided that

$$m \geq \max(67, 12K^2 \|C\|^{-1}_{\mathcal{L}(\mathcal{H})}) \log\left(\frac{m}{\delta}\right),$$

where c_1, c_2, c_3 are constants of order $K \log(1/\delta)$.

The constants c_1, c_2, c_3 are made explicit in the proof. A few remarks regarding Theorem 15 are in order. First, denoting by W the smallest branch of the Lambert's W function on $]-e^{-1}, 0[$ (Weisstein, 2002), the condition on the sub-sample size m can also be expressed as $m \geq -W(-\delta/c)c$ with $c = \max(67, 12K^2\|C\|^{-1}_{\mathcal{L}(\mathcal{H})})$ and can thus easily be checked numerically.

Then, the bound on the error is split in three parts: the first part corresponds to the usual rate one gets estimating the kernel mean embedding by its standard empirical counterpart, while the second part and the third part result from the approximation. Note that the first two terms already illustrate the trade-off between computational cost and statistical performance of our estimator: a small value of m (i.e $m < \sqrt{n}$) will reduce the computational burden, but yield a rate worse than $O(1/\sqrt{n})$; alternatively, taking $m > \sqrt{n}$ would not improve the overall error rate, but would require more computational and storage resources. The precise trade-off can be settled by the third term, which depends simultaneously on the subsample size m and on the effective dimension $d_{\mathrm{eff}}(\lambda)$. Extra assumptions about the

effective dimension – which depends both on the kernel and the probability distribution – are needed to obtain a more explicit bound. We thus specialize our result under Assumption 6 and Assumption 7, and present in both cases sufficient conditions on m and n to guarantee a $O(n^{-1/2})$ rate, and quantization rates w.r.t. m that are faster than the Monte-Carlo $O(m^{-1/2})$ rate.

Corollary 16 (Polynomial decay)**.** *Under the assumptions of Theorem 15, if the RKHS $\mathcal{H}$ and ρ satisfy Assumption 6, taking $m := n^{1/(2-\gamma)} \log(n/\delta)$ it holds*

$$\mathcal{E}(\mathcal{H}, \mathrm{I}_{\tilde{X},w}) = O\left(\frac{\log(m)^{1-\gamma/2}}{m^{1-\gamma/2}}\right).$$

According to Remark 8, we get the following result for Sobolev spaces.

Corollary 17 (Sobolev space)**.** *When $s > d/2$, under the assumptions of Theorem 15, taking $m := n^{1/(2-\gamma)} \log(n/\delta)$ it holds*

$$\mathcal{E}(\mathrm{H}^s(\mathcal{X}), \mathrm{I}_{\tilde{X},w}) = O\left(\frac{\log(m)^{1-d/(4s)}}{m^{1-d/(4s)}}\right).$$

The polynomial decay assumption always holds with $\gamma = 1$, but no compression is achieved in this setting. However as soon as $\gamma < 1$, we obtain rates that, despite not being optimal (the rate from Corollary 17 should be compared to the optimal rate $O(m^{-s/d})$ for Sobolevs that will be achieved with ARLS sampling below), are already faster-than-i.i.d. and obtained at a really contained computational cost. The rate goes up to order $O(\log(m)/m)$ when γ goes to zero, which corresponds to what we get when the spectrum of the covariance C decays exponentially, as formalized in the next corollary.

Corollary 18 (Exponential decay)**.** *Under the assumptions of Theorem 15 and Assumption 7, taking $m := \sqrt{n} \log(\sqrt{n} c_4)$ where c_4 is a constant, it holds*

$$\mathcal{E}(\mathcal{H}, \mathrm{I}_{\tilde{X},w}) = O\left(\frac{\log(m)}{m}\right).$$

The expresion of c_4 is provided in the proof, and this corollary holds for instance for the Gaussian kernel with a subgaussian probability distribution. Although not being optimal, these rates are nonetheless interesting because they still adapt to the spectral decay of the covariance operator, and thus outperform the standard $O(m^{-1/2})$ Monte-Carlo rate. We also stress that uniform sampling is, obviously, computationally extremely efficient - the overall complexity becoming then dominated by the cost of computing the quadrature weights. We will now show that improved rates can be obtained with leverage scores sampling.

4.3 Rates for Ridge Leverage Scores Sampling

In this section, we present quantization rates for ARLS sampling (as defined in Section 2.1). This result relies on a slightly different error decomposition w.r.t. to uniform sampling as detailed in Section D.

Theorem 19. *Let Assumptions 4 and 5 hold. Let the sub-samples $\tilde{X}_1, \dots, \tilde{X}_m$ be drawn with replacement proportionally to $(z, \lambda_0, \delta/6)$-approximate leverage scores from the dataset $\{X_1 \dots, X_n\}$, for some $z \geq 1, \lambda_0 > 0$, and w chosen as in (7). Assume $n \geq (1655 + 233 \log(12K^2/\delta))K^2$ and $\lambda_0 \leq \frac{19K^2 \log(\frac{8n}{\delta})}{n}$. Then, we have the two following results, depending on the assumption on the eigenvalue decay.*

- *Under Assumption 6 (polynomial decay), choosing $m = n^\gamma (\log \frac{32n}{\delta})^{1-\gamma} \frac{78cz^2}{(19K^2)^\gamma}$ guarantees that, with probability at least $1 - \delta$,*

$$\mathcal{E}(\mathcal{H}, \mathrm{I}_{\tilde{X},w}) = O\left(\frac{\log(m)^{1/(2\gamma)}}{m^{1/(2\gamma)}}\right) ,$$

  provided that n is large enough, i.e., $\frac{19K^2(\log \frac{32n}{\delta})}{n} \leq \min\left(\|C\|_{\mathcal{L}(\mathcal{H})}, \left(\frac{c_\gamma z^2}{5}\right)^{1/\gamma}\right)$ where $c_\gamma := a_\gamma/(1-\gamma)$ when $\gamma < 1$ and $c_\gamma := K^2$ when $\gamma = 1$.
- *Under Assumption 7 (exponential decay), choosing*

$$m = \max(334, 78z^2\beta^{-1}) \log\left(\max(\tfrac{2a_\beta}{19K^2}, \tfrac{48}{\delta})n\right)^2$$

  guarantees that, with probability at least $1 - \delta$,

$$\mathcal{E}(\mathcal{H}, \mathrm{I}_{\tilde{X},w}) = O\left(\frac{m^{1/4}}{\exp(\sqrt{m}/c)}\right) ,$$

  where c is a constant, provided that n is large enough: $\frac{19K^2 \log(\frac{8n}{\delta})}{n} \leq \min(a_\beta, \|C\|_{\mathcal{L}(\mathcal{H})})$.

We stress that the constant c appearing in the rate for the exponential decay setting is independent on the dimension. As one can see from the rates, ARLS sampling allows us to reach better rates both for polynomial and exponential decay. Again, the Sobolev case corresponds to a polynomial decay of the eigenvalues with $\gamma = d/(2s) < 1$, and we thus obtain the rate $\mathcal{E}(\mathrm{H}^s(\mathcal{X}), \mathrm{I}_{\tilde{X},w}) = O(\log(m)^{s/d} m^{-s/d})$ in this setting, which up to the logarithmic term matches the known optimal rates mentioned in Section 3.3.

Note that the condition on λ_0 can be satisfied by directly feeding the desired value to the algorithm used to estimate the approximate empirical leverage scores, and should therefore not be seen as a limitation.

4.4 Faster Rates Under a Source Condition

While previous rates were uniform over the RKHS $\mathcal{H}$, it is possible to obtain improved quadrature rates when considering fractional subspaces, i.e. nested subspaces of $\mathcal{H}$ of increasing smoothness. To our knowledge, this setting has never been studied in the literature so far.

Definition 20 (Fractional Subspaces)**.** *If $\mathcal{H}$ is an RKHS with covariance operator C, the fractional subspace of smoothness s of $\mathcal{H}$ for the data distribution ρ is defined as $\mathcal{H}^s_\rho = C^s\mathcal{H}$, and is endowed with the norm $\|f\|_s = \|g\|$ where g is the unique function satisfying $g \in (\ker C)^\perp$ and $C^s g = f$.*

Note that this definition depends on both $\mathcal{H}$ and ρ, i.e. not only on the properties of the base RKHS but also on its interaction with the data distribution. It is connected to the source condition hypothesis made in the inverse problem literature (Engl et al., 2000); the difference in our setting is that we are not interested in one single function, but rather in bounding the quadrature error uniformly over such fractional subspaces.

The fractional subspaces are themselves reproducing kernel Hilbert spaces and one could apply the previous result directly to them and define their associated kernels. However, in practice the smoothness is often unknown, and we obtain in this section improved rates without the need to estimate this smoothness: in particular the leverage scores are computed with respect to the base kernel κ.

Such improved rates are also reminiscent of the so-called superconvergence results in kernel interpolation, see e.g. Schaback (2018) and Wendland (2004, Sec. 11.5).

Theorem 21. *Let $s \in [0, 1/2]$. Let Assumption 5 hold. Furthermore, assume that the data points $X_1, \dots, X_n$ are drawn i.i.d. from the distribution ρ and that $m \leq n$ sub-samples $\tilde{X}_1, \dots, \tilde{X}_m$ are drawn using $(z, \lambda_0, \delta/4)$-approximate leverage scores sampling with replacement (for some $z \geq 1, \lambda_0 > 0$) from the dataset $\{X_1 \dots, X_n\}$. Let w chosen as in (7). Assume that:*

$$n \geq (1655 + 233 \log(8K^2/\delta))K^2$$
$$\lambda_0^{2s+1} \leq \frac{19K^2 \log(32n/\delta)}{n} \leq \min(1, \|C\|^{2s+1}_{\mathcal{L}(\mathcal{H})}).$$

- *Under Assumption 6 (polynomial decay), taking $m = \Theta\Big(n^{\gamma/(2s+1)} \log(32n/\delta)^{1-\gamma/(2s+1)}\Big)$, we get with probability $1-\delta$ the rate*

$$\mathcal{E}(\mathcal{H}^s_\rho, \mathrm{I}_{\tilde{X},w}) = O(m^{-(2s+1)/(2\gamma)})$$

  provided that n is large enough to additionally ensure $n \geq 19K^2\Big(\frac{334}{78z^2c_\gamma}\Big)^{(2s+1)/\gamma} \log\big(32n/\delta\big)$.

- *Under Assumption 7 (exponential decay), taking*

$$m := \max\left(\frac{c_m}{2s+1}\log\big(c'_m n\big), 334\right)\log\big(c'_m n\big) = O(\log(n)^2)$$
$$\text{where} \quad c_m := 78z^2\beta^{-1}, c'_m := \max\left(\frac{(2a_\beta)^{2s+1}}{19K^2}, \frac{32}{\delta}\right)$$

  it holds with probability $1-\delta$

$$\mathcal{E}(\mathcal{H}^s_\rho, \mathrm{I}_{\tilde{X},w}) = O\left(m^{1/4}\exp\Big(-\tfrac{2s+1}{2\sqrt{c_m}}\sqrt{m}\Big)\right),$$

  provided that n is large enough to additionally ensure $n \geq 19K^2 a_\beta^{-(2s+1)} \log(32n/\delta)$.

Note that depending on the constants, the conditions on n might always be satisfied, or reduce to lower bounds on n, but can always be satisfied for n large enough.

We observe under the polynomial decay assumption an improved rate of $O(m^{-(2s+1)/(2\gamma)})$, which should be compared to the rate $O(m^{-1/(2\gamma)})$ that we obtained (up to log terms) in Section 4.3. In the exponential decay setting, we still obtain an exponential dependence in $\sqrt{m}$, however the constant appearing inside the exponential is reduced due to the factor $2s+1$ and faster convergence can hence be obtained.

5. Numerical Experiments

In this section, we evaluate empirically the performance of our proposed method in two different setting. In Section 5.1, we consider periodic Sobolev spaces on $[0,1]$ and a uniform target distribution, a setting which has been extensively used to benchmark quadrature methods, and in Section 5.2 we use real datasets on $\mathbb{R}^d$ and consider spaces generated by Gaussian and Laplacian kernels.

Error computation Note that the (squared) error of a quadrature rule $\mathrm{I}_{\tilde{X},w}$ for the reproducing kernel Hilbert space $\mathcal{H}$ can be computed using Lemma 11 as follows:

$$\begin{aligned}\mathcal{E}(\mathcal{H}, \mathrm{I}_{\tilde{X},w})^2 &= \left\| \int \phi(x) \,\mathrm{d}\rho(x) - \sum_{j=1}^{m} w_j \phi(\tilde{X}_j) \right\|^2 \\ &= \iint \kappa(x,y) \,\mathrm{d}\rho(x) \,\mathrm{d}\rho(y) - 2 \sum_{1 \le j \le m} w_j \int \kappa(x, \tilde{X}_j) \,\mathrm{d}\rho(x) + w^T K_m w \end{aligned} \tag{14}$$

where we recall that K_m denotes the kernel matrix at the landmarks $\tilde{X}$. Hence, to compute the kernel mean embedding one only needs a closed form of the kernel κ and the Nyström landmarks, but to compute the error via (14) one needs a closed form for $\int \kappa(x, \tilde{X}_i) \,\mathrm{d}\rho(x)$ and $\iint \kappa(x,y) \,\mathrm{d}\rho(x) \,\mathrm{d}\rho(y)$. If ρ has a discrete support of size n, then evaluating the error requires only kernel evaluations and scales in $\Theta(n^2)$. For this reason, we restrict ourselves in Section 5.2 to datasets of moderate size, although the quadrature methods themselves do not suffer from this quadratic dependency in the dimension and could scale to larger datasets.

Optimal weights are always used in this section, for all landmark selection strategies.

5.1 Periodic Sobolev Spaces

We consider $\mathcal{X} = [0,1]$ and the translation-invariant kernel

$$\kappa_s(x,y) := 1 + 2 \sum_{n \in \mathbb{N}^*} \tfrac{1}{n^{2s}} \cos(2\pi n(x-y)) = 1 + \frac{(-1)^{s-1}(2\pi)^{2s}}{(2s)!} B_{2s}(\{x-y\})$$

where B_{2s} denotes the Bernoulli polynomial of order $2s$ and $\{\cdot\}$ the fractorial part. The expression involving Bernoulli polynomials is for instance mentioned in (Wahba, 1990, p.22). The associated reproducing kernel Hilbert space corresponds to the Sobolev space of periodic functions of order s satisfying the boundary conditions $f^{(i)}(0) = f^{(i)}(1)$ for $i = 0, \dots, s-1$, and we choose for ρ the uniform distribution on $\mathcal{X}$.

It holds $\int_0^1 \kappa(x, \tilde{x}) \,\mathrm{d}x = \iint_0^1 \kappa(x,y) \,\mathrm{d}x \,\mathrm{d}y = 1$ so the error can easily be computed using Equation (14). This RKHS has been used by multiple authors to benchmark quadrature methods because the eigendecomposition of the covariance operator is computable exactly, and we thus include this setting for completeness. However, we stress that the kernel mean embedding is the constant function $\mu(x) = \int_0^1 \kappa(x,y) \,\mathrm{d}y = 1$ (using the definition of the kernel as sum of cosines), and the continuous ridge leverage scores (of which the the leverage scores defined in (5) can be seen as a tractable approximation based on the empirical data)

are uniform in this setting as observed by Bach (2017, Sec. 4.4). As a consequence, no improvement over uniform sampling should be expected in this setting when using ARLS.

