Generalization Properties of Learning with Random Features

Alessandro Rudi *
INRIA - Sierra Project-team,
École Normale Supérieure, Paris,
75012 Paris, France
alessandro.rudi@inria.fr

Lorenzo Rosasco
University of Genova,
Istituto Italiano di Tecnologia,
Massachusetts Institute of Technology.
lrosasco@mit.edu

Abstract

We study the generalization properties of ridge regression with random features in the statistical learning framework. We show for the first time that $O(1/\sqrt{n})$ learning bounds can be achieved with only $O(\sqrt{n}\log n)$ random features rather than $O(n)$ as suggested by previous results. Further, we prove faster learning rates and show that they might require more random features, unless they are sampled according to a possibly problem dependent distribution. Our results shed light on the statistical computational trade-offs in large scale kernelized learning, showing the potential effectiveness of random features in reducing the computational complexity while keeping optimal generalization properties.

1 Introduction

Supervised learning is a basic machine learning problem where the goal is estimating a function from random noisy samples [1, 2]. The function to be learned is fixed, but unknown, and flexible non-parametric models are needed for good results. A general class of models is based on functions of the form,

$$f(x) \; = \; \sum_{i=1}^{M} \alpha_i \, q(x, \omega_i), \tag{1}$$

where q is a non-linear function, $\omega_1, \ldots, \omega_M \in \mathbb{R}^d$ are often called centers, $\alpha_1, \ldots, \alpha_M \in \mathbb{R}$ are coefficients, and $M = M_n$ could/should *grow* with the number of data points n. Algorithmically, the problem reduces to computing from data the parameters $\omega_1, \ldots, \omega_M, \alpha_1, \ldots, \alpha_M$ and M. Among others, one-hidden layer networks [3], or RBF networks [4], are examples of classical approaches considering these models. Here, parameters are computed by considering a non-convex optimization problem, typically hard to solve and analyze [5]. Kernel methods are another notable example of an approach [6] using functions of the form (1). In this case, q is assumed to be a positive definite function [7] and it is shown that choosing the centers to be the input points, hence $M = n$, suffices for optimal statistical results [8, 9, 10]. As a by product, kernel methods require only finding the coefficients $(\alpha_i)_i$, typically by convex optimization. While theoretically sound and remarkably effective in small and medium size problems, memory requirements make kernel methods unfeasible for large scale problems.

Most popular approaches to tackle these limitations are randomized and include sampling the centers at random, either in a data-dependent or in a data-independent way. Notable examples include Nyström [11, 12] and random features [13] approaches. Given random centers, computations still

---

reduce to convex optimization with potential big memory gains, provided that the centers are fewer than the data-points. In practice, the choice of the number of centers is based on heuristics or memory constraints, and the question arises of characterizing theoretically which choices provide optimal learning bounds. Answering this question allows to understand the statistical and computational trade-offs in using these randomized approximations. For Nyström methods, partial results in this direction were derived for example in [14] and improved in [15], but only for a simplified setting where the input points are fixed. Results in the statistical learning setting were given in [16] for ridge regression, showing in particular that $O(\sqrt{n}\log n)$ random centers uniformly sampled from n training points suffices to yield $O(1/\sqrt{n})$ learning bounds, *the same as full kernel ridge regression.*

A question motivating our study is whether similar results hold for random features approaches. While several papers consider the properties of random features for approximating the kernel function, see [17] and references therein, fewer results consider their generalization properties.

Several papers considered the properties of random features for approximating the kernel function, see [17] and references therein, an interesting line of research with connections to sketching [18] and non-linear (one-bit) compressed sensing [19]. However, only a few results consider the generalization properties of learning with random features.

An exception is one of the original random features papers, which provides learning bounds for a general class of loss functions [20]. These results show that $O(n)$ random features are needed for $O(1/\sqrt{n})$ learning bounds and choosing less random features leads to worse bounds. In other words, these results suggest that that computational gains come at the expense of learning accuracy. Later results, see e.g. [21, 22, 23], essentially confirm these considerations, albeit the analysis in [23] suggests that fewer random features could suffice if sampled in a problem dependent way.

In this paper, we focus on the least squares loss, considering random features within a ridge regression approach. Our main result shows, under standard assumptions, that the estimator obtained with a number of random features proportional to $O(\sqrt{n}\log n)$ achieves $O(1/\sqrt{n})$ learning error, that is the *same* prediction accuracy of the *exact* kernel ridge regression estimator. In other words, there are problems for which random features can allow to drastically reduce computational costs *without* any loss of prediction accuracy. To the best of our knowledge this is the first result showing that such an effect is possible. Our study improves on previous results by taking advantage of analytic and probabilistic results developed to provide sharp analyses of kernel ridge regression. We further present a second set of more refined results deriving fast convergence rates. We show that indeed fast rates are possible, but, depending on the problem at hand, a larger number of features might be needed. We then discuss how the requirement on the number of random features can be weakened at the expense of typically more complex sampling schemes. Indeed, in this latter case either some knowledge of the data-generating distribution or some potentially data-driven sampling scheme is needed. For this latter case, we borrow and extend ideas from [23, 16] and inspired from the theory of statical leverage scores [24]. Theoretical findings are complemented by numerical simulation validating the bounds.

The rest of the paper is organized as follows. In Section 2, we review relevant results on learning with kernels, least squares and learning with random features. In Section 3, we present and discuss our main results, while proofs are deferred to the appendix. Finally, numerical experiments are presented in Section 4.

2 Learning with random features and ridge regression

We begin recalling basics ideas in kernel methods and their approximation via random features.

Kernel ridge regression Consider the supervised problem of learning a function given a training set of n examples $(x_i, y_i)_{i=1}^n$, where $x_i \in X$, $X = \mathbb{R}^D$ and $y_i \in \mathbb{R}$. Kernel methods are nonparametric approaches defined by a *kernel* $K : X \times X \to \mathbb{R}$, that is a symmetric and positive definite (PD) function[2]. A particular instance is kernel ridge regression given by

$$\widehat{f}_\lambda(x) = \sum_{i=1}^n \alpha_i K(x_i, x), \quad \alpha = (\mathbf{K} + \lambda n I)^{-1} y. \tag{2}$$

---

[2]A kernel K is PD if for all $x_1, \dots, x_N$ the N by N matrix with entries $K(x_i, x_j)$ is positive semidefinite.

Here $\lambda > 0$, $y = (y_1, \ldots, y_n)$, $\alpha \in \mathbb{R}^n$, and $\mathbf{K}$ is the n by n matrix with entries $\mathbf{K}_{ij} = K(x_i, x_j)$. The above method is standard and can be derived from an empirical risk minimization perspective [6], and is related to Gaussian processes [3]. While KRR has optimal statistical properties– see later– its applicability to large scale datasets is limited since it requires $O(n^2)$ in space, to store $\mathbf{K}$, and roughly $O(n^3)$ in time, to solve the linear system in (2). Similar requirements are shared by other kernel methods [6].

To explain the basic ideas behind using random features with ridge regression, it is useful to recall the computations needed to solve KRR when the kernel is linear $K(x, x') = x^\top x'$. In this case, Eq. (2) reduces to standard ridge regression and can be equivalenty computed considering,

$$\widehat{f}_\lambda(x) = x^\top \widehat{w}_\lambda \qquad \widehat{w}_\lambda = (\widehat{X}^\top \widehat{X} + \lambda n I)^{-1} \widehat{X}^\top y. \tag{3}$$

where $\widehat{X}$ is the n by D data matrix. In this case, the complexity becomes $O(nD)$ in space, and $O(nD^2 + D^3)$ in time. Beyond the linear case, the above reasoning extends to inner product kernels

$$K(x, x') = \phi_M(x)^\top \phi_M(x') \tag{4}$$

where $\phi_M : X \to \mathbb{R}^M$ is a finite dimensional (feature) map. In this case, KRR can be computed considering (3) with the data matrix $\widehat{X}$ replaced by the n by M matrix $\widehat{S}_M^\top = (\phi(x_1), \ldots, \phi(x_n))$. The complexity is then $O(nM)$ in space, and $O(nM^2 + M^3)$ in time, hence much better than $O(n^2)$ and $O(n^3)$, as soon as $M \ll n$. Considering only kernels of the form (4) can be restrictive. Indeed, classic examples of kernels, e.g. the Gaussian kernel $e^{-\|x-x'\|^2}$, do not satisfy (4) with finite M. It is then natural to ask if the above reasoning can still be useful to reduce the computational burden for more complex kernels such as the Gaussian kernel. Random features, that we recall next, show that this is indeed the case.

Random features with ridge regression The basic idea of random features [13] is to relax Eq. (4) assuming it holds only approximately,

$$K(x, x') \approx \phi_M(x)^\top \phi_M(x'). \tag{5}$$

Clearly, if one such approximation exists the approach described in the previous section can still be used. A first question is then for which kernels an approximation of the form (5) can be derived. A simple manipulation of the Gaussian kernel provides one basic example.

Example 1 (Random Fourier features [13])**.** *If we write the Gaussian kernel as $K(x, x') = G(x - x')$, with $G(z) = e^{-\frac{1}{2\sigma^2}\|z\|^2}$, for a $\sigma > 0$, then since the inverse Fourier transform of G is a Gaussian, and using a basic symmetry argument, it is easy to show that*

$$G(x - x') \quad = \quad \frac{1}{2\pi Z} \int \int_0^{2\pi} \sqrt{2}\cos(w^\top x + b)\ \sqrt{2}\cos(w^\top x' + b)\ e^{-\frac{\sigma^2}{2}\|w\|^2} dw\ db$$

where Z is a normalizing factor. Then, the Gaussian kernel has an approximation of the form (5) with $\phi_M(x) = M^{-1/2} (\sqrt{2}\cos(w_1^\top x + b_1), \ldots, \sqrt{2}\cos(w_M^\top x + b_M))$, and $w_1, \ldots, w_M$ and $b_1, \ldots, b_M$ sampled independently from $\frac{1}{Z}e^{-\sigma^2\|w\|^2/2}$ and uniformly in $[0, 2\pi]$, respectively.

The above example can be abstracted to a general strategy. Assume the kernel K to have an integral representation,

$$K(x, x') = \int_\Omega \psi(x, \omega)\psi(x', \omega)d\pi(\omega), \quad \forall x, x' \in X, \tag{6}$$

where (Ω, π) is probability space and $\psi : X \times \Omega \to \mathbb{R}$. The random features approach provides an approximation of the form (5) where $\phi_M(x) = M^{-1/2} (\psi(x, \omega_1), \ldots, \psi(x, \omega_M))$, and with $\omega_1, \ldots, \omega_M$ sampled independently with respect to π. Key to the success of random features is that kernels, to which the above idea apply, abound– see Appendix E for a survey with some details.

Remark 1 (Random features, sketching and one-bit compressed sensing)**.** *We note that specific examples of random features can be seen as form of sketching [18]. This latter term typically refers to reducing data dimensionality by random projection, e.g. considering*

$$\psi(x, \omega) = x^\top \omega,$$

3

where $\omega \sim N(0, I)$ (or suitable bounded measures). From a random feature perspective, we are defining an approximation of the linear kernel since $\mathbb{E}[\psi(x, \omega)\psi(x', \omega)] = \mathbb{E}[x^\top \omega \omega^\top x'] = x^\top \mathbb{E}[\omega \omega^\top]x' = x^\top x'$. More general non-linear sketching can also be considered. For example in one-bit compressed sensing [19] the following random features are relevant,

$$\psi(x, \omega) = sign(x^\top \omega)$$

with $w \sim N(0, I)$ and $sign(a) = 1$ if $a > 0$ and -1 otherwise. Deriving the corresponding kernel is more involved and we refer to [25] (see Section E in the appendixes).

Back to supervised learning, combining random features with ridge regression leads to,

$$\widehat{f}_{\lambda,M}(x) := \phi_M(x)^\top \widehat{w}_{\lambda,M}, \quad \text{with} \quad \widehat{w}_{\lambda,M} := (\widehat{S}_M^\top \widehat{S}_M + \lambda I)^{-1} \widehat{S}_M^\top \widehat{y}, \tag{7}$$

for $\lambda > 0$, $\widehat{S}_M^\top := n^{-1/2} (\phi_M(x_1), \ldots, \phi_M(x_n))$ and $\widehat{y} := n^{-1/2} (y_1, \ldots, y_n)$.

Then, random features can be used to reduce the computational costs of full kernel ridge regression as soon as $M \ll n$ (see Sec. 2). However, since random features rely on an approximation (5), the question is whether there is a loss of prediction accuracy. This is the question we analyze in the rest of the paper.

3   Main Results

In this section, we present our main results characterizing the generalization properties of random features with ridge regression. We begin considering a basic setting and then discuss fast learning rates and the possible benefits of problem dependent sampling schemes.

3.1 $O(\sqrt{n}\log n)$ Random features lead to $O(1/\sqrt{n})$ learning error

We consider a standard statistical learning setting. The data $(x_i, y_i)_{i=1}^n$ are sampled identically and independently with respect to a probability ρ on $X \times \mathbb{R}$, with X a separable space (e.g. $X = \mathbb{R}^D$, $D \in \mathbb{N}$). The goal is to minimize the expected risk

$$\mathcal{E}(f) = \int (f(x) - y)^2 d\rho(x, y),$$

since this implies that f will generalize/predict well new data. Since we consider estimators of the form (2), (7) we are potentially restricting the space of possible solutions. Indeed, estimators of this form can be naturally related to the so called reproducing kernel Hilbert space (RKHS) corresponding to the PD kernel K. Recall that, the latter is the function space $\mathcal{H}$ defined as as the completion of the linear span of $\{K(x, \cdot) \; : \; x \in X\}$ with respect to the inner product $\langle K(x, \cdot), K(x', \cdot)\rangle := K(x, x')$ [7]. In this view, the best possible solution is $f_\mathcal{H}$ solving

$$\min_{f \in \mathcal{H}} \mathcal{E}(f). \tag{8}$$

We will assume throughout that $f_\mathcal{H}$ exists. We add one technical remark useful in the following.

Remark 2. *Existence of $f_\mathcal{H}$ is not ensured, since we consider a potentially infinite dimensional RKHS $\mathcal{H}$, possibly universal [26]. The situation is different if $\mathcal{H}$ is replaced by $\mathcal{H}_R = \{f \in \mathcal{H} \; : \; \|f\| \leq R\}$, with R fixed a priori. In this case a minimizer of risk $\mathcal{E}$ always exists, but R needs to be fixed a priori and $\mathcal{H}_R$ can't be universal. Clearly, assuming $f_\mathcal{H}$ to exist, implies it belongs to a ball of radius $R_{\rho,\mathcal{H}}$. However, our results do not require prior knowledge of $R_{\rho,\mathcal{H}}$ and hold uniformly over all finite radii.*

The following is our first result on the learning properties of random features with ridge regression.