For $\mathcal{X} = [0,1]^d$ with $d > 1$, we consider the product kernel $\kappa_s^d(x,y) = \Pi_{i=1}^d \kappa_s(x_i, y_i)$. This kernel does not induce a Sobolev space but rather consists in functions having square integrable mixed partial derivatives of order up to s in each variable. The eigenvalues of the associated integral operator for the uniform distribution are known to decay in $(\log i)^{2s(d-1)} i^{-2s}$, see e.g. Bach (2017).

We compare our approach to the method of (Belhadji et al., 2019) based on determinantal point processes sampling, as well as the method of (Hayakawa et al., 2022) which relies like us on a Nyström approximation but uses a recombination algorithm. We also include for comparison three greedy deterministic methods: greedy minimization of the norm of the residual $\|P_m^\perp \hat{\mu}_n\|$, orthogonal matching pursuit, and greedy maximization of $\det(K_m)$. Note that these three methods correspond in the kernel interpolation literature respectively to the so-called f/P-greedy, f-greedy and P-greedy methods applied on the function $\hat{\mu}_n$. For these methods, the non-convex optimization steps to select the new atoms are approximated by an exhaustive search over the empirical data. We provide additional details regarding these methods in Section H.2.

We implemented our approach as well as the three greedy methods in Julia[1], and rely on the Python authors' implementations of the other two methods. All implementations however use OpenBLAS as BLAS implementation with the same number of threads, see Section H.1 for technical details.

Results are reported in Figure 1 for $d = 1, s = 1$ and $d = 2, s = 3$. We observe that all methods seem to roughly follow the optimal $O(m^{-s})$ rate in dimension $d = 1$. This is expected for our method by Theorem 9 even though we are sampling uniformly, given that leverage scores are uniform in this setting. For $d = 2$, all methods appear to be slightly sub-optimal compared to the optimal theoretical rate, which is still $O(m^{-s})$ in this setting as discussed above. Although our method seems to suffer from a slightly larger error with respect to other methods for a fixed support size m , it outperforms all of them when looking at the tradeoff between approximation error and runtime. In particular, the three greedy methods suffer a lot from the linear search which is done at each iteration. The method from (Belhadji et al., 2019) is competitive with our approach in terms of accuracy-runtime tradeoff for $d = 1, s = 1$, but requires the knowledge of the covariance's eigendecomposition which is highly limiting for applications beyond this setting. Greedy maximization of $\det(K_m)$ seems to yield a better convergence rate than our method at a moderate computational cost, however this method is not adaptive to the target distribution and we will show in the following experiment that it performs poorly for a non-uniform distribution.

5.2 OpenML Datasets with Gaussian and Laplacian Kernels

We consider in this section multiple machine learning datasets from the OpenML database[2]. To better see the rates of the different methods, we do not use data splitting and report

1. See `https://gitlab.com/achatali/efficient-numerical-integration-in-rkhs-via-ls-sampling`, code released under the AGPL3 license.
2. `https://www.openml.org/`

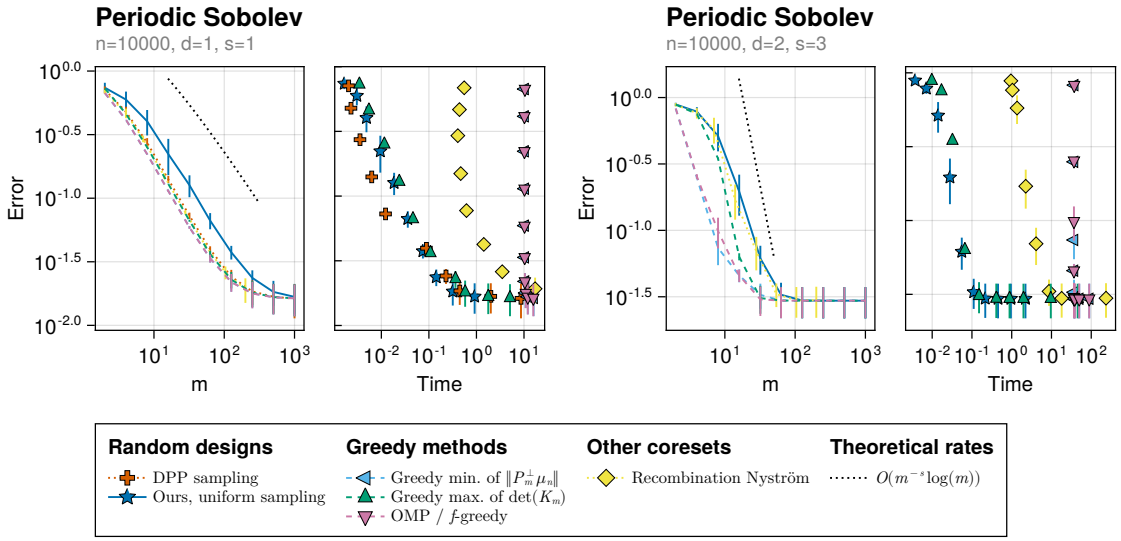

Figure 1: Periodic Sobolev setting. Medians and standard deviations over 50 trials.

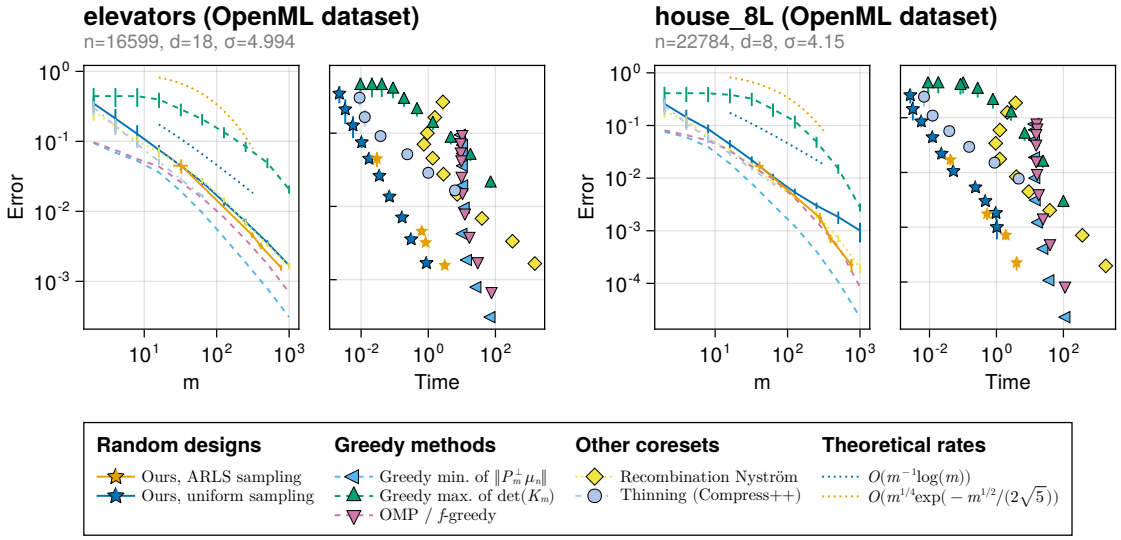

Figure 2: Gaussian kernel for two OpenML datasets. Medians and standard deviations over 40 trials.

the error computed using (14) taking ρ to be the discrete measure corresponding to the full dataset $\rho := \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)$.

We report here the error as a function of both the number of nodes m and running times, for the Gaussian (Figure 2) and Laplacian (Figure 3) kernels and for two datasets, but additional results on a wider selection of datasets are provided in Section H.3. The kernel scale is fixed by computing the median inter-point euclidean distance on a random subset of the data, and its value is reported on the figures for each dataset. We compare our methods to the algorithms mentioned in Section 5.1, at the exception of the methods which rely on the Mercer decomposition, as the latter is unknown in this setting. We also include the thinning method of (Shetty et al., 2022), for which we take as oversampling parameter $g = 4$, which corresponds to the author's choice in their experimentations, and start building the coreset from $m^2 \leq n$ samples drawn iid and uniformly from the dataset. Additional technical details are provided in Section H.1.

In Figure 2, we plot in dotted line the theoretical rates predicted by Theorem 9 under Assumption 7, picking for the exponential rate a constant matching the observations. We observe that on these two datasets uniform sampling indeed yields a fast $O(m^{-1})$ rate. Leverage scores sampling improves the converge rate as predicted by theory, however this is observed in practice only when $m \geq 100$; it should be noted however in this setting that (i) the tails of the target distribution might be too heavy to satisfy the hypotheses and (ii) the exponential decay is conditioned in Theorem 19 to having $n = \exp(m^{1/2})$, which is not satisfied for the larger values of m used in the plot as computing the error exactly would become prohibitive on very large datasets.

Here again, when looking at the error as a function of runtime, we see that our approach outperforms all the others algorithms. It is clear that the method which greedily fills the space in a uniform manner, which seemed to be competitive in the Sobolev setting, yields here a really poor accuracy; this should be expected as this method is not adaptive to the target distribution.

With a Laplacian kernel $\kappa(x, y) = \exp(-\lambda\|x - y\|)$, we do not observe any different between uniform and ARLS sampling, which matches our theoretical guarantees. Indeed due to the lack of smoothness of the Laplacian kernel, we expected to observe the rates for exponential decay with the weakest hypotheses (Assumption 6 with $\gamma \to 1$), which yields a rate of order $O(m^{-1/2})$ (i.e. no better than Monte-Carlo) for both uniform and ARLS sampling. All methods achieve the same rate, with slightly smaller constants for greedy methods - still at the price of a much larger computational cost.

6. Conclusion

In this article, we introduced an efficient quadrature method based on random subsampling, which is related to the Nyström approximation used for the discretization of linear integral equations and to build low-rank approximations of kernel matrices. We derived worst-case error bounds for RKHS for both uniform and approximate ridge leverage scores sampling, and showed that optimal rates can be obtained for Sobolev spaces in the latter case. Empirically, we showed that our method outperforms the state of the art in terms of accuracy-runtime tradeoff. Studying the performance of our approach in the misspecified

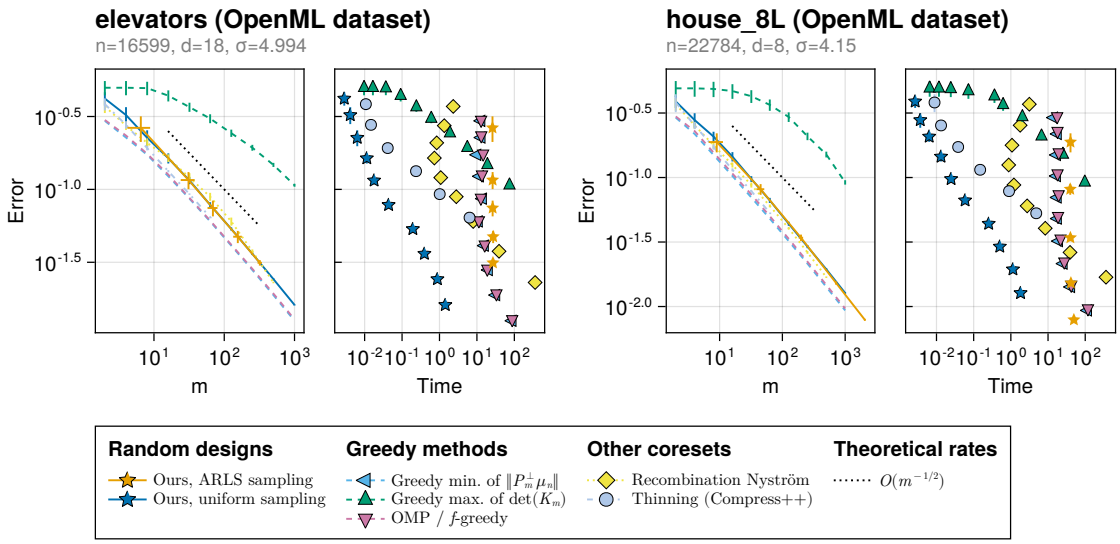

Figure 3: Laplacian kernel for two OpenML datasets. Medians and standard deviations over 30 trials.

setting, i.e. when the integrand do not belong to the considered RKHS, would be of interest for future works.

Acknowledgments

We would like to thank the anonymous referees as well as Lester Mackey and Mathias Sonnleitner for their helpful and constructive comments. Lorenzo Rosasco acknowledges the financial support of the European Research Council (grant SLING 819789), the AFOSR projects FA9550-18-1-7009, FA9550-17-1-0390 and BAA-AFRL-AFOSR-2016-0007 (European Office of Aerospace Research and Development), the EU H2020-MSCA-RISE project NoMADS - DLV-777826, and the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. The research by Ernesto De Vito and Lorenzo Rosasco has been supported by the MIUR grant PRIN 202244A7YL. The research by Ernesto De Vito has been supported by the MIUR Excellence Department Project awarded to Dipartimento di Matematica, Università di Genova, CUP D33C23001110001. Ernesto De Vito is a member of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM).

Structure of the Appendix

We begin by introducing additional notations in Section B. Then, we prove in Section C the expression of optimal weights. A deterministic error decomposition is derived in Section D, and then used in Section E to prove our main results. Section G contains the concentration results that our proof of Theorem 15 rely on, and we also recall in Section F some key results on the effective dimensions and the Nyström approximation. Eventually we provide additional details regarding numerical experiments in Section H.

Appendix A. Table of Notations

$\mathcal{X}$	Input space
$\mathcal{H}$	Generic RKHS
$\mathrm{H}^s(\mathcal{X})$	Sobolev space (see Example 1)
$\mathcal{H}^s_\rho$	Subspace of $\mathcal{H}$ corresponding to a source condition
ρ	Target/data distribution
$\tilde{X}$	Quadrature nodes (= Nyström landmarks in our case)
w	Quadrature weights
$\mathrm{I}_{\tilde{X},w}$	Quadrature rule
$C : \mathcal{H} \to \mathcal{H}$	(Uncentered) covariance operator
$(\sigma_i)_{i\in\mathbb{N}}$	Eigenvalues of C
γ, a_γ	Parameter and constant for polynomial decay (Assumption 6)
β, a_β	Parameter and constant for exponential decay (Assumption 7)
$\mathcal{E}(\mathcal{H}, \mathrm{I}_{\tilde{X},w})$	Worst-case quadrature error on the unit ball of $\mathcal{F}$ (cf. (3))

Appendix B. Additional Notations

We define the operator $\Phi : L^2(\rho) \to \mathcal{H}$ for any $f \in L^2(\rho)$ as

$$\Phi f = \int_{\mathcal{X}} f(x)\phi(x)\,\mathrm{d}\rho(x).$$

Its adjoint Φ^* is defined by $\Phi^* h = \langle h, \phi(\cdot)\rangle$ for any $h \in \mathcal{H}$ and corresponds to the inclusion operator from $\mathcal{H}$ into $L^2(\rho)$.