Theorem 1. *Assume that K is a kernel with an integral representation (6). Assume ψ continuous, such that $|\psi(x, \omega)| \leq \kappa$ almost surely, with $\kappa \in [1, \infty)$ and $|y| \leq b$ almost surely, with $b > 0$. Let $\delta \in (0, 1]$. If $n \geq n_0$ and $\lambda_n = n^{-1/2}$, then a number of random features M_n equal to*

$$M_n = c_0 \; \sqrt{n} \; \log \frac{108\kappa^2 \sqrt{n}}{\delta},$$

is enough to guarantee, with probability at least $1 - \delta$, that

$$\mathcal{E}(\widehat{f}_{\lambda_n, M_n}) - \mathcal{E}(f_\mathcal{H}) \leq \frac{c_1 \log^2 \frac{18}{\delta}}{\sqrt{n}}.$$

In particular the constants c_0, c_1 do not depend on n, λ, δ, and n_0 does not depends on $n, \lambda, f_\mathcal{H}, \rho$.

The above result is presented with some simplifications (e.g. the assumption of bounded output) for sake of presentation, while it is proved and presented in full generality in the Appendix. In particular, the values of all the constants are given explicitly. Here, we make a few comments. The learning bound is the same achieved by the *exact* kernel ridge regression estimator (2) choosing $\lambda = n^{-1/2}$, see e.g. [10]. The theorem derives a bound in a worst case situation, where no assumption is made besides existence of $f_{\mathcal{H}}$, and is optimal in a minmax sense [10]. This means that, in this setting, as soon as the number of features is order $\sqrt{n} \log n$, the corresponding ridge regression estimator has optimal generalization properties. This is remarkable considering the corresponding gain from a computational perspective: from roughly $O(n^3)$ and $O(n^2)$ in time and space for kernel ridge regression to $O(n^2)$ and $O(n\sqrt{n})$ for ridge regression with random features (see Section 2). Consider that taking $\delta \propto 1/n^2$ changes only the constants and allows to derive bounds in expectation and almost sure convergence (see Cor. 1 in the appendix, for the result in expectation).

The above result shows that there is a whole set of problems where computational gains are achieved without having to trade-off statistical accuracy. In the next sections we consider what happens under more benign assumptions, which are standard, but also somewhat more technical. We first compare with previous works since the above setting is the one more closely related.

Comparison with [20]. This is one of the original random features paper and considers the question of generalization properties. In particular they study the estimator

$$\widehat{f}_R(x) = \phi_M(x)^\top \widehat{\beta}_{R,\infty}, \quad \widehat{\beta}_{R,\infty} = \underset{\|\beta\|_\infty \leq R}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \ell(\phi_M(x_i)^\top \beta, y_i),$$

for a fixed R, a Lipshitz loss function ℓ, and where $\|w\|_\infty = \max\{|\beta_1|, \cdots, |\beta_M|\}$. The largest space considered in [20] is

$$\mathcal{G}_R = \left\{ \int \psi(\cdot, \omega)\beta(\omega)d\pi(\omega) \;\middle|\; |\beta(\omega)| < R \text{ a.e.} \right\}, \tag{9}$$

rather than a RKHS, where R is fixed a priori. The best possible solution is $f_{\mathcal{G}_R}^*$ solving $\min_{f \in \mathcal{G}_R} \mathcal{E}(f)$, and the main result in [20] provides the bound

$$\mathcal{E}(\widehat{f}_R) - \mathcal{E}(f_{\mathcal{G}_R}^*) \lesssim \frac{R}{\sqrt{n}} + \frac{R}{\sqrt{M}}, \tag{10}$$

This is the first and still one the main results providing a statistical analysis for an estimator based on random features for a wide class of loss functions. There are a few elements of comparison with the result in this paper, but the main one is that to get $O(1/\sqrt{n})$ learning bounds, the above result requires $O(n)$ random features, while a smaller number leads to worse bounds. This shows the main novelty of our analysis. Indeed we prove that, considering the square loss, fewer random features are sufficient, hence allowing computational gains without loss of accuracy. We add a few more tehcnical comments explaining : 1) how the setting we consider covers a wider range of problems, and 2) why the bounds we obtain are sharper. First, note that the functional setting in our paper is more general in the following sense. It is easy to see that considering the RKHS $\mathcal{H}$ is equivalent to consider $\mathcal{H}_2 = \left\{ \int \psi(\cdot, \omega)\beta(\omega)d\pi(\omega) \;\middle|\; \int |\beta(\omega)|^2 d\pi(\omega) < \infty \right\}$ and the following inclusions hold $\mathcal{G}_R \subset \mathcal{G}_\infty \subset \mathcal{H}_2$. Clearly, assuming a minimizer of the expected risk to exists in $\mathcal{H}_2$ *does not* imply it belongs to $\mathcal{G}_\infty$ or $\mathcal{G}_R$, while the converse is true. In this view, our results cover a wider range of problems. Second, note that, this gap is not easy to bridge. Indeed, even if we were to consider $\mathcal{G}_\infty$ in place of $\mathcal{G}_R$, the results in [20] could be used to derive the bound

$$\mathbb{E} \; \mathcal{E}(\widehat{f}_R) - \mathcal{E}(f_{\mathcal{G}_\infty}^*) \lesssim \frac{R}{\sqrt{n}} + \frac{R}{\sqrt{M}} + A(R), \tag{11}$$

where $A(R) := \mathcal{E}(f_{\mathcal{G}_R}^*) - \mathcal{E}(f_{\mathcal{G}_\infty}^*)$ and $f_{\mathcal{G}_\infty}^*$ is a minimizer of the expected risk on $\mathcal{G}_\infty$. In this case we would have to balance the various terms in (11), which would lead to a worse bound. For example, we could consider $R := \log n$, obtaining a bound $n^{-1/2} \log n$ with an extra logarithmic term, but the result would hold only for n larger than a number of examples n_0 at least *exponential* with respect to the norm of f_∞. Moreover, to derive results uniform with respect to f_∞, we would have to keep into account the decay rate of $A(R)$ and this would get bounds slower than $n^{-1/2}$.
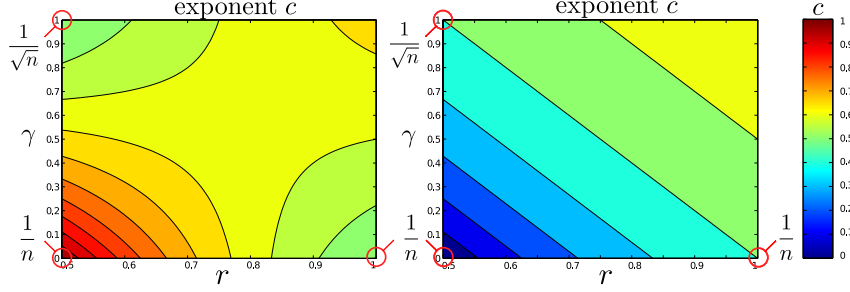
Figure 1: Random feat. $M = O(n^c)$ required for optimal generalization. Left: $\alpha = 1$. Right: $\alpha = \gamma$.

Comparison with other results. Several other papers study the generalization properties of random features, see [23] and references therein. For example, generalization bounds are derived in [21] from very general arguments. However, the corresponding generalization bound requires a number of random features much larger than the number of training examples to give $O(1/\sqrt{n})$ bounds. The basic results in [23] are analogous to those in [20] with the set $\mathcal{G}_R$ replaced by $\mathcal{H}_R$. These results are closer, albeit more restrictive then ours (see Remark 8) and especially like the bounds in [20] suggest $O(n)$ random features are needed for $O(1/\sqrt{n})$ learning bounds. A novelty in [23] is the introduction of more complex problem dependent sampling that can reduce the number of random features. In Section 3.3, we show that using possibly-data dependent random features can lead to rates much faster than $n^{-1/2}$, and using much less than $\sqrt{n}$ features.

Remark 3 (Sketching and randomized numerical linear algebra (RandLA))**.** *Standard sketching techniques from RandLA [18] can be recovered, when X is a bounded subset of $\mathbb{R}^D$, by selecting $\psi(x, \omega) = x^\top \omega$ and ω sampled from suitable bounded distribution (e.g. $\omega = (\zeta_1, \ldots, \zeta_d)$ independent Rademacher random variables). Note however that the final goal of the analysis in the randomized numerical linear algebra community is to minimize the empirical error instead of $\mathcal{E}$.*

3.2 Refined Results: Fast Learning Rates

Faster rates can be achieved under favorable conditions. Such conditions for kernel ridge regression are standard, but somewhat technical. Roughly speaking they characterize the "size" of the considered RKHS and the regularity of $f_{\mathcal{H}}$. The key quantity needed to make this precise is the integral operator defined by the kernel K and the marginal distribution ρ_X of ρ on X, that is

$$(Lg)(x) = \int_X K(x, z)g(z)d\rho_X(z), \quad \forall g \in L^2(X, \rho_X),$$

seen as a map from $L^2(X, \rho_X) = \{f : X \to \mathbb{R} \mid \|f\|_\rho^2 = \int |f(x)|^2 d\rho_X < \infty\}$ to itself. Under the assumptions of Thm. 1, the integral operator is positive, self-adjoint and trace-class (hence compact) [27]. We next define the conditions that will lead to fast rates, and then comment on their interpretation.

Assumption 1 (Prior assumptions)**.** *For $\lambda > 0$, let the effective dimension be defined as $\mathcal{N}(\lambda) := \mathrm{Tr}\left((L + \lambda I)^{-1} L\right)$, and assume, there exists $Q > 0$ and $\gamma \in [0, 1]$ such that,*

$$\mathcal{N}(\lambda) \leq Q^2 \lambda^{-\gamma}. \tag{12}$$

Moreover, assume there exists $r \geq 1/2$ and $g \in L^2(X, \rho_X)$ such that

$$f_{\mathcal{H}}(x) = (L^r g)(x) \quad a.s. \tag{13}$$

We provide some intuition on the meaning of the above assumptions, and defer the interested reader to [10] for more details. The effective dimension can be seen as a "measure of the size" of the RKHS $\mathcal{H}$. Condition (12) allows to control the variance of the estimator and is equivalent to conditions on covering numbers and related capacity measures [26]. In particular, it holds if the eigenvalues σ_i's of L decay as $i^{-1/\gamma}$. Intuitively, a fast decay corresponds to a smaller RKHS, whereas a slow decay corresponds to a larger RKHS. The case $\gamma = 0$ is the more benign situation, whereas $\gamma = 1$ is the worst case, corresponding to the basic setting. A classic example, when $X = \mathbb{R}^D$, corresponds to

considering kernels of smoothness s, in which case $\gamma = D/(2s)$ and condition (12) is equivalent to assuming $\mathcal{H}$ to be a Sobolev space [26]. Condition (13) allows to control the bias of the estimator and is common in approximation theory [28]. It is a regularity condition that can be seen as form of weak sparsity of $f_{\mathcal{H}}$. Roughly speaking, it requires the expansion of $f_{\mathcal{H}}$, on the the basis given by the the eigenfunctions L, to have coefficients that decay faster than σ_i^r. A large value of r means that the coefficients decay fast and hence many are close to zero. The case $r = 1/2$ is the worst case, and can be shown to be equivalent to assuming $f_{\mathcal{H}}$ exists. This latter situation corresponds to setting considered in the previous section. We next show how these assumptions allow to derive fast rates.

Theorem 2. *Let $\delta \in (0, 1]$. Under Asm. 1 and the same assumptions of Thm. 1, if $n \geq n_0$, and $\lambda_n = n^{-\frac{1}{2r+\gamma}}$, then a number of random features M equal to*

$$M_n \;=\; c_0\, n^{\frac{1+\gamma(2r-1)}{2r+\gamma}}\; \log \frac{108\kappa^2 n}{\delta},$$

is enough to guarantee, with probability at least $1 - \delta$, that

$$\mathcal{E}(\widehat{f}_{\lambda_n, M_n}) - \mathcal{E}(f_{\mathcal{H}}) \leq c_1 \log^2 \frac{18}{\delta}\, n^{-\frac{2r}{2r+\gamma}},$$

for $r \leq 1$, and where c_0, c_1 do not depend on n, τ, while n_0 does not depends on $n, f_{\mathcal{H}}, \rho$.

The above bound is the same as the one obtained by the full kernel ridge regression estimator and is optimal in a minimax sense [10]. For large r and small γ it approaches a $O(1/n)$ bound. When $\gamma = 1$ and $r = 1/2$ the worst case bound of the previous section is recovered. Interestingly, the number of random features in different regimes is typically smaller than n but can be larger than $O(\sqrt{n})$. Figure. 1 provides a pictorial representation of the number of random features needed for optimal rates in different regimes. In particular $M \ll n$ random features are enough when $\gamma > 0$ and $r > 1/2$. For example for $r = 1, \gamma = 0$ (higher regularity/sparsity and a small RKHS) $O(\sqrt{n})$ are sufficient to get a rate $O(1/n)$. But, for example, if $r = 1/2, \gamma = 0$ (not too much regularity/sparsity but a small RKHS) $O(n)$ are needed for $O(1/n)$ error. The proof suggests that this effect can be a byproduct of sampling features in a data-independent way. Indeed, in the next section we show how much fewer features can be used considering problem dependent sampling schemes.

3.3 Refined Results: Beyond uniform sampling

We show next that fast learning rates can be achieved with fewer random features if they are somewhat *compatible* with the data distribution. This is made precise by the following condition.

Assumption 2 (Compatibility condition). *Define the* maximum random features dimension *as*

$$\mathcal{F}_\infty(\lambda) \;=\; \sup_{\omega \in \Omega}\, \|(L + \lambda I)^{-1/2}\psi(\cdot, \omega)\|_{\rho_X}^2, \quad \lambda > 0. \tag{14}$$

Assume there exists $\alpha \in [0, 1]$, and $F > 0$ such that $\mathcal{F}_\infty(\lambda) \leq F\lambda^{-\alpha}, \quad \forall \lambda > 0$.

The above assumption is abstract and we comment on it before showing how it affects the results. The maximum random features dimension (14) relates the random features to the data-generating distribution through the operator L. It is always satisfied for $\alpha = 1$ ands $F = \kappa^2$. e.g. considering any random feature satisfying (6). The favorable situation corresponds to random features such that case $\alpha = \gamma$. The following theoretical construction borrowed from [23] gives an example.

Example 2 (Problem dependent RF). *Assume K is a kernel with an integral representation (6). For $s(\omega) = \|(L + \lambda I)^{-1/2}\psi(\cdot, \omega)\|_{\rho_X}^{-2}$ and $C_s := \int \frac{1}{s(\omega)} d\pi(\omega)$, consider the random features $\psi_s(x, \omega) = \psi(x, \omega)\sqrt{C_s s(\omega)}$, with distribution $\pi_s(\omega) := \frac{\pi(\omega)}{C_s s(\omega)}$. We show in the Appendix that these random features provide an integral representation of K and satisfy Asm. 2 with $\alpha = \gamma$.*

We next show how random features satisfying Asm. 2 can lead to better resuts.