We define the (uncentered) covariance operator $C : \mathcal{H} \to \mathcal{H}$ as

$$C \coloneqq \int \phi(x) \otimes \phi(x) d\rho(x)$$

where $(\phi(\boldsymbol{x}) \otimes \phi(\boldsymbol{x}))(f) \coloneqq \langle f, \phi(\boldsymbol{x})\rangle\phi(\boldsymbol{x})$. One can easily check that $C = \Phi\Phi^*$. Moreover, Assumption 5 implies that the operator C is a positive trace class operator on $\mathcal{H}$ and allows to leverage tools from spectral theory. Positivy derives from the fact that $\phi(x) \otimes \phi(x)$ is positive for any x. Indeed, for any $f \in \mathcal{H}$, by applying twice Hille's theorem (Diestel and Uhl, 1977, Theorem 6) on the linear bounded (and thus closed (Kreyszig, 1989, 4.13.5 (a)))

operators $M \mapsto Mf$ and $v \mapsto \langle v, f\rangle$

$$\langle Cf, f\rangle = \left\langle \left(\int \phi(x) \otimes \phi(x) \,\mathrm{d}\rho(x) \right) f, f \right\rangle \tag{15}$$

$$= \left\langle \int \phi(x) f(x) \,\mathrm{d}\rho(x), f \right\rangle \tag{16}$$

$$= \int f(x)^2 \,\mathrm{d}\rho(x) \geq 0. \tag{17}$$

The empirical covariance operator is defined as

$$\hat{C}_n = \sum_{i=1}^{n} \phi(X_i) \otimes \phi(X_i).$$

For any operator $Q : \mathcal{H} :\to \mathcal{H}$ and any real number $\lambda > 0$, we denote by $Q_\lambda : \mathcal{H} \to \mathcal{H}$ the regularized operator $Q_\lambda = Q + \lambda I$. We denote the (Moore-Penrose) pseudo-inverse of an operator A by A^+.

Given a random variable X, we write $\operatorname{ess\,sup} X$ to denote its essential supremum.

We write $1_n \in \mathbb{R}^n$ for the n-dimensional vector of ones.

We recall the definition of the effective dimension, and also introduce the notation $d_\infty(\lambda)$:

$$d_{\mathrm{eff}}(\lambda) := \mathbf{E}_{x\sim\rho} \|C_\lambda^{-1/2}\phi(x)\|^2 = \operatorname{tr}(CC_\lambda^{-1}), \tag{18a}$$

$$d_\infty(\lambda) := \operatorname*{ess\,sup}_{x\sim\rho} \|C_\lambda^{-1/2}\phi(x)\|^2. \tag{18b}$$

It holds for any $\lambda > 0$ that $d_{\mathrm{eff}}(\lambda) \leq d_\infty(\lambda) \leq K^2/\lambda < \infty$.

Appendix C. Derivation of the Weights

This section provides a proof for the expression of the optimal weights claimed in Equation (7). For ease of exposition, let us introduce the operators

$$\Phi_m : \mathbb{R}^m \to \mathcal{H}_m,\ w \mapsto \sum_{j=1}^{m} w_j \phi(\tilde{X}_j),$$

$$\Phi_n : \mathbb{R}^n \to \mathcal{H},\ w \mapsto \sum_{i=1}^{n} w_i \phi(X_i).$$

Since, by definition, $\tilde{\mu}_m$ is the orthogonal projection of $\hat{\mu}_n$ onto the space $\mathcal{H}_m$, it can be expressed as $\tilde{\mu}_m = \Phi_m w^*$ where the weights $w^* \in \mathbb{R}^m$ minimize the mapping $w \mapsto \|\hat{\mu}_n - \Phi_m w\|^2$. Setting the gradient of this mapping to zero, we obtain that w must satisfy

$$\Phi_m^*\Phi_m w = \Phi_m^*\hat{\mu}_n.$$

The minimum norm solution of the above equation is given by $w = (\Phi_m^*\Phi_m)^+\Phi_m^*\hat{\mu}_n$ (Laub, 2004). Noting that the empirical kernel mean embedding $\hat{\mu}_n$ can be expressed as $\hat{\mu}_n = \frac{1}{n}\Phi_n 1_n$ and using the fact that $\Phi_m^*\Phi_m = K_m$, $\Phi_m^*\Phi_n = K_{mn}$, we obtain the claimed equality

$$w = K_m^+\Phi_m^*(n^{-1}\Phi_n 1_n) = \frac{1}{n} K_m^+ K_{mn} 1_n.$$

Appendix D. Deterministic Error Bound

In order to break down the approximation error, we introduce the quantity

$$\hat{\mu}_m = \frac{1}{m}\sum_{j=1}^{m} \phi(\tilde{X}_j) \in \mathcal{H}_m,$$

which is an unbiased estimate of the empirical kernel mean embedding $\hat{\mu}_n$ when sampling uniformly the landmarks.

Our main results rely on the following deterministic error decompositions.

Lemma 22 (Error decomposition)**.** *For any $\lambda > 0$, it holds (almost surely)*

$$\|\mu - \tilde{\mu}_m\| \leq \|\mu - \hat{\mu}_n\| + \|P_m^{\perp} C_\lambda^{1/2}\|_{\mathcal{L}(\mathcal{H})} \|C_\lambda^{-1/2}(\hat{\mu}_n - \hat{\mu}_m)\| \tag{19}$$

$$\|\mu - \tilde{\mu}_m\| \leq \|\mu - \hat{\mu}_n\| + \|P_m^{\perp} C_\lambda^{1/2}\|_{\mathcal{L}(\mathcal{H})}. \tag{20}$$

While the decomposition (19) is convenient, it is not well suited for the analysis when sampling proportionally to leverage scores as described in Section 2.1, and we will see that the decomposition (20) is easier to work with in this setting.

Proof We rely for both inequalities on the decomposition

$$\|\mu - \tilde{\mu}_m\| \leq \|\mu - \hat{\mu}_n\| + \|\hat{\mu}_n - \tilde{\mu}_m\|$$

First bound (19) Note that

$$\|\hat{\mu}_n - \tilde{\mu}_m\| = \|P_m^{\perp}\hat{\mu}_n\| = \|P_m^{\perp}(\hat{\mu}_n - \hat{\mu}_m)\|$$

where the last inequality follows from $P_m^{\perp}\hat{\mu}_m = 0$. Hence we get

$$\begin{aligned}\|\mu - \tilde{\mu}_m\| &\leq \|\mu - \hat{\mu}_n\| + \|P_m^{\perp}(\hat{\mu}_n - \hat{\mu}_m)\| \\ &\leq \|\mu - \hat{\mu}_n\| + \|P_m^{\perp} C_\lambda^{1/2}\|_{\mathcal{L}(\mathcal{H})} \|C_\lambda^{-1/2}(\hat{\mu}_n - \hat{\mu}_m)\|.\end{aligned}$$

Second bound (20) We use the alternative decomposition

$$\begin{aligned}\|\mu - \tilde{\mu}_m\| &= \|\mu - P_m\hat{\mu}_n\| \\ &\leq \|\mu - P_m\mu\| + \|P_m(\mu - \hat{\mu}_n)\| \\ &\leq \|P_m^{\perp} C_\lambda^{1/2}\|_{\mathcal{L}(\mathcal{H})} \|C_\lambda^{-1/2}\mu\| + \|\mu - \hat{\mu}_n\|\end{aligned}$$

Note that because μ is a mean embedding, it can be written $\mu = \Phi 1$ where $1 \in L^2(\rho)$ denotes the constant function, and Φ admits a polar decomposition of the form $\Phi = C^{1/2}U$ where U is a partial isometry from $L^2(\rho)$ to $\mathcal{H}$. Hence we have

$$\|C_\lambda^{-1/2}\mu\| = \|C_\lambda^{-1/2} C^{1/2} U 1\| \leq \|C_\lambda^{-1/2} C^{1/2}\|_{\mathcal{L}(\mathcal{H})} \|1\|_{L^2(\rho)} \leq 1.$$

■

Appendix E. Proofs of the Main Results

E.1 Proofs for Uniform Sampling (Section 4.2)

Theorem 15 is a consequence of a more general result which we state now.

Theorem 23. *Let Assumption 5 hold. Furthermore, assume that the data points $X_1, \dots, X_n$ are drawn i.i.d. from the distribution ρ and that $m \leq n$ sub-samples $\tilde{X}_1, \dots, \tilde{X}_m$ are drawn uniformly with replacement from the dataset $\{X_1 \dots, X_n\}$. Then, for any $\lambda \in]0, \|C\|_{\mathcal{L}(\mathcal{H})}]$ and $\delta \in]0,1[$, with probability at least $1-\delta$*

$$\|\mu - \tilde{\mu}_m\| \leq \frac{2K\sqrt{2\log(6/\delta)}}{\sqrt{n}} + \sqrt{\lambda}\left(\frac{4\sqrt{3d_\infty(\lambda)}\log(12/\delta)}{m} + 6\sqrt{\frac{d_{\text{eff}}(\lambda)\log(12/\delta)}{m}}\right),$$

provided that

- $m \geq \max(67, 5d_\infty(\lambda))\log\left(\frac{12K^2}{\lambda\delta}\right)$,
- $\lambda n \geq 12K^2\log(12/\delta)$.

Proof Let $\delta \in (0,1)$ be the desired confidence level. Let $\lambda > 0$, $m \in \mathbb{N}$ and $n \in \mathbb{N}$ satisfy the conditions of the theorem. Using the error decomposition of Lemma 22, we get

$$\|\mu - \tilde{\mu}_m\| \leq \|\mu - \hat{\mu}_n\| + \|P_m^\perp C_\lambda^{1/2}\|_{\mathcal{L}(\mathcal{H})}\|C_\lambda^{-1/2}(\hat{\mu}_n - \hat{\mu}_m)\|.$$

Controlling the first term amounts to measuring the concentration of an empirical mean around its true mean in a Hilbert space. Multiple variants of such results can be found in the literature (see, e.g., (Pinelis, 1994)). We apply here Lemma 31 on the random variables $\eta_i \coloneqq \phi(X_i) - \mu, 1 \leq i \leq n$. Note that they are indeed bounded since, for any index $1 \leq i \leq n$, $\|\eta_i\| \leq 2\sup_{x\in\mathcal{X}}\|\phi(x)\| = 2K$. Thus, it holds with probability at least $1 - \delta/3$ on the draw of the the dataset $X_1, \dots, X_n$ that

$$\|\mu - \hat{\mu}_n\| \leq \frac{2K\sqrt{2\log(6/\delta)}}{\sqrt{n}}.$$

Next, we rely on Lemma 28 to bound the term $\|P_m^\perp C_\lambda^{1/2}\|_{\mathcal{L}(\mathcal{H})}$ with high probability. Since the Nyström landmarks are uniformly drawn and $m \geq \max(67, 5d_\infty(\lambda))\log\frac{12K^2}{\lambda\delta}$, we have, for any $\lambda > 0$, with probability at least $1 - \delta/3$ on the draw of the landmarks $\tilde{X}_1, \dots, \tilde{X}_m$,

$$\|P_m^\perp C_\lambda^{1/2}\|_{\mathcal{L}(\mathcal{H})} \leq \sqrt{3\lambda}.$$

Finally, the last term can be bounded using Lemma 35 which implies that, since λ satisfies $0 < \lambda \leq \|C\|_{\mathcal{L}(\mathcal{H})}$ and $\lambda n \geq 12K^2\log(4/\delta)$, it holds with probability at least $1 - \delta/3$

$$\left\|C_\lambda^{-1/2}(\hat{\mu}_n - \hat{\mu}_m)\right\| \leq \frac{4\sqrt{d_\infty(\lambda)}\log(12/\delta)}{m} + \sqrt{\frac{12d_{\text{eff}}(\lambda)\log(12/\delta)}{m}}.$$

Taking the union bound over the three events yields the desired result: with probability at least $1-\delta$ (over all sources of randomness), it holds that

$$\|\mu - \tilde{\mu}_m\| \leq \frac{2K\sqrt{2\log(6/\delta)}}{\sqrt{n}} + \sqrt{3\lambda}\left(\frac{4\sqrt{d_\infty(\lambda)}\log(12/\delta)}{m} + \sqrt{\frac{12d_{\text{eff}}(\lambda)\log(12/\delta)}{m}}\right).$$

■

Proof Assuming that our choice of m and λ satisfies the constraints

$$\begin{cases} m \geq \max(67, 5d_\infty(\lambda)) \log \frac{12K^2}{\lambda\delta} \\ \lambda n \geq 12K^2 \log(12/\delta) \\ 0 < \lambda \leq \|C\|_{\mathcal{L}(\mathcal{H})} \end{cases}, \tag{21}$$

we can apply Theorem 23 and use the fact that $d_\infty(\lambda) \leq K^2/\lambda$ to get

$$\|\mu - \tilde{\mu}_m\| \leq \frac{2K\sqrt{2\log(6/\delta)}}{\sqrt{n}} + \frac{4\sqrt{3}K\log(12/\delta)}{m} + 6\sqrt{\log(12/\delta)}\sqrt{\frac{\lambda d_{\text{eff}}(\lambda)}{m}}.$$

Setting $\lambda = \frac{12K^2\log(m/\delta)}{m}$ we obtain by Lemma 11 the claimed result with constants $c_1 = 2K\sqrt{2\log(6/\delta)}$, $c_2 = 4\sqrt{3}K\log(12/\delta)$, and $c_3 = 12\sqrt{3\log(12/\delta)}K$.

Let us now check that our choices are consistent with the constraints. We will also obtain a more user-friendly expression for the constraints and express the sub-sample size m as a function of the sample size n. Using the fact that $d_\infty(\lambda) \leq K^2/\lambda$, one can easily check that a sufficient set of conditions to satisfy (21) is given by

$$\begin{cases} m \geq 67 \log\left(\frac{1}{\delta}\frac{m}{\log(m/\delta)}\right) \\ m \geq \frac{5m}{12\log\left(\frac{m}{\delta}\right)} \log\left(\frac{1}{\delta}\frac{m}{\log(m/\delta)}\right) \\ \frac{\log(12/\delta)}{n} \leq \frac{\log(m/\delta)}{m} \\ 12K^2\frac{\log(m/\delta)}{m} \leq \|C\|_{\mathcal{L}(\mathcal{H})} \end{cases}.$$

As $m \leq n$, the third condition is satisfied as soon as $m \geq 12$. Moreover, with this choice of m, we have $\log(m/\delta) > 1$, hence the second constraint always holds and it is sufficient to show that

$$m \geq \max(67, 12K^2\|C\|^{-1}_{\mathcal{L}(\mathcal{H})}) \log\left(\frac{m}{\delta}\right).$$

■

Proof Under Assumption 6, by Lemma 26 it holds $d_{\text{eff}}(\lambda) \leq c_\gamma \lambda^{-\gamma}$. Under the assumptions of Theorem 15, setting $m := n^{1/(2-\gamma)}\log(n/\delta)$, we get

$$\mathcal{E}(\mathcal{H}, \mathrm{I}_{\tilde{X},w}) = \|\mu - \tilde{\mu}_m\| \leq \frac{c_1}{\sqrt{n}} + \frac{c_2}{m} + \frac{c_3\sqrt{\log(m/\delta)}}{m}\sqrt{d_{\text{eff}}\left(\frac{12K^2\log(m/\delta)}{m}\right)} \tag{22}$$

$$\leq \frac{c_1}{\sqrt{n}} + \frac{c_2}{m} + c_3\sqrt{c_\gamma}(12K^2)^{-\gamma/2}\frac{\log(m/\delta)^{\frac{1-\gamma}{2}}}{m^{\frac{2-\gamma}{2}}} \tag{23}$$

$$= O\left(\frac{\log(m/\delta)^{1-\frac{\gamma}{2}}}{m^{1-\frac{\gamma}{2}}}\right) \tag{24}$$

■

Proof Under Assumption 7, it holds by Lemma 27 $d_{\text{eff}}(\lambda) \leq \log(1 + a_\beta/\lambda)/\beta$. We apply Theorem 15. Taking $m \geq \frac{12K^2 \log(m/\delta)}{a_\beta}$, and using the fact that $\log(1 + x) \leq \log(2x)$ for $x > 1$, the last term of (13) can be bounded by

$$\begin{aligned}\frac{\sqrt{\log(m/\delta)}}{\sqrt{\beta}m}\sqrt{\log\left(1 + \frac{a_\beta m}{12K^2\log(m/\delta)}\right)} &\leq \frac{1}{\sqrt{\beta}m}\sqrt{\log(m/\delta)\log\left(\frac{a_\beta m}{6K^2\log(m/\delta)}\right)} \\ &\leq \frac{1}{\sqrt{\beta}m}\log(m\max(1/\delta, a_\beta/(6K^2)))\end{aligned}$$

which is bounded by $\frac{1}{\sqrt{\beta n}}$ by taking $m := \sqrt{n}\log(\sqrt{n}c_4)$ with $c_4 := \max(1/\delta, a_\beta/(6K^2))$. Plugging the latter bound in (13), we obtain

$$\mathcal{E}(\mathcal{H}, \mathrm{I}_{\tilde{X},w}) \leq \frac{c_1}{\sqrt{n}} + \frac{c_2}{\sqrt{n}\log(\sqrt{n}\max(1/\delta, a_\beta/(6K^2)))} + \frac{c_3}{\sqrt{\beta n}} = O\left(\frac{1}{\sqrt{n}}\right).$$

The claimed quantization rate follows

$$\frac{1}{\sqrt{n}} \leq \frac{\log(c_4\sqrt{n}\log(c_4\sqrt{n})))}{\sqrt{n}\log(c_4\sqrt{n})} = O\left(\frac{\log(m)}{m}\right).$$

■

E.2 Proofs for Leverage Scores Sampling (Section 4.3)

Theorem 24. *Let Assumptions 4 and 5 hold. Let $\delta \in (0, 1)$. Let the m sub-samples $\tilde{X}_1, \ldots, \tilde{X}_m$ be drawn according to $(z, \lambda_0, \delta/4)$-approximate leverage scores from the dataset $\{X_1 \ldots, X_n\}$ for some $z \geq 1$ and $\lambda_0 > 0$. Then, for any $\lambda \in (\lambda_0 \vee \frac{19K^2}{n}\log(\frac{8n}{\delta}), \|C\|_{\mathcal{L}(\mathcal{H})}]$, it holds, with probability at least $1 - \delta$,*

$$\|\mu - \tilde{\mu}_m\| \leq \frac{2K\sqrt{2\log(4/\delta)}}{\sqrt{n}} + \sqrt{3\lambda}\ ,$$

provided that

- $n \geq (1655 + 233\log(8K^2/\delta))K^2$;
- $m \geq \max\big(334, 78z^2 d_{\text{eff}}(\lambda)\big)\log\frac{32n}{\delta}$.

Proof Let the assumptions of the theorem hold. Let $\delta \in (0, 1)$ be the desired confidence level. Let the integers $m \in \mathbb{N}$ and $n \in \mathbb{N}$ satisfy the conditions of the theorem and let $\lambda \in (\lambda_0 \vee \frac{19K^2}{n}\log(\frac{8n}{\delta}), \|C\|_{\mathcal{L}(\mathcal{H})}]$. Recall the error decomposition from Equation (20),

$$\|\mu - \tilde{\mu}_m\| \leq \|\mu - \hat{\mu}_n\| + \|P_m^\perp C_\lambda^{1/2}\|_{\mathcal{L}(\mathcal{H})}.$$

We apply here Lemma 31 on the random variables $\eta_i := \phi(X_i) - \mu, 1 \leq i \leq n$. Note that they are indeed bounded since, for any index $1 \leq i \leq n$, $\|\eta_i\| \leq 2\sup_{x\in\mathcal{X}}\|\phi(x)\| = 2K$.

Thus, it holds with probability at least $1 - \delta/2$ on the draw of the the dataset $X_1, \dots, X_n$ that

$$\|\mu - \hat{\mu}_n\| \leq \frac{2K\sqrt{2\log(4/\delta)}}{\sqrt{n}} \ .$$

Next, we rely on Lemma 29 to bound the term $\|P_m^{\perp} C_\lambda^{1/2}\|_{\mathcal{L}(\mathcal{H})}$ with high probability. Since the sub-samples are drawn according to $(z, \lambda_0, \delta/4)$-approximate leverage scores from the full dataset $\{X_1, \dots, X_n\}$, we have, with probability at least $1 - \delta/2$ on the draw of the sub-samples $\tilde{X}_1, \dots, \tilde{X}_m$,

$$\|P_m^{\perp} C_\lambda^{1/2}\|_{\mathcal{L}(\mathcal{H})} \leq \sqrt{3\lambda} \ .$$

Taking the union bound over the two events yields the claimed result. ■

We now justify how the parameters λ, m are chosen to yield the result claimed in Theorem 19.

Proof We apply Theorem 24, use the fact that $d_{\text{eff}}(\lambda) \leq d_\infty(\lambda) \leq K^2/\lambda$ to get (without hypotheses on the eigenvalues decay)

$$\|\mu - \tilde{\mu}_m\| \leq \frac{2K\sqrt{2\log(4/\delta)}}{\sqrt{n}} + \sqrt{3\lambda} \tag{25}$$

We now need to pick m and λ that ensure

$$\begin{cases} \lambda_0 < \lambda \leq \|C\|_{\mathcal{L}(\mathcal{H})} & \text{(26a)} \\ \lambda \geq \frac{19K^2}{n} \log(\frac{8n}{\delta}) & \text{(26b)} \\ m \geq 334 \log\left(\frac{32n}{\delta}\right) & \text{(26c)} \\ m \geq 78 z^2 d_{\text{eff}}(\lambda) \log\left(\frac{32n}{\delta}\right) & \text{(26d)} \end{cases}$$

Polynomial decay Under Assumption 6, by Lemma 26 it holds $d_{\text{eff}}(\lambda) \leq c_\gamma \lambda^{-\gamma}$ for some constant $c_\gamma > 0$. In this setting, a sufficient condition to satisfy (26d) is to take

$$\lambda := \left(\frac{78 c_\gamma z^2 \log\frac{32n}{\delta}}{m}\right)^{1/\gamma}. \tag{27}$$

One can easily check that choosing additionally

$$m := n^\gamma \frac{78 c_\gamma z^2 (\log\frac{32n}{\delta})^{1-\gamma}}{(19K^2)^\gamma}, \tag{28}$$

we get

$$\lambda = \frac{19K^2 \log\frac{32n}{\delta}}{n}$$

and the sufficient conditions (26) are satisfied as long as n is large enough and λ_0 is small enough, i.e.,

$$\begin{cases} \lambda_0 \leq \dfrac{19K^2\Big(\log\frac{32n}{\delta}\Big)}{n} \leq \|C\|_{\mathcal{L}(\mathcal{H})} & \text{(29a)} \\ n^{\gamma}(\log\dfrac{32n}{\delta})^{-\gamma} \geq \dfrac{334(19K^2)^{\gamma}}{78c_{\gamma}z^2} & \text{(29b)} \end{cases}$$

In this regime, the error (25) follows the rate $O\left(\frac{\log(n)^{1/2}}{n^{1/2}}\right)$. From (28), one can observe that $\log(n) \leq \log(n\log(32n/\delta)^{(1-\gamma)/\gamma}) = \mathrm{cst} + \log(m^{1/\gamma})$ so that the error (25) also follows a quantization rate of order $O(\sqrt{\lambda}) = O\left(\frac{\log(m)^{1/(2\gamma)}}{m^{1/(2\gamma)}}\right)$ with respect to m.

Exponential decay Under Assumption 7, by Lemma 27 it holds $d_{\mathrm{eff}}(\lambda) \leq \beta^{-1}\log(1+\frac{a_\beta}{\lambda})$. Given that $\log(1+x) \leq \log(2x)$ whenever $x \geq 1$, the following conditions are sufficient to enforce (26):

$$\begin{cases} \lambda_0 \leq \lambda \leq \min(a_\beta, \|C\|_{\mathcal{L}(\mathcal{H})}) \\ \lambda \geq \frac{19K^2}{n}\log\Big(\frac{8n}{\delta}\Big) \\ m \geq 334\log\Big(\frac{32n}{\delta}\Big) \\ m \geq 78z^2\beta^{-1}\log(2\frac{a_\beta}{\lambda})\log\frac{32n}{\delta} \end{cases} \tag{30}$$

One can easily check that the choice

$$\lambda := \frac{19K^2}{n}\log\left(\frac{8n}{\delta}\right), \quad m := \max(334, 78z^2\beta^{-1}\log\Big(\tfrac{2a_\beta}{19K^2}n\Big))\log\left(\frac{32n}{\delta}\right)$$

satisfies (30) as long as

$$\max(a_\beta^{-1}, \|C\|^{-1}_{\mathcal{L}(\mathcal{H})}) \leq \tfrac{n}{19K^2\log(\frac{8n}{\delta})} \leq \lambda_0^{-1}. \tag{31}$$

With these choices of parameters, we get a rate of order $O(\sqrt{\lambda}) = O(\log(n)^{1/2}n^{-1/2})$. Moreover, if we assume for simplicity $m := c_m\log(n)^2$, this yields the quantization rate $O(\sqrt{\lambda}) = O(m^{1/4}\exp(-\sqrt{m}/c)$ with $c = 2\sqrt{c_m}$. ■

E.3 Proofs With Source Condition (Section 4.4)

Lemma 25 (Faster rate with source condition)**.** *Let Assumptions 4 and 5 hold. Let the sub-samples $\tilde{X}_1, \ldots, \tilde{X}_m$ be drawn according to $(z, \lambda_0, \delta/4)$-approximate leverage scores from the dataset $\{X_1 \ldots, X_n\}$, for some $z \geq 1$. The for any $\lambda \in [\max(\lambda_0, \frac{19K^2\log(\frac{8n}{\delta})}{n}); \|C\|_{\mathcal{L}(\mathcal{H})}]$, $\delta \in (0,1)$, $s \in [0, 1/2]$, it holds with probability at least $1-\delta$,*

$$\mathcal{E}(\mathcal{H}^s_\rho, \mathrm{I}_{\tilde{X},w}) \leq \frac{2K^{1+2s}\sqrt{2\log(6/\delta)}}{\sqrt{n}} + (3\lambda)^{s+1/2}$$

provided that

$$n \geq (1655 + 233 \log(8K^2/\delta))K^2$$
$$m \geq \max\Big(334, 78z^2 d_{\text{eff}}(\lambda)\Big) \log\big(32n/\delta\big).$$

Proof Let $g \in \mathcal{H}$ such that $\|g\| \leq 1$ and let $f = C^s g$. Using the reproducing property of the RKHS $\mathcal{H}$, the fact that the operator C^s is self-adjoint and Cauchy-Schwarz inequality, we have

$$\left| \int f \, \mathrm{d}\rho - \sum_{j=1}^{m} f(\tilde{X}_j) \right| = \left| \left\langle C^s g, \int \phi \, \mathrm{d}\rho - \sum_{j=1}^{m} w_j \phi(\tilde{X}_j) \right\rangle \right| \leq \left\| C^s(\mu - \tilde{\mu}_m) \right\| \ .$$

Using the same decomposition as in the proof of Lemma 22 for (20), we have

$$\left\| C^s(\mu - \tilde{\mu}_m) \right\| \leq \left\| C^s P_m^{\perp} \right\|_{\mathcal{L}(\mathcal{H})} \left\| P_m^{\perp} C_\lambda^{1/2} \right\|_{\mathcal{L}(\mathcal{H})} + K^{2s} \left\| \mu - \hat{\mu}_n \right\| \tag{32}$$

Since $P_m^{\perp}$ and $C^{1/2}$ are positive bounded operators (respectively as a projection, and as the square root of a positive operator, cf. Section B), it holds by Cordes inequality (Theorem 30)

$$\left\| C^s P_m^{\perp} \right\|_{\mathcal{L}(\mathcal{H})} = \left\| (P_m^{\perp})^{2s} (C^{1/2})^{2s} \right\|_{\mathcal{L}(\mathcal{H})} \leq \left\| P_m^{\perp} C^{1/2} \right\|_{\mathcal{L}(\mathcal{H})}^{2s}.$$

so that

$$\left\| C^s(\mu - \tilde{\mu}_m) \right\| \leq \left\| P_m^{\perp} C_\lambda^{1/2} \right\|_{\mathcal{L}(\mathcal{H})}^{2s+1} + K^{2s} \left\| \mu - \hat{\mu}_n \right\| \ . \tag{33}$$

To control the second term, we apply Lemma 31 on the random variables $\eta_i \coloneqq \phi(X_i) - \mu, 1 \leq i \leq n$. For any index $1 \leq i \leq n$, it holds $\|\eta_i\| \leq 2 \sup_{x \in \mathcal{X}} \|\phi(x)\| = 2K$. Thus, with probability at least $1 - \delta/2$ on the draw of the the dataset $X_1, \ldots, X_n$,

$$\left\| \mu - \hat{\mu}_n \right\| \leq \frac{2K\sqrt{2\log(4/\delta)}}{\sqrt{n}} \ .$$

Next, we rely on Lemma 29 to bound the term $\|P_m^{\perp} C_\lambda^{1/2}\|_{\mathcal{L}(\mathcal{H})}$ with high probability. Under the hypothesis of the Lemma, we have, for any $\lambda \in]0, \|C\|_{\mathcal{L}(\mathcal{H})}]$, with probability at least $1 - \delta/2$ on the draw of the landmarks $\tilde{X}_1, \ldots, \tilde{X}_m$,

$$\|P_m^{\perp} C_\lambda^{1/2}\|_{\mathcal{L}(\mathcal{H})} \leq \sqrt{3\lambda}.$$

The proof is concluded by taking the union bound over the two high-probability events on which we control the first and the second term of (33). ■

We now prove the quantization rates claimed with a source condition.

Proof By Lemma 25, we have

$$\left\| \mu - \tilde{\mu}_m \right\| \leq \frac{2K^{1+2s}\sqrt{2\log(4/\delta)}}{\sqrt{n}} + (3\lambda)^{s+1/2}.$$

We now need to pick λ, m ensuring

$$\begin{cases} m \geq \max\Big(334, 78z^2 d_{\text{eff}}(\lambda)\Big) \log\left(\dfrac{32n}{\delta}\right) & \text{(34a)} \\ \lambda \geq \frac{19K^2 \log(\frac{8n}{\delta})}{n} & \text{(34b)} \\ \lambda_0 \leq \lambda \leq \|C\|_{\mathcal{L}(\mathcal{H})}. & \text{(34c)} \end{cases}$$

We pick $\boxed{\lambda = \left(\dfrac{19K^2 \log(32n/\delta)}{n}\right)^{1/(2s+1)}}$, which is the largest choice for λ allowing to get (up to log term) a global rate of order $\Theta(n^{-1/2})$ while satisfying (34b) (by assumption it holds $\frac{19K^2 \log(32n/\delta)}{n} < 1$). Note that as soon as $s > 0$, the logarithmic term in n can be avoided provided n is large enough and one recovers exactly the optimal rate $O(n^{-1/2})$. We opt here for a unified analysis with simplified constraints at the cost of achieving only the rate $O(\log(n) n^{-1/2})$.