Theorem 3. *Let $\delta \in (0, 1]$. Under Asm. 2 and the same assumptions of Thm. 1, 2, if $n \geq n_0$, and $\lambda_n = n^{-\frac{1}{2r+\gamma}}$, then a number of random features M_n equal to*

$$M_n \;=\; c_0\, n^{\frac{\alpha+(1+\gamma-\alpha)(2r-1)}{2r+\gamma}}\; \log \frac{108\kappa^2 n}{\delta},$$
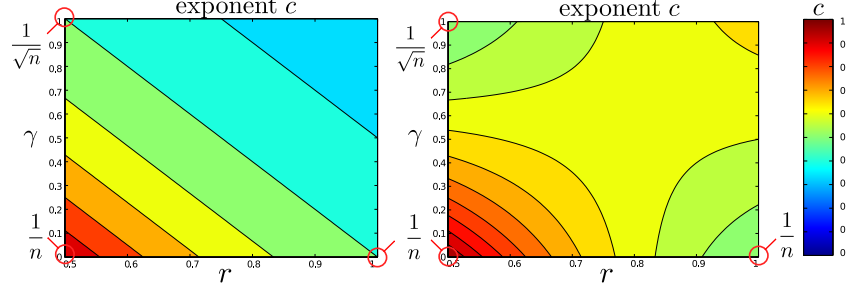
Figure 2: Comparison between the number of features $M = O(n^c)$ required by Nyström (uniform sampling, left) [16] and Random Features ($\alpha = 1$, right), for optimal generalization.

is enough to guarantee, with probability at least $1 - \delta$, that

$$\mathcal{E}(\widehat{f}_{\lambda_n, M_n}) - \mathcal{E}(f_{\mathcal{H}}) \leq c_1 \log^2 \frac{18}{\delta} \, n^{-\frac{2r}{2r+\gamma}},$$

where c_0, c_1 do not depend on n, τ, while n_0 does not depends on $n, f_{\mathcal{H}}, \rho$.

The above learning bound is the same as Thm. 2, but the number of random features is given by a more complex expression depending on α. In particular, in the slow $O(1/\sqrt{n})$ rates scenario, that is $r = 1/2$, $\gamma = 1$, we see that $O(n^{\alpha/2})$ are needed, recovering $O(\sqrt{n})$, since $\gamma \leq \alpha \leq 1$. On the contrary, for a small RKHS, that is $\gamma = 0$ and random features with $\alpha = \gamma$, a constant (!) number of feature is sufficient. A similar trend is seen considering fast rates. For $\gamma > 0$ and $r > 1/2$, if $\alpha < 1$ then the number of random features is always smaller, and potentially much smaller, then the number of random features sampled in a problem independent way, that is $\alpha = 1$. For $\gamma = 0$ and $r = 1/2$, the number of number of features is $O(n^\alpha)$ and can be again just constant if $\alpha = \gamma$. Figure 1 depicts the number of random features required if $\alpha = \gamma$. The above result shows the potentially dramatic effect of problem dependent random features. However the construction in Ex. 2 is theoretical. We comment on this in the next remark.

Remark 4 (Random features leverage scores). *The construction in Ex. 2 is theoretical, however empirical random features leverage scores $\widehat{s}(\omega) = \widehat{v}(\omega)^\top (\mathbf{K} + \lambda n I)^{-1} \widehat{v}(\omega)$, with $\widehat{v}(\omega) \in \mathbb{R}^n$, $(\widehat{v}(\omega))_i = \psi(x_i, \omega)$, can be considered. Statistically, this requires considering an extra estimation step. It seems our proof can be extended to account for this, and we will pursue this in a future work. Computationally, it requires devising approximate numerical strategies, like standard leverage scores* [24].

Comparison with Nyström. This question was recently considered in [22] and our results offer new insights. In particular, recalling the results in [16], we see that in the slow rate setting there is essentially no difference between random features and Nyström approaches, neither from a statistical nor from a computational point of view. In the case of fast rates, Nyström methods with uniform sampling requires $O(n^{-\frac{1}{2r+\gamma}})$ random centers, which compared to Thm. 2, suggests Nyström methods can be advantageous in this regime. While problem dependent random features provide a further improvement, it should be compared with the number of centers needed for Nyström with leverage scores, which is $O(n^{-\frac{\gamma}{2r+\gamma}})$ and hence again better, see Thm. 3. In summary, both random features and Nyström methods achieve optimal statistical guarantees while reducing computations. They are essentially the same in the worst case, while Nyström can be better for benign problems.
Finally we add a few words about the main steps in the proof.

Steps of the proof. The proofs are quite technical and long and are collected in the appendices. They use a battery of tools developed to analyze KRR and related methods. The key challenges in the analysis include analyzing the bias of the estimator, the effect of noise in the outputs, the effect of random sampling in the data, the approximation due to random features and a notion of orthogonality between the function space corresponding to random features and the full RKHS. The last two points are the main elements on novelty in the proof. In particular, compared to other studies, we identify and study the quantity needed to assess the effect of the random feature approximation if the goal is prediction rather than the kernel approximation itself.
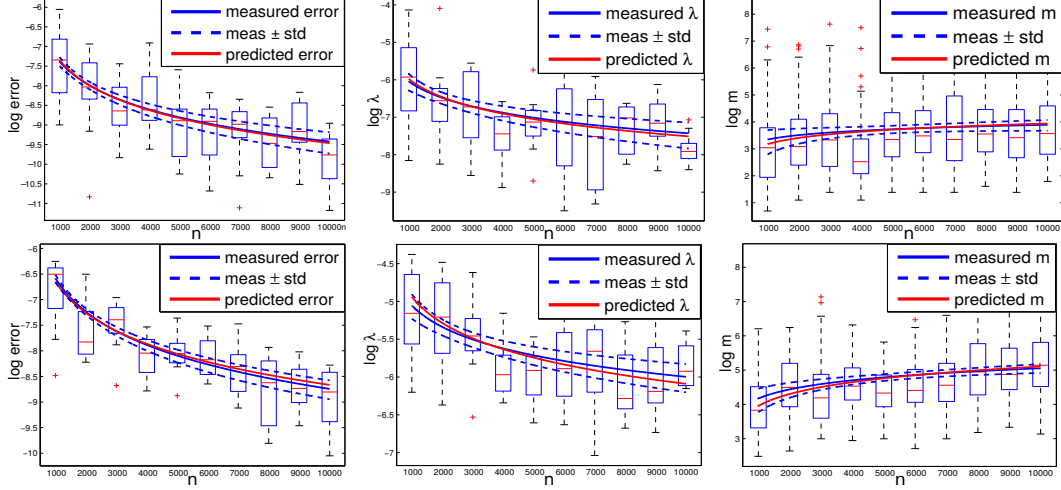
Figure 3: Comparison of theoretical and simulated rates for: excess risk $\mathcal{E}(\widehat{f}_{\lambda,M}) - \inf_{f \in \mathcal{H}} \mathcal{E}(f)$, λ, M, w.r.t. n (100 repetitions). Parameters $r = 11/16, \gamma = 1/8$ (top), and $r = 7/8, \gamma = 1/4$ (bottom).