Condition (34c) holds as soon as

$$\lambda_0^{2s+1} \leq \frac{19K^2 \log(32n/\delta)}{n} \leq \|C\|_{\mathcal{L}(\mathcal{H})}^{2s+1}.$$

We now consider the settings of polynomial and exponential decay of the spectrum, and detail how to choose m in order to satisfy the remaining constraints (34a), which we rewrite as:

$$\begin{cases} m \geq 334 \log \dfrac{32n}{\delta} & \text{(35a)} \\ m \geq 78z^2 d_{\text{eff}}(\lambda) \log \dfrac{32n}{\delta} & \text{(35b)} \end{cases}$$

Polynomial decay Under Assumption 6, by Lemma 26 it holds $d_{\text{eff}}(\lambda) \leq c_\gamma \lambda^{-\gamma}$. We choose

$$\boxed{m := c_m \log\big(32n/\delta\big)^{1-\gamma/(2s+1)} n^{\gamma/(2s+1)} \quad \text{where} \quad c_m := 78z^2 c_\gamma (19K^2)^{-\gamma/(2s+1)}}$$

which satisfies (35b). Condition (35a) is satisfied whenever

$$n \geq \left(\tfrac{334}{c_m}\right)^{(2s+1)/\gamma} \log\big(32n/\delta\big).$$

The quantization rate can be derived by noting that

$$\begin{aligned} m^{-(2s+1)/(2\gamma)} &= \Theta\left(\left(\log\big(32n/\delta\big)^{1-\gamma/(2s+1)} n^{\gamma/(2s+1)}\right)^{-(2s+1)/(2\gamma)}\right) \\ &= \Theta\left(n^{-1/2} \log\big(32n/\delta\big)^{1/2 - \frac{2s+1}{2\gamma}}\right) \end{aligned}$$

with $\frac{2s+1}{2\gamma} \geq 1/2$. Thus get the rate

$$\mathcal{E}(\mathcal{H}_\rho^s, \mathrm{I}_{\tilde{X},w}) = \Theta(\lambda^{(2s+1)/2}) = \Theta\left(\frac{\log(n)^{1/2}}{n^{1/2}}\right) = O(m^{-(2s+1)/(2\gamma)}).$$

Exponential decay Under Assumption 7 holds, by Lemma 27 it holds $d_{\text{eff}}(\lambda) \leq \beta^{-1} \log(1+\frac{a_\beta}{\lambda})$. Using that $\log(2x) \geq \log(1+x)$ whenever $x \geq 1$, the constraint (35) is satisfied as soon as

$$\begin{cases} m \geq \dfrac{78z^2}{\beta(2s+1)} \log\left((2a_\beta)^{2s+1} \dfrac{n}{19K^2 \log(32n/\delta)} \right) \log\big(32n/\delta\big) & \text{(36a)} \\ n \geq 19K^2 a_\beta^{-(2s+1)} \log(32n/\delta) & \text{(36b)} \\ m \geq 334 \log\big(32n/\delta\big) & \text{(36c)} \end{cases}$$

We choose

$$\boxed{m := \max\left(\frac{c_m}{2s+1} \log\big(c'_m n\big), 334 \right) \log\big(c'_m n\big)}$$

$$\text{where} \quad c_m := 78z^2\beta^{-1}, c'_m := \max\left(\tfrac{(2a_\beta)^{2s+1}}{19K^2}, \tfrac{32}{\delta} \right)$$

in order to enforce both (36a) and (36c), while (36b) is satisfied by assumption. Note that with this definition, there exists $N \in \mathbb{N}$ such that for any $n \geq N$, it holds

$$m = \frac{c_m}{2s+1} \log\big(c'_m n\big)^2$$

so that asymptotically $n = \exp(\sqrt{(2s+1)m/c_m})/c'_m$, and the quantization rate can be expressed as

$$\mathcal{E}(\mathcal{H}^s_\rho, \mathrm{I}_{\tilde{X},w}) = \Theta(\lambda^{(2s+1)/2}) = \Theta\left(\frac{\log(n)^{1/2}}{n^{1/2}} \right) = O\left(m^{1/4} \exp\Big(-\tfrac{\sqrt{2s+1}}{2\sqrt{c_m}} \sqrt{m} \Big) \right).$$

■

Appendix F. Auxiliary Results

F.1 Bounds on the Effective Dimension

We now recall how the effective dimension can be bounded under any of Assumption 6 or Assumption 7.

Lemma 26 (Effective dimension, polynomial decay)**.** *Under Assumptions 5 and 6 it holds*

$$d_{\text{eff}}(\lambda) \leq c_\gamma \lambda^{-\gamma} \text{ where } c_\gamma := \begin{cases} \dfrac{a_\gamma}{1-\gamma}, & \text{if } \gamma < 1 \\ K^2, & \text{if } \gamma = 1 \end{cases}. \tag{37}$$

It is well known, see e.g. Fischer and Steinwart (2020, Lemma 11), that the existence of a constant c_γ such that the first part of (37) holds implies in return a polynomial decay of the spectrum, i.e. $\sigma_i \lesssim i^{-1/\gamma}$.

Proof The case $\gamma < 1$ is covered in (Caponnetto and De Vito, 2007, Proposition 3 with $b \to 1/\gamma$ and $\gamma \to c$). The case $\gamma = 1$ follows from the observation $d_{\text{eff}}(\lambda) \leq d_\infty(\lambda) =$

$\operatorname{ess\,sup}_{x\sim\rho}\|C_\lambda^{-1/2}\phi(x)\|^2 \leq \|C_\lambda^{-1/2}\|^2 \operatorname{ess\,sup}_{x\sim\rho}\|\phi(x)\|^2 \leq K^2/\lambda.$ ■

For the exponential decay setting (Assumption 7), we use the following result of Della Vecchia et al. (2021, Proposition 5).

Lemma 27 (Effective dimension, exponential decay)**.** *Under Assumption 7 it holds*

$$d_{\text{eff}}(\lambda) \leq \log(1 + a_\beta/\lambda)/\beta \tag{38}$$

F.2 Nyström Approximation Result

To control the term involving $P_m^\perp$, we rely on the following lemma from Rudi et. al (Rudi et al., 2015, Lemma 6).

Lemma 28 (Uniform Nyström approximation)**.** *When the set of m landmarks is drawn uniformly from all partitions of cardinality m, for any $\lambda \in]0, \|C\|_{\mathcal{L}(\mathcal{H})}]$ we have*

$$\|P_m^\perp (C + \lambda I)^{1/2}\|^2_{\mathcal{L}(\mathcal{H})} \leq 3\lambda$$

with probability at least $1-\delta$ provided

$$m \geq \max(67, 5d_\infty(\lambda)) \log \frac{4K^2}{\lambda\delta}.$$

The next lemma is a restatement of (Rudi et al., 2015, Lemma 7).

Lemma 29 (ALS Nyström approximation)**.** *Let $z \geq 1$, $\lambda_0 > 0$ and $\delta \in]0,1[$. Let $(\hat{\ell}_t(i))_{1\leq i\leq n}$ be a collection of $(z, \lambda_0, \delta/2)$-approximate leverage scores. Let $\lambda < \|C\|_{\mathcal{L}(\mathcal{H})}$, and p_λ be a probability distribution on the set of indexes $\{1,\dots,n\}$ defined as $p_\lambda(i) := \hat{\ell}_\lambda(i)/(\sum_{i=1}^n \hat{\ell}_\lambda(i))$. Let $\{i_1,\dots,i_m\}$ be a collection of indices sampled independently with replacement from p_λ, and P_m the orthogonal projection on $\mathcal{H}_m = \operatorname{span}\{\phi(\boldsymbol{x}_{i_1}),\dots,\phi(\boldsymbol{x}_{i_m})\}$. We have with probability at least $1-\delta$*

$$\|P_m^\perp (C + \lambda I)^{1/2}\|_{\mathcal{L}(\mathcal{H})} \leq \sqrt{3\lambda}\ ,$$

provided that

- $m \geq \max\big(334, 78z^2 d_{\text{eff}}(\lambda)\big) \log \frac{16n}{\delta}$;
- $n \geq (1655 + 233\log(4K^2/\delta))K^2$;
- $19K^2 \log(\frac{4n}{\delta}) \leq \lambda n$;
- $\lambda_0 \leq \lambda$.

F.3 Misc. Results

Theorem 30 (Cordes Inequality (Cordes, 1987, Lemma 5.1))**.** *Let A, B be two positive bounded linear operators on a Hilbert space $\mathcal{H}$. Then for any $s \in [0,1]$, it holds*

$$\|A^s B^s\| \leq \|AB\|^s$$

Appendix G. Concentration Inequalities

This section contains concentration results that we rely on to prove our main result. These results are standard, and we include proofs for completeness.

The first lemma provides a high-probability control on the norm of the average of bounded random variables taking values in a separable Hilbert space.

Lemma 31. *Let $X_1, \dots, X_n$ be i.i.d. random variables on a separable Hilbert space $(\mathcal{X}, \|\cdot\|)$ such that $\sup_{i=1,\dots,n}\|X_i\| \leq A$ almost surely, for some real number $A > 0$. Then, for any $\delta \in (0,1)$, it holds with probability at least $1-\delta$ that*

$$\left\| \frac{1}{n}\sum_{i=1}^{n} X_i \right\| \leq A\frac{\sqrt{2\log(2/\delta)}}{\sqrt{n}}.$$

The proof of Lemma 31 relies on (Pinelis, 1994, Theorem 3.5) which we recall now for clarity of exposition.

Lemma 32. *Let $M = (M_i)_{i\in\mathbb{N}}$ be a martingale on a $(2, D)$-smooth separable Banach space $(\mathcal{X}, \|\cdot\|)$. Define $\sum_{j=1}^{\infty} \operatorname{ess\,sup}\|M_j - M_{j-1}\|^2 \leq b_*^2$, for some real number $b_* > 0$. Then, for all $r \geq 0$,*

$$\Pr\left[\sup_{j\in\mathbb{N}}\|M_j\| \geq r\right] \leq 2\exp\left(-\frac{r^2}{2D^2b_*^2}\right).$$

We now prove Lemma 31.

Proof Since $\mathcal{X}$ is a Hilbert space, it is 2-smooth with 2-smoothness constant $D = 1$. We define the martingale $(M_n)_{n\in\mathbb{N}}$ as $M_0 = 0$, $M_k = \sum_{1\leq i\leq k} X_k$ for $1 \leq k \leq n$ and $M_k = M_n$ for $k \geq n$, so that

$$d_k \coloneqq M_k - M_{k-1} = \begin{cases} X_k, & \text{if } 1 \leq k \leq n \\ 0, & \text{otherwise} \end{cases}.$$

As a consequence, we have $\sum_{j=1}^{\infty} \operatorname{ess\,sup}\|d_j\|^2 = \sum_{j=1}^{n} \operatorname{ess\,sup}\|X_j\|^2 \leq nA^2$. Applying Pinelis' inequality (Lemma 32) with $b_*^2 = nA^2$ yields

$$\Pr\left[\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\right\| > \epsilon\right] = \Pr\big[\|M_n\| > n\epsilon\big] \leq \Pr\left[\sup_{1\leq j\leq n}\big\|M_j\big\| > n\epsilon\right] \leq 2\exp\left(-\frac{n\epsilon^2}{2A^2}\right).$$

We get the desired result by choosing $\epsilon = A\sqrt{2\log(2/\delta)}n^{-1/2}$. ∎

The next result is a Bernstein-type inequality for random vectors defined in a Hilbert space.

Lemma 33 (Bernstein inequality for Hilbert space-valued random vectors)**.** *Let $X_1, \dots, X_n$ be i.i.d. random variables in a Hilbert space $(\mathcal{H}, \|\cdot\|)$ such that*

- $\forall i \in [n], \mathbf{E}X_i = \mu$,
- $\exists \sigma > 0, \exists H > 0, \forall i \in [n], \forall p \geq 2$, $\mathbf{E}\|X_i - \mu\|^p \leq 1/2 p!\sigma^2 H^{p-2}$.

Then, for any $\delta \in]0,1[$, we have with probability at least $1-\delta$,

$$\left\| \frac{1}{n}\sum_{i=1}^{n} X_i - \mu \right\| \leq \frac{2H\log(2/\delta)}{n} + \sqrt{\frac{2\sigma^2\log(2/\delta)}{n}}.$$

Proof Fix a confidence level $\delta \in (0,1)$. Applying (Yurinsky, 1995, Theorem 3.3.4) on the i.i.d. centered random variables $\xi_i = X_i - \mu$ with $B^2 = \sigma^2 n$, we get

$$\begin{aligned}\Pr\left[\left\|\frac{1}{n}\sum_{j=1}^{n}\xi_j - \mu\right\| \geq t\right] &\leq \Pr\left[\max_{1\leq k\leq n} k\left\|\frac{1}{k}\sum_{j=1}^{k}\xi_j - \mu\right\| \geq \left(\frac{tn}{B}\right)B\right] \\ &\leq 2\exp\left(-1/2\frac{(tn)^2}{B^2}\left(1+\frac{tHn}{B^2}\right)^{-1}\right).\end{aligned}$$

The RHS of the above is smaller than δ if and only if

$$t^2n^2 - t(2Hn\log(2/\delta)) - 2B^2\log(2/\delta) \geq 0.$$

Denoting $\Delta = 4H^2n^2\log(2/\delta)^2 + 8n^2B^2\log(2/\delta) > 0$, this holds in particular if $t \geq \frac{H\log(2/\delta)}{n} + \frac{\sqrt{\Delta}}{2n^2}$, and thus a fortiori (using $\sqrt{\Delta} \leq \sqrt{4H^2n^2\log(2/\delta)^2} + \sqrt{8n^2B^2\log(2/\delta)}$) when

$$t \geq \frac{2H\log(2/\delta)}{n} + \sqrt{\frac{2\sigma^2\log(2/\delta)}{n}}.$$

■

The following lemma provides a Bernstein-type bound for the empirical mean of Hilbert space-valued centered random variables 'whitened" by regularized linear operator.

Lemma 34. *Let $X_1,\ldots,X_n$ be i.i.d. random variables taking values in a separable Hilbert space $(\mathcal{H},\langle\cdot,\cdot\rangle)$ with associated norm $\|\cdot\|$. We denote their mean by $\mu_X \coloneqq \mathbf{E}X_1$ and their covariance by $C \coloneqq \mathbf{E}[X_1\otimes X_1]$.*

Let $Q:\mathcal{H}\to\mathcal{H}$ be a linear operator. For any $\lambda > 0$ and $\delta \in]0,1[$, it holds with probability at least $1-\delta$ that

$$\left\| Q_\lambda^{-1/2}\left(\frac{1}{n}\sum_{i=1}^{n}X_i - \mu_X\right)\right\| \leq \frac{4\,\mathrm{ess\,sup}\left\|Q_\lambda^{-1/2}X_1\right\|\log(2/\delta)}{n} + \sqrt{\frac{4\,\mathrm{tr}(Q_\lambda^{-1}C)\log(2/\delta)}{n}}.$$

Proof To prove the stated result we will apply Lemma 33 on the random variables $(\zeta_i)_{1\leq i\leq n}$ defined by $\zeta_i = Q_\lambda^{-1/2}X_i$. Let $N_Q(\lambda) = \mathrm{tr}(Q_\lambda^{-1}C)$ and $N_{Q,\infty}(\lambda) \coloneqq \mathrm{ess\,sup}\left\|Q_\lambda^{-1/2}X_1\right\|$.

For any index $1\leq i\leq n$, we have $\mathbf{E}\zeta_1 = Q_\lambda^{-1/2}\mu_X$,

$$\mathrm{ess\,sup}\|\zeta_i - \mathbf{E}[\zeta_i]\| \leq 2\,\mathrm{ess\,sup}\|\zeta_i\| = 2N_\infty(\lambda)^{1/2},$$

and,

$$\begin{aligned}\mathbf{E}\|\zeta_i - \mathbf{E}[\zeta_i]\|^2 &= \operatorname{tr}(\mathbf{E}\langle \zeta_i - \mathbf{E}[\zeta_i], \zeta_i - \mathbf{E}[\zeta_i]\rangle) \\ &= \operatorname{tr}(\mathbf{E}\big[(\zeta_i - \mathbf{E}[\zeta_i]) \otimes (\zeta_i - \mathbf{E}[\zeta_i])\big]) \\ &= \operatorname{tr}(\mathbf{E}[\zeta_i \otimes \zeta_i] - \mathbf{E}\zeta_i \otimes \mathbf{E}\zeta_i) \\ &\leq \operatorname{tr}(\mathbf{E}[\zeta_i \otimes \zeta_i]) \\ &= \operatorname{tr}(Q_\lambda^{-1/2} C Q_\lambda^{-1/2}) \\ &= N_Q(\lambda).\end{aligned}$$

Moreover, for any $p \geq 2$,

$$\begin{aligned}\|\zeta_i - \mathbf{E}[\zeta_i]\|^p &\leq (\mathbf{E}\|\zeta_i - \mathbf{E}[\zeta_i]\|^2)(\operatorname{ess\,sup}\|\zeta_i - \mathbf{E}[\zeta_i]\|^{p-2}) \\ &\leq 1/2(2N_Q(\lambda))(2N_{Q,\infty}(\lambda)^{1/2})^{p-2} \\ &\leq 1/2 p!(2N_Q(\lambda))(2N_{Q,\infty}(\lambda)^{1/2})^{p-2}.\end{aligned}$$

The result follows from Lemma 33 with constants $\sigma^2 = 2N_Q(\lambda)$ and $H = 2N_{Q,\infty}(\lambda)^{1/2}$. ■

Lemma 35 is a specialization of Lemma 34 to bound the last term appearing in Lemma 22 in our setting of Nyström uniform sampling.

Lemma 35. *Assume that the $m \geq 1$ Nyström landmarks are sampled uniformly with replacement from the dataset $X_1, \dots, X_n$. If $0 < \lambda \leq \|C\|_{\mathcal{L}(\mathcal{H})}$ and $\lambda n \geq 12K^2 \log(4/\delta)$, it holds with probability at least $1 - \delta$,*

$$\left\| C_\lambda^{-1/2}(\hat{\mu}_n - \hat{\mu}_m) \right\| \leq \frac{4\sqrt{d_\infty(\lambda)}\log(4/\delta)}{m} + \sqrt{\frac{12 d_{\text{eff}}(\lambda)\log(4/\delta)}{m}}.$$

Proof Fix the desired confidence level $\delta \in (0,1)$. Let us begin by conditioning w.r.t. to the dataset $X_1, \dots, X_n$. As the landmarks are assumed to be drawn i.i.d., we can apply Lemma 34 with $Q = C$ on the i.i.d. random variables $h_j \coloneqq \phi(\tilde{X}_j), 1 \leq j \leq m$, satisfying $\mathbf{E}[h_1] = \hat{\mu}_n$, $\mathbf{E}[h_1 \otimes h_1] = \hat{C}_n$ and $\operatorname{ess\,sup}\|C_\lambda^{-1/2} h_1\|^2 \leq d_\infty(\lambda)$: it holds with probability at least $1 - \delta/2$ (over the drawing of the landmarks) that

$$\left\| Q_\lambda^{-1/2}(\mu_X - \hat{\mu}_X) \right\| \leq \frac{4\sqrt{d_\infty(\lambda)}\log(4/\delta)}{m} + \sqrt{\frac{4\operatorname{tr}(C_\lambda^{-1}\hat{C}_n)\log(4/\delta)}{m}}.$$

Then, since we assumed $\lambda \leq \|C\|_{\mathcal{L}(\mathcal{H})}$ and $\lambda n \geq 12K^2\log(4/\delta)$, Lemma 36 ensures that $\operatorname{tr}(C_\lambda^{-1}\hat{C}_n) \leq 3d_{\text{eff}}(\lambda)$ with probability at least $1 - \delta/2$ w.r.t. the dataset $X_1, \dots, X_n$.

Finally, since the drawing of dataset and that of the indexes of the landmark are independent, the claimed bound holds with probability at least $(1 - \delta/2)(1 - \delta/2) \geq 1 - \delta$. ■

The next lemma bounds the trace term involving the empirical covariance appearing in Lemma 35 by the effective dimension.

Lemma 36. *Let $\delta > 0$, $\lambda > 0$ and $n \in \mathbb{N}$ be such that $\lambda \leq \|C\|_{\mathcal{L}(\mathcal{H})}$ and $n \geq 12 d_\infty(\lambda)\log(2/\delta)$. Then it holds with probability at least $1 - \delta$ that*

$$\operatorname{tr}(C_\lambda^{-1}\hat{C}_n) \leq 3 d_{\text{eff}}(\lambda).$$

Proof Let us control the deviation of $\mathrm{tr}(C_\lambda^{-1}\hat{C}_n)$ from its expectation $d_{\mathrm{eff}}(\lambda)$. We have

$$\mathrm{tr}(C_\lambda^{-1}\hat{C}_n) - d_{\mathrm{eff}}(\lambda) = \mathrm{tr}(C_\lambda^{-1}(\hat{C}_n - C)) = \frac{1}{n}\sum_{i=1}^{n} \xi_i - \mathbf{E}[\xi_i],$$

where we define $\xi_i \coloneqq \mathrm{tr}(C_\lambda^{-1}\phi(X_i)\otimes\phi(X_i)), i = 1,\dots,n$. The random variables $\xi_i, 1 \le i \le n$, satisfy

$$|\xi_i - \mathbf{E}[\xi_i]| = \left|\mathrm{tr}(C_\lambda^{-1}(\phi(X_i)\otimes\phi(X_i) - C))\right| \le \left\|C_\lambda^{-1/2}\phi(X_i)\right\|^2 + d_{\mathrm{eff}}(\lambda) \le 2d_\infty(\lambda)$$

and

$$\mathbf{E}[(\xi_i - \mathbf{E}[\xi_i])^2] = \mathbf{E}[\xi_i^2] - (\mathbf{E}\xi_i)^2 \le \operatorname{ess\,sup}|\xi_i|\,\mathbf{E}[\xi_i] \le 2d_\infty(\lambda)d_{\mathrm{eff}}(\lambda).$$

Lemma 33 with $H = 2d_\infty(\lambda)$ and $\sigma^2 = 2d_\infty(\lambda)d_{\mathrm{eff}}(\lambda)$ ensures that with probability at least $1-\delta$,

$$|\,\mathrm{tr}(C_\lambda^{-1}\hat{C}_n) - d_{\mathrm{eff}}(\lambda)| \le \frac{4d_\infty(\lambda)\log(2/\delta)}{n} + \sqrt{\frac{4d_\infty(\lambda)d_{\mathrm{eff}}(\lambda)\log(2/\delta)}{n}}.$$

Since $\lambda \le \|C\|_{\mathcal{L}(\mathcal{H})}$, we have $d_{\mathrm{eff}}(\lambda) = \mathrm{tr}(CC_\lambda^{-1}) \ge \left\|CC_\lambda^{-1}\right\|_{\mathcal{L}(\mathcal{H})} = \frac{\|C\|_{\mathcal{L}(\mathcal{H})}}{\|C\|_{\mathcal{L}(\mathcal{H})}+\lambda} \ge 1/2$. Furthermore, using the assumption $n \ge 12d_\infty(\lambda)\log(2/\delta)$, it holds with probability at least $1-\delta$,

$$\mathrm{tr}(C_\lambda^{-1}\hat{C}_n) \le d_{\mathrm{eff}}(\lambda)\left(1 + \frac{1}{3d_{\mathrm{eff}}(\lambda)} + \sqrt{\frac{1}{3d_{\mathrm{eff}}(\lambda)}}\right) \le d_{\mathrm{eff}}(\lambda)\left(1 + \frac{2}{3} + \sqrt{\frac{2}{3}}\right) \le 2.5\mathcal{N}(\lambda).$$

■

Appendix H. Experiments

H.1 Implementation Details

Experiments in Section 5.1 have been run on a Intel(R) Core(TM) i7-7700HQ CPU @ 2.80GHz (4 cores, 8 threads) with 4 BLAS threads. Experiments in Section 5.2 have been run on a AMD EPYC 7301 16-Core Processor @ 2.20GHz (32 cores, 64 threads) with 32 BLAS threads. We did not use GPUs to make it easier to fairly compare the different methods and measure runtimes. Note however that some methods, such as the BLESS algorithm that we use to compute approximate leverage scores, have a GPU implementation and could be accelerated in this way.

The datasets can be freely downloaded from `https://www.openml.org/`, however to ensure reproducibility we provide the `CuratedDataset`[3] Julia package which takes care of downloading, preprocessing and loading the data. All datasets have been centered and reduced.

The method of Belhadji et al. (2019) is reported in Section 5.1 only in dimension $d = 1$ because the code for the setting $d > 1$ is not publicly available.

3. `https://gitlab.com/dzla/CuratedDatasets.jl`

H.2 Implementation of the Greedy Methods

In Section 5 we considered three kernel-based greedy methods to compute quadratures rules. We provide here a few details on how such methods can be implemented. Note that we do not solve the (usually non-convex) optimization problem to select the new atom on $\mathcal{X}$, but rather do an approximate exhaustive search over the data samples. For generality, we denote $f \in \mathcal{H}$ the function to approximate, although in our context we always use these methods on $f = \hat{\mu}_n$. In the following, we denote P_t the orthogonal projection on the space $\text{span}\{\phi(\tilde{X}_1), \dots, \phi(\tilde{X}_t)\}$ spanned by the features of the so-far selected landmarks, $\Phi_t = [\phi(\tilde{X}_1), \dots, \phi(\tilde{X}_t)]$ the operator induced by their features and K_t their kernel matrix. The three considered methods are the following:

- **Greedy minimization of the residual** $P_t^\perp f$, also known as the f-greedy method:

$$\tilde{X}_{t+1} := \arg\min_{x \in X} |(P_t^\perp f)(x)|.$$

  Note that as we are optimizing over the dataset here (and not e.g. over $\mathcal{X}$), this algorithm can be seen as **orthogonal matching pursuit** with the finite dictionary $\{\phi(x_1), \dots, \phi(x_n)\}$, assuming the latter is normalized for the chosen kernel (which holds for instance for translation-invariant kernels).

- **Greedily maximization of** $\det(K_m)$**.** This method is also known as the **P-greedy method** in the kernel interpolation literature as it consists in maximizing the so-called power function:

$$\tilde{X}_{t+1} := \arg\max_{x \in X} \left\| P_t^\perp \phi(x) \right\|$$

  Note however that using the formula for the determinant of block matrices,

$$\begin{aligned} \left\| P_t^\perp \phi(x) \right\|^2 &= \langle \phi(x), (I - \Phi_t K_t^{-1} \Phi_t^*) \phi(x) \rangle \\ &= \kappa(x,x) - \kappa(x, \tilde{X}_t) \kappa(\tilde{X}_t, \tilde{X}_t)^{-1} \kappa(\tilde{X}_t, x) \\ &= \frac{\det\left(K_{t \cup \{x\}}\right)}{\det(K_t)} \qquad \text{where} \qquad K_{t \cup \{x\}} := \begin{bmatrix} K_t & \Phi_t^* \phi(x) \\ \phi(x)^* \Phi_t & \kappa(x,x) \end{bmatrix} \end{aligned} \tag{39}$$

  so that this indeed corresponds to greedily maximizing the determinant of selected points. This method has also been proposed in (De Marchi et al., 2005) and used in multiple works such as (Chen et al., 2018). It is the only of the 3 mentioned methods that does not depend on the function f to approximate.

- **Greedy minimization of** $\|P_m^\perp f\|$**:**

$$\tilde{X}_{t+1} \in \arg\min_{x \in X} \left\| P_{t,x}^\perp f \right\| \text{ where } P_{t,x} \text{ is the othogonal projection on } \text{span}(\phi(\tilde{X}_1), \dots, \phi(\tilde{X}_t), \phi(x)).$$

  This method is also known as f/P **greedy interpolation**, as the new landmark chosen at each iteration is the one minimizing the residual over power function ratio. A rewriting of $P_{t,x}$ indeed yields the following relation:

$$\left\| P_{t,x}^\perp f \right\|^2 = \left\| P_t^\perp f \right\|^2 - \left(\frac{(P_t^\perp f)(x)}{\| P_t^\perp \phi(x) \|} \right)^2.$$

Algorithm These three methods can be implemented as shown in Algorithm H.1, and we implemented this algorithm in Julia[4].

Algorithm H.1: Greedy algorithms (f-greedy,P-greedy,f/P-greedy) for kernel interpolation

Input: Kernel κ, number of landmarks l, function evaluations $f_{|X} \in \mathbb{R}^n$, data $X \in \mathbb{R}^{d\times n}$

Output: Quadrature points $X[:, S]$

$C \leftarrow \text{zeros}(l, n)$; // Size $l \times n$

$\text{powfun}^2 \leftarrow [\kappa(X[:, i], X[:, i])$ for i in 1:n] ; // $(\|P_t^\perp \phi(X_i)\|^2)_{1\le i\le n}$, $O(nc_\kappa)$ time

$r \leftarrow f_{|X}$; // Residual, size n. $O(n^2)$ time when $f = \hat{\mu}_n$.

$c_f \leftarrow \text{zeros}(l)$; // Coefficients of f in U, size l

$S \leftarrow []$; // Support (set of indexes in $\{1, \dots, n\}$)

$k \leftarrow 0$;

while $k < l$ **do**

  newatom_criterion $\leftarrow$ **if** *P-greedy* **then**

    powfun^2 ;

  else if *f-greedy* **then**

    r ;

  else

    $r.^2/\text{powfun}^2$;

  $j \leftarrow \arg\max_{i\in\{1,\dots,n\}\setminus S}$ newatom_criterion ;

  $S \leftarrow S \cup \{j\}$;

  $k \leftarrow k + 1$;

  $K_j \leftarrow \text{kernelmatrix}(\kappa, x_j, X)$; // Size $1 \times n$, $O(nc_\kappa)$ time

  idxs $\leftarrow \text{powfun}^2. > 1\text{e-}10$; // For stability, update only points which are not already in the subspace

  $C[k, \text{idxs}] \leftarrow (K_j[\text{idxs}] - \text{vec}(C[:, j]' * C[:, \text{idxs}]))/\text{sqrt}(\text{powfun}^2[j])$; // $O(nl)$ time

  $c_f[k] \leftarrow r[j]/\text{sqrt}(\text{powfun}^2[j])$; // Update coefficients of f

  $r \leftarrow r\,. - c_f[k] * C[k, :]$; // Update the residual

  $\text{powfun}^2 \leftarrow \text{powfun}^2 - (C[k, :]^T).^2$; // Update power function, $O(n)$ time

return $X[:, S]$

Computational cost The algorithm has a cost of $O(nm(m+c_\kappa))$ time complexity, where c_κ denotes the kernel evaluation time and is typically of order $c_\kappa = O(d)$. Note that this cost does not include the computation of weights. Although we write the three algorithms together for conciseness, note that the method consisting in greedily maximizing $\det(K_m)$ does not require to compute the residual (the method being then independent of the function to approximate). In particular in our setting $f = \hat{\mu}_n$ and this would avoid the $O(n^2)$ cost of initializing the residual. The cost for computing the weights is $O(nm + m^3)$ and is the same for all methods (we use the same expression as for all other quadratures methods in

4. `https://gitlab.com/achatali/greedykernelmethods.jl`

the paper). With a small modification, the algorithm above can maintain an estimation of the weights, however the overall complexity of the algorithm remains unchanged.