4 Numerical results

While the learning bounds we present are optimal, there are no lower bounds on the number of random features, hence we present numerical experiments validating our bounds. Consider a spline kernel of order q (see [29] Eq. 2.1.7 when q integer), defined as $\Lambda_q(x, x') = \sum_{k=-\infty}^{\infty} e^{2\pi i k x} e^{-2\pi i k z} |k|^{-q}$, almost everywhere on $[0, 1]$, with $q \in \mathbb{R}$, for which we have $\int_0^1 \Lambda_q(x, z) \Lambda_{q'}(x', z) dz = \Lambda_{q+q'}(x, x')$, for any $q, q' \in \mathbb{R}$. Let $X = [0, 1]$, and ρ_X be the uniform distribution. For $\gamma \in (0, 1)$ and $r \in [1/2, 1]$ let, $K(x, x') = \Lambda_{\frac{1}{\gamma}}(x, x')$, $\psi(\omega, x) = \Lambda_{\frac{1}{2\gamma}}(\omega, x)$, $f_*(x) = \Lambda_{\frac{r}{\gamma} + \frac{1}{2} + \epsilon}(x, x_0)$ with $\epsilon > 0, x_0 \in X$. Let $\rho(y|x)$ be a Gaussian density with variance σ^2 and mean $f^*(x)$. Then Asm 1, 2 are satisfied and $\alpha = \gamma$. We compute the KRR estimator for $n \in \{10^3, \ldots, 10^4\}$ and select λ minimizing the excess risk computed analytically. Then we compute the RF-KRR estimator and select the number of features M needed to obtain an excess risk within 5% of the one by KRR. In Figure 3, the theoretical and estimated behavior of the excess risk, λ and M with respect to n are reported together with their standard deviation over 100 repetitions. The experiment shows that the predictions by Thm. 3 are accurate, since the theoretical predictions estimations are within one standard deviation from the values measured in the simulation.

5 Conclusion

In this paper, we provide a thorough analyses of the generalization properties of random features with ridge regression. We consider a statistical learning theory setting where data are noisy and sampled at random. Our main results show that there are large classes of learning problems where random features allow to reduce computations while preserving optimal statistical accuracy of exact kernel ridge regression. This in contrast with previous state of the art results suggesting computational gains needs to be traded-off with statistical accuracy. Our results open several venues for both theoretical and empirical work. As mentioned in the paper, it would be interesting to analyze random features with empirical leverage scores. This is immediate if input points are fixed, but our approach should allow to also consider the statistical learning setting. Beyond KRR, it would be interesting to analyze random features together with other approaches, in particular accelerated and stochastic gradient methods, or distributed techniques. It should be possible to extend the results in the paper to consider these cases. A more substantial generalization would be to consider loss functions other than quadratic loss, since this require different techniques from empirical process theory.

References

[1] V. Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

[2] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the AMS*, 39:1–49, 2002.

[3] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[4] T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 1990.

[5] A. Pinkus. Approximation theory of the mlp model in neural networks. *Acta Numerica*, 8:143–195, 1999.

[6] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. MIT Press, 2002.

[7] N. Aronszajn. Theory of reproducing kernels. *Transactions of the AMS*, 68(3):337–404, 1950.

[8] G. S. Kimeldorf and G. Wahba. A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502, 1970.

[9] B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *Computational learning theory*, pages 416–426. Springer, 2001.

[10] A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *FoCM*, 2007.

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

[12] C. Williams and M. Seeger. Using the nyström method to speed up kernel machines. In *NIPS*, 2000.

[13] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *NIPS*, 2007.

[14] F. Bach. Sharp analysis of low-rank kernel matrix approximations. In *COLT*, 2013.

[15] A. Alaoui and M. Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *NIPS*. 2015.

[16] A. Rudi, R. Camoriano, and L. Rosasco. Less is more: Nyström computational regularization. In *NIPS*. 2015.

[17] B. K. Sriperumbudur and Z. Szabo. Optimal rates for random fourier features. *ArXiv e-prints*, June 2015.

[18] N. Halko, P. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[19] Yaniv Plan and Roman Vershynin. Dimension reduction by random hyperplane tessellations. *Discrete & Computational Geometry*, 51(2):438–461, 2014.

[20] Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *NIPS*, 2009.

[21] C. Cortes, M. Mohri, and A. Talwalkar. On the impact of kernel approximation on learning accuracy. In *AISTATS*, 2010.

[22] T. Yang, Y. Li, M. Mahdavi, R. Jin, and Z. Zhou. Nyström method vs random fourier features: A theoretical and empirical comparison. In *NIPS*, pages 485–493, 2012.

[23] F. Bach. On the equivalence between quadrature rules and random features. *ArXiv e-prints*, February 2015.

[24] P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *JMLR*, 13:3475–3506, 2012.

[25] Y. Cho and L. K. Saul. Kernel methods for deep learning. In Y. Bengio, D. Schuurmans, J.D. Lafferty, C.K.I. Williams, and A. Culotta, editors, *NIPS*, pages 342–350. 2009.

[26] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer New York, 2008.

[27] S. Smale and D. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

[28] S. Smale and D. Zhou. Estimating the approximation error in learning theory. *Analysis and Applications*, 1(01):17–41, 2003.

[29] G. Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia, 1990.

[30] E. De Vito, L. Rosasco, A. Caponnetto, U. D. Giovannini, and F. Odone. Learning from examples as an inverse problem. In *JMLR*, pages 883–904, 2005.

[31] S. Boucheron, G. Lugosi, and O. Bousquet. Concentration inequalities. In *Advanced Lectures on Machine Learning*. 2004.

[32] V. V. Yurinsky. *Sums and Gaussian vectors*. 1995.

[33] J. A. Tropp. User-friendly tools for random matrices: An introduction. 2012.

[34] S. Minsker. On some extensions of bernstein's inequality for self-adjoint operators. *arXiv*, 2011.

[35] J. Fujii, M. Fujii, T. Furuta, and R. Nakamoto. Norm inequalities equivalent to heinz inequality. *Proceedings of the American Mathematical Society*, 118(3), 1993.

[36] Andrea Caponnetto and Yuan Yao. Adaptation for regularization operators in learning theory. Technical report, DTIC Document, 2006.

[37] Rajendra Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

[38] M. Raginsky and S. Lazebnik. Locality-sensitive binary codes from shift-invariant kernels. In *NIPS*, 2009.

[39] P. Kar and H. Karnick. Random feature maps for dot product kernels. In *AISTATS*, 2012.

[40] N. Pham and R. Pagh. Fast and scalable polynomial kernels via explicit feature maps. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 239–247. ACM, 2013.

[41] Q. Le, T. Sarlós, and A. Smola. Fastfood - computing hilbert space expansions in loglinear time. In *ICML*, 2013.

[42] J. Yang, V. Sindhwani, Q. Fan, H. Avron, and M. Mahoney. Random laplace feature maps for semigroup kernels on histograms. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 971–978. IEEE, 2014.

[43] R. Hamid, Y. Xiao, A. Gittens, and D. Decoste. Compact random feature maps. In *ICML*, pages 19–27, 2014.

[44] J. Yang, V. Sindhwani, H. Avron, and M. W. Mahoney. Quasi-monte carlo feature maps for shift-invariant kernels. In *ICML*, volume 32 of *JMLR Proceedings*, pages 485–493. JMLR.org, 2014.

[45] Andrew Cotter, Joseph Keshet, and Nathan Srebro. Explicit approximations of the gaussian kernel. *arXiv preprint arXiv:1109.4603*, 2011.