Implementation We define the following quantities for any $1 \leq t \leq m$, which match the notations in Algorithm H.1 when relevant:

- $\tilde{\Phi}_t := [\phi(\tilde{X}_1), \dots, \phi(\tilde{X}_t)] : \mathbb{R}^t \to \mathcal{H}$.
- $U_t = [u_1, \dots, u_t] : \mathbb{R}^t \to \mathcal{H}$ is the Gram-Schmidt basis obtained from $\tilde{\Phi}_t$, i.e. for any t it holds
$$u_{t+1} := \frac{P_t^\perp \phi(\tilde{X}_{t+1})}{\|P_t^\perp \phi(\tilde{X}_{t+1})\|} \tag{40}$$
- $C \in \mathbb{R}^{m \times n}$ whose columns contain at step t the coefficients in U_t of the projected data features $(P_t\phi(X_i))_{1 \leq i \leq n}$, i.e. the block of the first t columns of C is $C_{1:t,:} = U_t^*[\phi(x_1), \dots, \phi(x_n)]$.
- S is a set containing the indexes of the so-far selected landmarks.

The algorithm then derives from the following observations.

- Line 19 derives from (40), indeed for any $i \in \{1, \dots, n\}$:
$$\begin{aligned}\langle u_{t+1}, \phi(X_i)\rangle &= \frac{\langle (I - P_t)\phi(\tilde{X}_{t+1}), \phi(X_i)\rangle}{\|P_t^\perp \phi(\tilde{X}_{t+1})\|} \\ &= \frac{\kappa(\tilde{X}_{t+1}, X_i) - \langle P_t\phi(\tilde{X}_{t+1}), P_t\phi(X_i)\rangle}{\|P_t^\perp \phi(\tilde{X}_{t+1})\|}\end{aligned}$$
and using the fact that at any iteration the index j is updated such that $\tilde{X}_{t+1} = x_j$.
- Line 20 follows from
$$\langle f, u_{t+1}\rangle = \frac{(P_t^\perp \phi(\tilde{X}_{t+1}))^* f}{\|P_t^\perp \phi(\tilde{X}_{t+1})\|} = \frac{(P_t^\perp f)(\tilde{X}_{t+1})}{\|P_t^\perp \phi(\tilde{X}_{t+1})\|}$$
Not in particular that no evaluations of f are required for this operation.
- Eventually Line 22 corresponds to the joint update for all $i \in \{1, \dots, n\}$ of the power function:
$$\begin{aligned}\|P_{t+1}^\perp \phi(X_i)\|^2 &= \|(P_t^\perp - u_{t+1}u_{t+1}^*)\phi(X_i)\|^2 \\ &= \|P_t^\perp \phi(X_i) - u_{t+1}u_{t+1}^*\phi(X_i)\|^2 \\ &= \|P_t^\perp \phi(X_i)\|^2 + \|u_{t+1}u_{t+1}^*\phi(X_i)\|^2 - 2\langle (I - P_t)\phi(X_i), u_{t+1}u_{t+1}^*\phi(X_i)\rangle \\ &= \|P_t^\perp \phi(X_i)\|^2 - \langle u_{t+1}, \phi(X_i)\rangle^2\end{aligned}$$
where we used the fact that $P_t u_{t+1} = 0$ and $\|u_{t+1}\| = 1$.

H.3 Additional Experimental Results

We here provide empirical results for the setting of Section 5.2, but on more datasets. Results are reported in Figure 4 for the Gaussian kernel and Figure 5 for the Laplacian kernel.

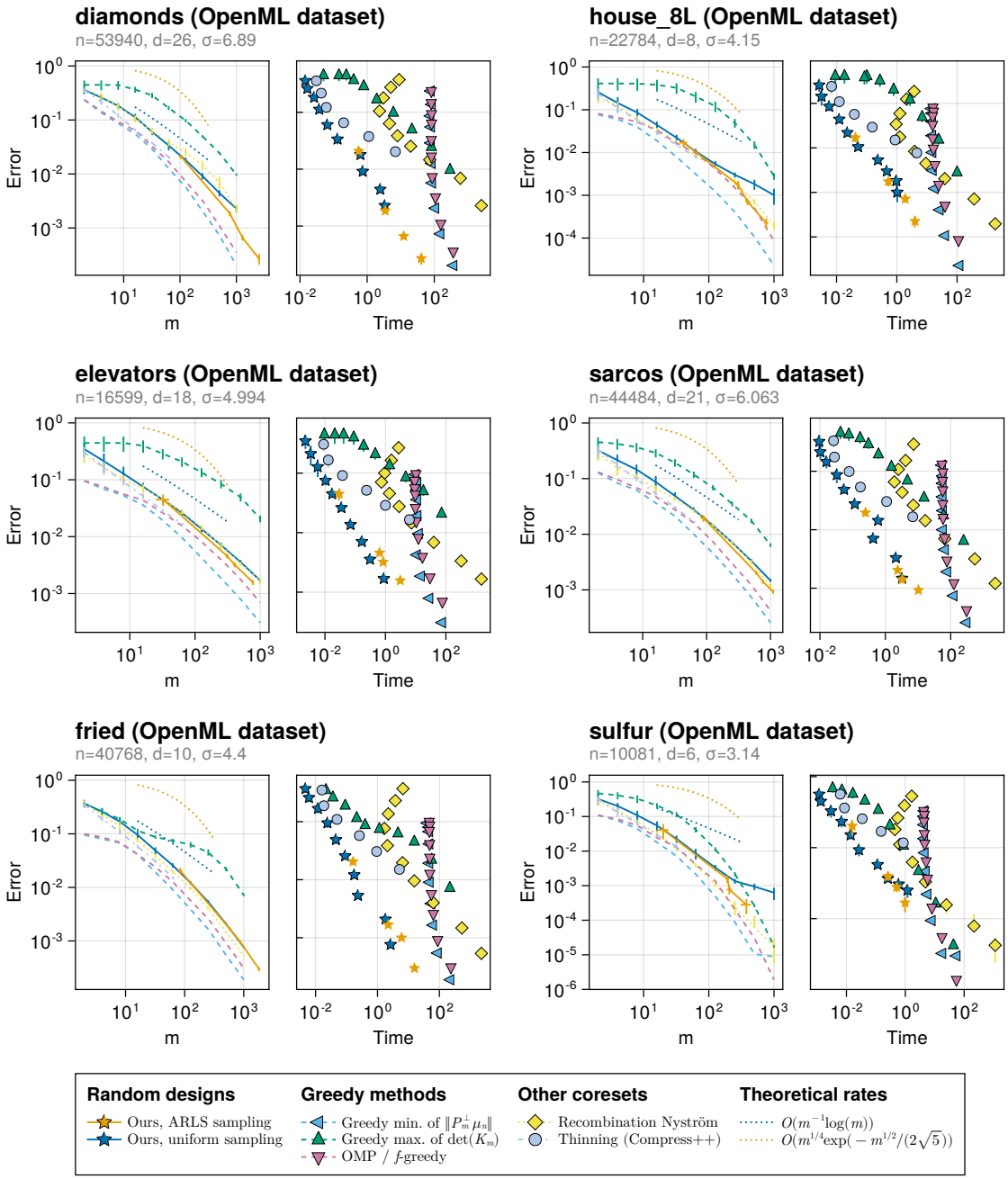

Figure 4: Empirical results for OpenML datasets, Gaussian kernel. Each point is a median over 40 trials.

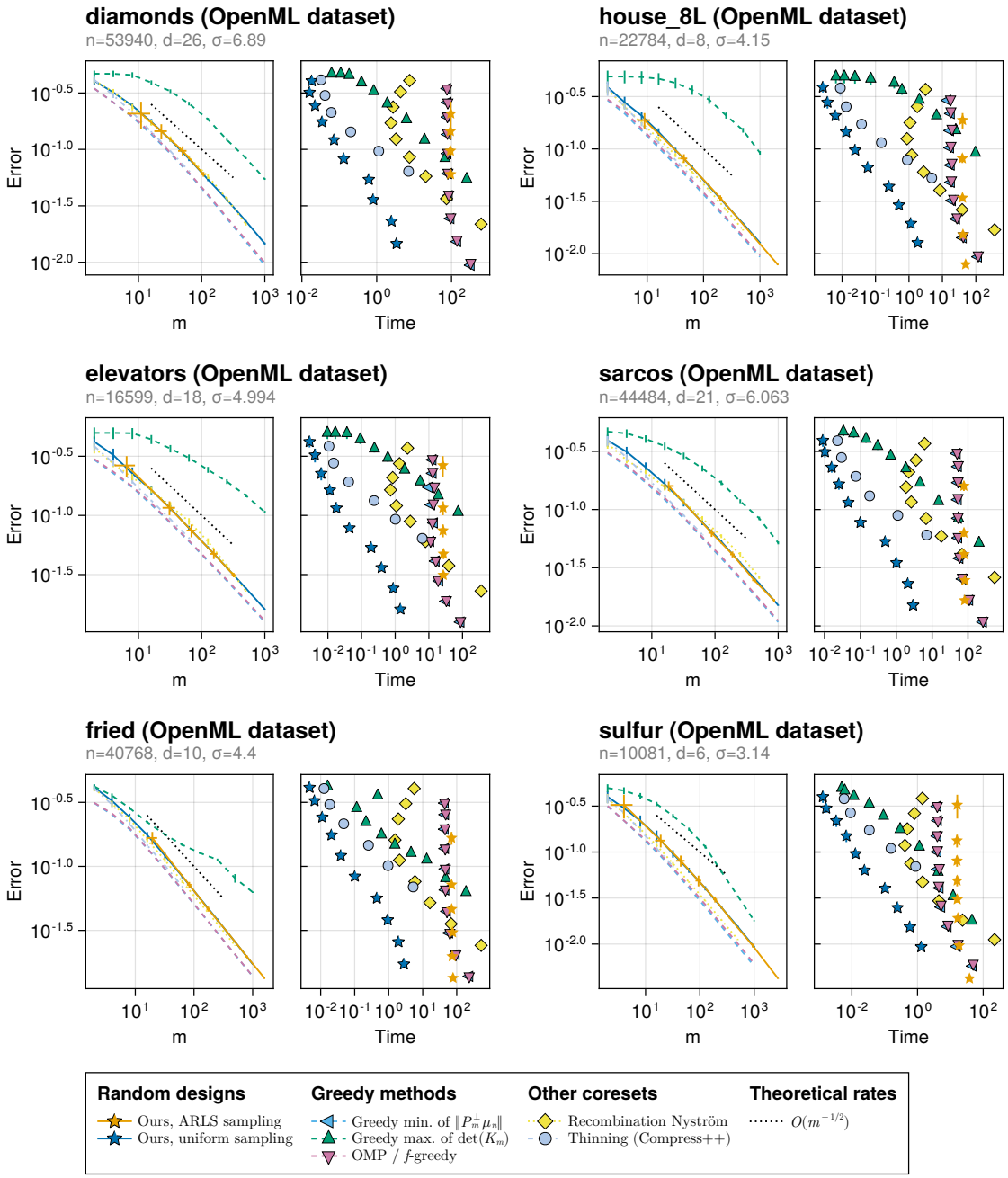

Figure 5: Empirical results for OpenML datasets, Laplacian kernel. Each point is a median over 40 trials.

References

Ahmed Alaoui and Michael W. Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 775–783, 2015.

Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum Mean Discrepancy Gradient Flow, December 2019.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

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

Francis Bach, Simon Lacoste-Julien, and Guillaume Obozinski. On the equivalence between herding and conditional gradient algorithms. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pages 1355–1362, Madison, WI, USA, June 2012. Omnipress. ISBN 978-1-4503-1285-1.

Matej Balog, Ilya Tolstikhin, and Bernhard Schölkopf. Differentially private database release via kernel mean embeddings. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 414–422. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/balog18a.html`.

Ayoub Belhadji. An analysis of Ermakov-Zolotukhin quadrature using kernels. In *Advances in Neural Information Processing Systems*, volume 34, pages 27278–27289. Curran Associates, Inc., 2021.

Ayoub Belhadji, Rémi Bardenet, and Pierre Chainais. Kernel quadrature with DPPs. In *arXiv:1906.07832 [Cs, Stat]*, volume 32, pages 12927–12937, December 2019.

Ayoub Belhadji, Rémi Bardenet, and Pierre Chainais. Kernel interpolation with continuous volume sampling. In *International Conference on Machine Learning*, pages 725–735. PMLR, November 2020.

Karsten M. Borgwardt, Arthur Gretton, Malte J. Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alexander J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. In *Proceedings 14th International Conference on Intelligent Systems for Molecular Biology 2006, Fortaleza, Brazil, August 6-10, 2006*, pages 49–57, 2006. doi: 10.1093/bioinformatics/btl242. URL `https://doi.org/10.1093/bioinformatics/btl242`.

L. Bos, S. De Marchi, A. Sommariva, and M. Vianello. Computing Multivariate Fekete and Leja Points by Numerical Linear Algebra. *SIAM Journal on Numerical Analysis*, 48(5): 1984–1999, January 2010. ISSN 0036-1429, 1095-7170. doi: 10.1137/090779024.

F. X. Briol, C. J. Oates, M. Girolami, M. A. Osborne, and D. Sejdinovic. Probabilistic integration: A role in statistical computation? *Statistical Science*, 34(1):1–22, February 2019. ISSN 0883-4237.

François-Xavier Briol, Chris Oates, Mark Girolami, and Michael A. Osborne. Frank-Wolfe Bayesian quadrature: Probabilistic integration with theoretical guarantees. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Francois-Xavier Briol, Chris J Oates, Jon Cockayne, Wilson Ye Chen, and Mark Girolami. On the Sampling Problem for Kernel Quadrature. In *International Conference on Machine Learning*, page 10, 2017.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Luigi Carratino, Stefano Vigogna, Daniele Calandriello, and Lorenzo Rosasco. ParK: Sound and Efficient Kernel Ridge Regression by Feature Space Partitions. In *Advances in Neural Information Processing Systems 34 Pre-Proceedings*, June 2021.

Antoine Chatalic, Vincent Schellekens, Florimond Houssiau, Yves-Alexandre De Montjoye, Laurent Jacques, and Rémi Gribonval. Compressive learning with privacy guarantees. *Information and Inference: A Journal of the IMA*, (iaab005), May 2021. ISSN 2049-8772. doi: 10.1093/imaiai/iaab005.

Antoine Chatalic, Luigi Carratino, Ernesto De Vito, and Lorenzo Rosasco. Mean Nyström Embeddings for Adaptive Compressive Learning. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 9869–9889. PMLR, May 2022a.

Antoine Chatalic, Nicolas Schreuder, Lorenzo Rosasco, and Alessandro Rudi. Nyström Kernel Mean Embeddings. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 3006–3024. PMLR, July 2022b.

Laming Chen, Guoxin Zhang, and Eric Zhou. Fast Greedy MAP Inference for Determinantal Point Process to Improve Recommendation Diversity. *Advances in Neural Information Processing Systems*, 31:5627–5638, 2018.

Yutian Chen, Max Welling, and Alex Smola. Super-samples from kernel herding. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, UAI'10, pages 109–116, Arlington, Virginia, USA, July 2010. AUAI Press. ISBN 978-0-9749039-6-5.

Badr-Eddine Chérief-Abdellatif and Pierre Alquier. Finite sample properties of parametric MMD estimation: Robustness to misspecification and dependence. *Bernoulli*, 28(1), February 2022. ISSN 1350-7265. doi: 10.3150/21-BEJ1338.

Heinz Otto Cordes. *Spectral Theory of Linear Differential Operators and Comparison Algebras*, volume 76. Cambridge University Press, 1987.

Philip J Davis and Philip Rabinowitz. *Methods of numerical integration*. Courier Corporation, 2007.

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

Andrea Della Vecchia, Jaouad Mourtada, Ernesto De Vito, and Lorenzo Rosasco. Regularized ERM on random subspaces. *arXiv:2006.10016 [cs, stat]*, February 2021.

Bernard Delyon and François Portier. Integral approximation by kernel smoothing. 2016.

Quentin Denoyelle, Vincent Duval, Gabriel Peyré, and Emmanuel Soubies. The sliding Frank–Wolfe algorithm and its application to super-resolution microscopy. *Inverse Problems*, 36(1):014001, December 2019. ISSN 0266-5611. doi: 10.1088/1361-6420/ab2a29.

R. A. DeVore and V. N. Temlyakov. Some remarks on greedy algorithms. *Advances in Computational Mathematics*, 5(1):173–187, December 1996. ISSN 1019-7168, 1572-9044. doi: 10.1007/BF02124742.

Josef Dick and Friedrich Pillichshammer. *Digital Nets and Sequences: Discrepancy Theory and Quasi-Monte Carlo Integration*. Cambridge University Press, 2010. ISBN 978-0-511-90197-3 0-511-90197-6 978-0-521-19159-3 0-521-19159-9.

Josef Dick, Peter Kritzer, and Friedrich Pillichshammer. *Lattice Rules: Numerical Integration, Approximation, and Discrepancy*, volume 58 of *Springer Series in Computational Mathematics*. Springer International Publishing, Cham, 2022. ISBN 978-3-031-09950-2 978-3-031-09951-9. doi: 10.1007/978-3-031-09951-9.

Joseph Diestel and John Jerry Uhl. *Vector Measures*. Mathematical Surveys and Monographs 15. American Mathematical Soc., 1977. ISBN 0-8218-1515-6 978-0-8218-1515-1.

Raaz Dwivedi and Lester Mackey. Kernel Thinning. *arXiv:2105.05842 [cs, math, stat]*, November 2021.

Raaz Dwivedi and Lester Mackey. Generalized Kernel Thinning, July 2022.

Heinz Werner Engl, Martin Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Mathematics and Its Applications. Springer Netherlands, 2000. ISBN 978-0-7923-4157-4.

Michaël Fanuel, Joachim Schreurs, and Johan A. K. Suykens. Nyström landmark sampling and regularized Christoffel functions. *Machine Learning*, 111(6):2213–2254, June 2022. ISSN 1573-0565. doi: 10.1007/s10994-022-06165-0.

Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *The Journal of Machine Learning Research*, 21(1):8464–8501, 2020.

Sushant S. Garud, Iftekhar A. Karimi, and Markus Kraft. Design of computer experiments: A review. *Computers & Chemical Engineering*, 106:71–95, November 2017. ISSN 00981354. doi: 10.1016/j.compchemeng.2017.05.010.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012. URL `http://dl.acm.org/citation.cfm?id=2188410`.

Zaid Harchaoui, Francis Bach, and Eric Moulines. Testing for Homogeneity with Kernel Fisher Discriminant Analysis, April 2008.

Satoshi Hayakawa, Harald Oberhauser, and Terry Lyons. Positively Weighted Kernel Quadrature via Subsampling. In *Advances in Neural Information Processing Systems*, volume 35, pages 6886–6900, October 2022.

Satoshi Hayakawa, Harald Oberhauser, and Terry Lyons. Sampling-based Nyström Approximation and Kernel Quadrature. In *Proceedings of the 40th International Conference on Machine Learning*, pages 12678–12699. PMLR, July 2023.

Saeed Hayati, Kenji Fukumizu, and Afshin Parvardeh. Kernel mean embedding of probability measures and its applications to functional data analysis. *arXiv preprint arXiv:2011.02315*, 2020.

Ferenc Huszár and David Duvenaud. Optimally-weighted herding is Bayesian quadrature. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI'12, pages 377–386, Arlington, Virginia, USA, August 2012. AUAI Press. ISBN 978-0-9749039-8-9.

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pages 427–435. PMLR, 2013.

Motonobu Kanagawa and Philipp Hennig. Convergence guarantees for adaptive Bayesian quadrature methods. *Advances in neural information processing systems*, 32, 2019.

Motonobu Kanagawa, Bharath K. Sriperumbudur, and Kenji Fukumizu. Convergence Analysis of Deterministic Kernel-Based Quadrature Rules in Misspecified Settings. *Foundations of Computational Mathematics*, 20(1):155–194, February 2020. ISSN 1615-3383. doi: 10.1007/s10208-018-09407-7.

Zohar Karnin and Edo Liberty. Discrepancy, Coresets, and Sketches in Machine Learning. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pages 1975–1993. PMLR, June 2019.

Toni Karvonen, Simo Särkkä, and Ken'ichiro Tanaka. Kernel-based interpolation at approximate Fekete points. *Numerical Algorithms*, 87(1):445–468, May 2021. ISSN 1572-9265. doi: 10.1007/s11075-020-00973-y.

Nicolas Keriven, Nicolas Tremblay, Yann Traonmilin, and Rémi Gribonval. Compressive K-means. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017.

Rajiv Khanna, Liam Hodgkinson, and Michael W. Mahoney. Geometric rates of convergence for kernel-based sampling algorithms. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 2156–2164. PMLR, December 2021.

Kee-Eung Kim and Hyun Soo Park. Imitation learning via kernel mean embedding. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3415–3422. AAAI Press, 2018. URL `https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16807`.

Rainer Kress. *Linear Integral Equations*. Applied Mathematical Sciences (Switzerland) 82. Springer-Verlag New York, 3 edition, 2014. ISBN 1-4614-9592-X 978-1-4614-9592-5 978-1-4614-9593-2 1-4614-9593-8.

Erwin Kreyszig. *Introductory Functional Analysis with Applications*. Wiley Classics Library. Wiley, 1 edition, 1989. ISBN 0-471-50459-9 978-0-471-50459-7.

Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Sampling methods for the Nyström method. *The Journal of Machine Learning Research*, 13(1):981–1006, 2012.

Simon Lacoste-Julien, Fredrik Lindsten, and Francis Bach. Sequential Kernel Herding: Frank-Wolfe Optimization for Particle Filtering. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 544–552. PMLR, 2015.

Alan J. Laub. *Matrix Analysis for Scientists and Engineers*. SIAM: Society for Industrial and Applied Mathematics, 2004. ISBN 978-0-89871-576-7.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. MMD GAN: towards deeper understanding of moment matching network. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 2203–2213, 2017. URL `https://proceedings.neurips.cc/paper/2017/hash/dfd7468ac613286cdbb40872c8ef3b06-Abstract.html`.

Francesco Locatello, Rajiv Khanna, Michael Tschannen, and Martin Jaggi. A Unified Optimization View on Generalized Matching Pursuit and Frank-Wolfe. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 860–868. PMLR, April 2017.

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

Krikamol Muandet, Kenji Fukumizu, Francesco Dinuzzo, and Bernhard Schölkopf. Learning from distributions via support measure machines. In Peter L. Bartlett,

Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 10–18, 2012. URL https://proceedings.neurips.cc/paper/2012/hash/9bf31c7ff062936a96d3c8bd1f8f2ff3-Abstract.html.

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Arthur Gretton, and Bernhard Schoelkopf. Kernel Mean Estimation and Stein Effect. In *Proceedings of the 31st International Conference on Machine Learning*, pages 10–18. PMLR, January 2014.

Krikamol Muandet, Kenji Fukumizu, Bharath K. Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Found. Trends Mach. Learn.*, 10(1-2):1–141, 2017. doi: 10.1561/2200000060. URL https://doi.org/10.1561/2200000060.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

Stefan Müller. *Komplexität und Stabilität von kernbasierten Rekonstruktionsmethoden*. doctoralThesis, March 2009.

Mark EJ Newman and Gerard T Barkema. *Monte Carlo methods in statistical physics*. Clarendon Press, 1999.

Erich Novak. *Deterministic and Stochastic Error Bounds In Numerical Analysis*. Springer, 1988.

Erich Novak and Hans Triebel. Function Spaces in Lipschitz Domains and Optimal Rates of Convergence for Sampling. *Constructive Approximation*, 23(3):325–350, February 2006. ISSN 1432-0940. doi: 10.1007/s00365-005-0612-y.

Erich Novak and Henryk Wozniakowski. *Tractability of Multivariate Problems, Volume II: Standard Information for Functionals*. EMS Tracts in Mathematics. European Mathematical Society, 2010. ISBN 3-03719-084-1 978-3-03719-084-5.

E. J. Nyström. Über Die Praktische Auflösung von Integralgleichungen mit Anwendungen auf Randwertaufgaben. *Acta Mathematica*, 54:185–204, 1930. ISSN 0001-5962, 1871-2509. doi: 10.1007/BF02547521.

Chris Oates, Steven Niederer, Angela Lee, François-Xavier Briol, and Mark Girolami. Probabilistic Models for Integration Error in the Assessment of Functional Cardiac Models. In *31st Conference on Neural Information Processing Systems*, Long Beach, CA, USA, 2017.

Brooks Paige, Dino Sejdinovic, and Frank Wood. Super-sampling with a reservoir. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, pages 567–576, Arlington, Virginia, USA, June 2016. AUAI Press. ISBN 978-0-9966431-1-5.

Mijung Park, Wittawat Jitkrittum, and Dino Sejdinovic. K2-ABC: approximate bayesian computation with kernel embeddings. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016, Cadiz, Spain, May 9-11, 2016*, volume 51 of *JMLR Workshop and Conference Proceedings*, pages 398–407. JMLR.org, 2016. URL `http://proceedings.mlr.press/v51/park16.html`.

Edouard Pauwels, Francis Bach, and Jean-Philippe Vert. Relating Leverage Scores and Density Using Regularized Christoffel Functions. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 1670–1679, Red Hook, NY, USA, 2018. Curran Associates Inc.

Iosif Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *The Annals of Probability*, pages 1679–1706, 1994.

Alfio Quarteroni and Alberto Valli. *Numerical Approximation of Partial Differential Equations (Springer Series in Computational Mathematics)*. Springer Series in Computational Mathematics. Springer, 1st ed. 1994. 2nd printing edition, 2008. ISBN 978-3-540-85267-4 3-540-85267-0.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes in Machine Learning*, volume 2. MIT press Cambridge, MA, 2006.

Michael Reed and Barry Simon. *Functional Analysis, Volume 1*. Methods of Modern Mathematical Physics. Academic Press, rev and enl edition, 1981. ISBN 0-12-585050-6 978-0-12-585050-6.

Jon A. Rivera, Jamie M. Taylor, Ángel J. Omella, and David Pardo. On quadrature rules for solving Partial Differential Equations using Neural Networks. *Computer Methods in Applied Mechanics and Engineering*, 393:114710, April 2022. ISSN 0045-7825. doi: 10.1016/j.cma.2022.114710.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nyström computational regularization. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 1657–1665, Cambridge, MA, USA, 2015. MIT Press.

Alessandro Rudi, Daniele Calandriello, Luigi Carratino, and Lorenzo Rosasco. On fast leverage score sampling and optimal learning. In *Advances in Neural Information Processing Systems*, pages 5672–5682, 2018.

Gabriele Santin, Toni Karvonen, and Bernard Haasdonk. Sampling based approximation of linear functionals in reproducing kernel Hilbert spaces. *BIT Numerical Mathematics*, 62(1):279–310, March 2022. ISSN 1572-9125. doi: 10.1007/s10543-021-00870-3.

Robert Schaback. Superconvergence of kernel-based interpolation. *Journal of Approximation Theory*, 235:1–19, November 2018. ISSN 0021-9045. doi: 10.1016/j.jat.2018.05.002.

Abhishek Shetty, Raaz Dwivedi, and Lester Mackey. Distribution Compression in Near-linear Time. In *ICLR 2022*. arXiv, October 2022. doi: 10.48550/arXiv.2111.07941.

Alexander J. Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A hilbert space embedding for distributions. In Marcus Hutter, Rocco A. Servedio, and Eiji Takimoto, editors, *Algorithmic Learning Theory, 18th International Conference, ALT 2007, Sendai, Japan, October 1-4, 2007, Proceedings*, volume 4754 of *Lecture Notes in Computer Science*, pages 13–31. Springer, 2007. doi: 10.1007/978-3-540-75225-7_5. URL https://doi.org/10.1007/978-3-540-75225-7_5.

Le Song, Kenji Fukumizu, and Arthur Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Process. Mag.*, 30(4):98–111, 2013. doi: 10.1109/MSP.2013.2252713. URL https://doi.org/10.1109/MSP.2013.2252713.

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert Space Embeddings and Metrics on Probability Measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561, 2010. ISSN 1533-7928.

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

Danica J. Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alexander J. Smola, and Arthur Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=HJWHIKqgl.

Ilya Tolstikhin, Bharath K Sriperumbudur, and Krikamol Muandet. Minimax estimation of kernel mean embeddings. *The Journal of Machine Learning Research*, 18(1):3002–3048, 2017.

Kazuma K. Tsuji, Ken'Ichiro Tanaka, and Sebastian Pokutta. Pairwise Conditional Gradients without Swap Steps and Sparser Kernel Herding. In *Proceedings of the 39th International Conference on Machine Learning*, pages 21864–21883. PMLR, June 2022.

Grace Wahba. *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM: Society for Industrial and Applied Mathematics, illustrated edition edition, 1990. ISBN 978-0-89871-244-5 0-89871-244-0.

Eric W Weisstein. Lambert w-function. *https://mathworld. wolfram. com/*, 2002.

Holger Wendland. *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2004. ISBN 0-521-84335-9 978-0-521-84335-5 978-0-511-26579-2.

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

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

Christopher Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, pages 682–688, 2001.

Vadim Yurinsky. *Sums and Gaussian Vectors*. Lecture Notes in Mathematics 1617. Springer-Verlag Berlin Heidelberg, 1 edition, 1995. ISBN 978-3-540-60311-5.

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 819–827. JMLR.org, 2013. URL `http://proceedings.mlr.press/v28/zhang13d.html`.

Shaofeng Zou, Yingbin Liang, H. Vincent Poor, and Xinghua Shi. Nonparametric detection of anomalous data via kernel mean embedding. *CoRR*, abs/1405.2294, 2014. URL `http://arxiv.org/abs/1405.2294`.