[46] Ingo Steinwart, Don Hush, and Clint Scovel. An explicit description of the reproducing kernel hilbert spaces of gaussian rbf kernels. *IEEE Transactions on Information Theory*, 52(10):4635–4643, 2006.

[47] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(3):480–492, 2012.

Generalization Properties of Learning with Random Features
Supplementary Materials

The supplementary materials are divided in the following four section

A. *Proofs* - where the proofs for Section 3 are provided

B. *Concentration Inequalities* - where probabilistic tools necessary for the proofs are recalled

C. *Operator Inequalities* - where some analytic inequalities used in the proofs are recalled

D. *Auxiliary Results* - where some technical lemmas necessary to the proof are derived

E. *Examples of Random Features* - where examples of random features expansion are recalled

A   Proofs

In Sect. A.1, the notation is introduced and some standard identities are recalled. In Sect. A.2, the excess risk is decomposed in five terms (Eq. (17)-(21)) that are further simplified in Lemma 2, 3, 4, 5. The complete decomposition is presented in Thm. 4. In Sect. A.3, the terms in decomposition are bounded in probability, in particular Lemma 7 bounds the variance term, Lemma 8 the computational error term, while Lemma 10 controls the constants. Finally the proofs of the main results are presented in Section A.4 together with the more general results of Thm. 5.

First we recall the assumptions needed to derive the results. They are already presented or implied in the main text, here we collect and number them.

Assumption 2 (Compatibility condition) *There exists $\alpha \in [0, 1]$ and $F > 0$ such that*

$$\mathcal{F}_\infty(\lambda) \le F\lambda^{-\alpha}, \quad \forall \lambda > 0.$$

Assumption 3 (Random Features are bounded and continuous). *The kernel K has an integral representation as in Eq. 6, with ψ continuous in both variables and bounded, that is, there exists $\kappa \ge 1$ such that $|\psi(x, \omega)| \le \kappa$ for any $x, \in X$ and $\omega \in \Omega$. The associated RKHS $\mathcal{H}$ is separable.*

Note that the assumption above is satisfied when the random feature is continuous and bounded and the space X is separable (e.g. $\mathbb{R}^d$, $d \in \mathbb{N}$ or any Polish space). Indeed the continuity of ψ implies the continuity of K, which, together with the separability of X implies the separability of $\mathcal{H}$.

Assumption 4 (Noise on the y is sub-exponential, and there exists $f_\mathcal{H}$). *For any $x \in X$*

$$\mathbb{E}[|y|^p \mid x] \le \frac{1}{2}p!\sigma^2 B^{p-2}, \quad \forall p \ge 2.$$

Moreover there exists $f_\mathcal{H} \in \mathcal{H}$ such that $\mathcal{E}(f_\mathcal{H}) = \inf_{f \in \mathcal{H}} \mathcal{E}(f)$.

Note that the above assumption on y is satisfied when y is bounded, sub-gaussian or sub-exponential. In particular, if $|y| \in [-\frac{b}{2}, \frac{b}{2}]$ almost surely, with $b \in (0, \infty)$ then the assumption above is satisfied with $\sigma = B = b$.

Assumption 5 (Effective dimension). *Let $\lambda > 0$. There exists $Q > 0$ and $\gamma \in [0, 1]$ such that, for any $\lambda > 0$*

$$\mathcal{N}(\lambda) \le Q^2\lambda^{-\gamma}.$$

It is the first part of Asm. 1, for the sake of clarity we need to split it in two, since many results depend either on the first or on the second part.

Assumption 6 (Source condition). *There exists $1/2 \le r \le 1$ and $g \in L^2(X, \rho_X)$ such that*

$$f_\mathcal{H}(x) = (L^r g)(x) \quad a.s.$$

We denote with R the quantity $1 \vee \|g\|_{\rho_X}$.

A.1 Kernel and Random Features Operators

In this section, we provide the notation, recall some useful facts and define some operators used in the rest of the appendix. In the rest of the paper we denote with $\|\cdot\|$ the operatorial norm and with $\|\cdot\|_{HS}$ the Hilbert-Schmidt norm. Let $\mathcal{L}$ be a Hilbert space, we denote with $\langle \cdot, \cdot \rangle_{\mathcal{L}}$ the associated inner product, with $\|\cdot\|_{\mathcal{L}}$ the norm and with $\mathrm{Tr}(\cdot)$ the trace. Let Q be a bounded self-adjoint linear operator on a separable Hilbert space $\mathcal{L}$, we denote with $\lambda_{\max}(Q)$ the biggest eigenvalue of Q, that is $\lambda_{\max}(Q) = \sup_{\|f\|_{\mathcal{L}} \leq 1} \langle f, Qf \rangle_{\mathcal{L}}$. Moreover, we denote with Q_λ the operator $Q + \lambda I$, where Q is a linear operator, $\lambda \in \mathbb{R}$ and I the identity operator, so for example $\widehat{C}_{M,\lambda} := \widehat{C}_M + \lambda I$. Moreover we recall some basic properties of norms in Hilbert spaces.

Remark 5. *Let $V_0, \ldots, V_t$ with $t \in \mathbb{N}$ be Hilbert spaces. Let $q \in V_0$ and $A_i : V_i \to V_{i-1}$ bounded linear operators and $f \in V_t$. We recall that the identity $q = (A_1) \cdots (A_t)(f)$, implies $\|q\|_{V_0} \leq \|A_1\| \ldots \|A_t\| \|f\|_{V_t}$.*

Let X be a probability space and ρ be a probability distribution on $X \times \mathbb{R}$ satisfying Assumption 3. We denote ρ_X its marginal on X and $\rho(y|x)$ the conditional distribution on $\mathbb{R}$. Let $L^2(X, \rho_X)$ be the Lebesgue space of square ρ_X-integrable functions, with the canonical inner product

$$\langle g, h \rangle_{\rho_X} = \int_X g(x)h(x)d\rho_X(x), \quad \forall g, h \in L^2(X, \rho_X),$$

and the norm $\|g\|_{\rho_X}^2 = \langle g, g \rangle_{\rho_X}$, for all $g \in L^2(X, \rho_X)$. Let (Ω, π) be a probability space and $\psi : \Omega \times X \to \mathbb{R}$ be a continuous and bounded map as in Asm. 3. Moreover let the kernel K be defined by Eq. (6). We denote with K_x the function $K(x, \cdot)$, for any $x \in X$. Then the Reproducing Kernel Hilbert Space $\mathcal{H}$ induced by K is defined by

$$\mathcal{H} = \overline{\mathrm{span}\{K_x \mid x \in X\}}, \quad \text{completed with} \quad \langle K_x, K_{x'} \rangle_{\mathcal{H}} = K(x, x') \ \forall x, x' \in X.$$

We now define the operators needed in the rest of the proofs. Let $n \in \mathbb{N}$, and $(x_1, y_1), \ldots, (x_n, y_n) \in X \times \mathbb{R}$ be sampled independently according to ρ.

Definition 1. *Let $P : L^2(X, \rho_X) \to L^2(X, \rho_X)$ be the projection operator with the same range of L. Let $f_\rho : X \to \mathbb{R}$ be defined as*

$$f_\rho(x) = \int y d\rho(y|x) \ a. \ e.$$

We now recall a useful characterization of the excess risk, in term of the quantities defined above.

Remark 6 (from [2, 30]). *When $\int y^2 d\rho$ is finite, then $f_\rho \in L^2(X, \rho_X)$ and f_ρ is the minimizer of $\mathcal{E}$ over all the measurable functions. When $\int K(x, x)d\rho_X$ is finite, the range of P is the closure of $\mathcal{H}$ in $L^2(X, \rho_X)$, that this the closure of the range of L. When both conditions hold, for any $f \in L^2(X, \rho_X)$ the following hold*

$$\mathcal{E}(f) - \inf_{g \in \mathcal{H}} \mathcal{E}(g) = \|f - Pf_\rho\|_{\rho_X}^2.$$

Moreover if there exists $f_{\mathcal{H}} \in \mathcal{H}$ minimizing $\mathcal{E}$, then Asm. 6 is equivalent to requiring the existence of $r \geq 1/2$, $g \in L^2(X, \rho_X)$ such that

$$Pf_\rho = L^r g, \tag{15}$$

with $R := \|g\|_{L^2(X, \rho_X)}$.

In the following we define analogous operators for the approximated kernel $K_M := \phi_M(x)^\top \phi_M(x')$, with

$$\phi_M(x) := M^{-1/2}(\psi(x, \omega_1), \ldots, \psi(x, \omega_M)),$$

for any $x, x' \in X$, where $M \in \mathbb{N}$ and $\omega_1, \ldots, \omega_M \in \Omega$ are sampled independently according to π. We denote with ψ_ω the function $\psi(\cdot, \omega)$ for any $\omega \in \Omega$. According to the following remark, we have that $\psi_{\omega_i} \in L^2(X, \rho_X)$ almost surely.

Remark 7. *Under Asm. 3 and the fact that ρ is a finite measure, $\psi_\omega \in L^2(X, \rho_X)$ almost surely.*

Now we are ready for defining the following operators, depending on ϕ_M or K_M.

Definition 2. *For all $g \in L^2(X, \rho_X)$, $\beta \in \mathbb{R}^M$, $\alpha \in \mathbb{R}^n$ and $i \in \{1, \ldots, M\}$, we have*

- $S_M : \mathbb{R}^M \to L^2(X, \rho_X)$, $(S_M \beta)(\cdot) = \phi_M(\cdot)^\top \beta$,

- $S_M^* : L^2(X, \rho_X) \to \mathbb{R}^M$, $(S_M^* g)_i = \frac{1}{\sqrt{M}} \int_X \psi_{\omega_i}(x) g(x) d\rho_X(x)$,

- $L_M : L^2(X, \rho_X) \to L^2(X, \rho_X)$, $(L_M g)(\cdot) = \int_X K_M(\cdot, z) g(z) d\rho_X(z)$.

- $C_M : \mathbb{R}^M \to \mathbb{R}^M$, $C_M = \int_X \phi_M(x) \phi_M(x)^\top d\rho_X(x)$,

- $\widehat{C}_M : \mathbb{R}^M \to \mathbb{R}^M$, $\widehat{C}_M = \frac{1}{n} \sum_{i=1}^n \phi_M(x_i) \phi_M(x_i)^\top$.

Note that the operators above satisfy the properties in the following remark.

Remark 8 (from [10])**.** *Under Asm. 3 the linear operators L is trace class and $L_M, C_M, S_M, \widehat{C}_M, \widehat{S}_M$ are finite dimensional. Moreover we have that $L = SS^*$, $L_M = S_M S_M^*$, $C_M = S_M^* S_M$ and $\widehat{C}_M = \widehat{S}_M^* \widehat{S}_M$. Finally $L, L_M, C_M, \widehat{C}_M$ are self-adjoint and positive operators, with spectrum is $[0, \kappa^2]$.*

In the next remark we rewrite $\widehat{f}_{\lambda, M}$ in terms of the operators introduced above.

Remark 9. *Let $\widehat{f}_{\lambda, M}$ defined as in Eq. 7. Under Assumption 3, $\widehat{f}_{\lambda, M} \in L^2(X, \rho_X)$ almost surely, since ψ_ω is in $L^2(X, \rho_X)$ almost surely (Rem. 7) and $\widehat{f}_{\lambda, M}$ is a linear combination of $\psi_{\omega_1}, \ldots, \psi_{\omega_M}$. In particular,*

$$\widehat{f}_{\lambda, M} = S_M \widehat{C}_{M, \lambda}^{-1} \widehat{S}_M^* \widehat{y}.$$

A.2   Analytic Result

In this subsection we decompose analytically the excess risks in different terms, that will be bounded via concentration inequalities in the next section. Under Asm. 3, since $\widehat{f}_{\lambda, M} \in L^2(X, \rho_X)$ almost surely, we have

$$\mathcal{E}(\widehat{f}_{\lambda, M}) - \inf_{f \in \mathcal{H}} \mathcal{E}(f) = \|\widehat{f}_{\lambda, M} - P f_\rho\|_{\rho_X}^2, \tag{16}$$

(for more details see Rem. 6, 9). In our analysis we decompose the excess risk in the following five terms

$$\widehat{f}_{\lambda, M} - P f_\rho = \widehat{f}_{\lambda, M} - S_M \widehat{C}_{M, \lambda}^{-1} S_M^* f_\rho \tag{17}$$

$$+ \ S_M \widehat{C}_{M, \lambda}^{-1} S_M^* (I - P) f_\rho \tag{18}$$

$$+ \ S_M \widehat{C}_{M, \lambda}^{-1} S_M^* P f_\rho - L_M L_{M, \lambda}^{-1} P f_\rho \tag{19}$$

$$+ \ L_M L_{M, \lambda}^{-1} P f_\rho - L L_\lambda^{-1} P f_\rho \tag{20}$$

$$+ \ L L_\lambda^{-1} P f_\rho - P f_\rho. \tag{21}$$

The first controls the variance of the outputs y, the second the interaction between the space of models spanned by $\psi_{\omega_1}, \ldots, \psi_{\omega_M}$ and $\mathcal{H}$, the third the approximation of the inverse covariance operator $\widehat{C}_{M, \lambda}^{-1}$, the fourth controls how close is the integral operator L_M to L, while the last controls the approximation error of the models in $\mathcal{H}$. The $L^2(X, \rho_X)$ norm of $\widehat{f}_{\lambda, M} - P f_\rho$ is bounded by the sum of the $L^2(X, \rho_X)$ norms of the terms, that are further bounded in Lemma. 2, 3, 4, 5. The final analytical decomposition is given in Thm. 4. First we need a preliminary result.

Lemma 1. *Under Asm. 3, the operator L is characterized by*

$$L = \int \psi_\omega \otimes \psi_\omega d\pi(\omega).$$

Proof. By Asm. 3, we have that $\psi_\omega \in L^2(X, \rho_X)$ almost surely and uniformly bounded. By using the kernel expansion of Eq. (6), the linearity of the Bochner integral and of the dot product, we have

that for any $f, g \in L^2(X, \rho_X)$ the following holds

$$\langle f, Lg \rangle_{\rho_X} = \int f(x) K(x, z) g(z) d\rho_X(x) d\rho_X(z)$$

$$= \int f(x) \psi(x, \omega) \psi(z, \omega) g(z) d\rho_X(x) d\rho_X(z) d\pi(\omega)$$

$$= \int \langle f, \psi_\omega \rangle_{\rho_X} \langle g, \psi_\omega \rangle_{\rho_X} \ d\pi(\omega) = \left\langle f, \int \psi_\omega \langle g, \psi_\omega \rangle_{\rho_X} \ d\pi(\omega) \right\rangle_{\rho_X}$$

$$= \left\langle f, \left(\int \psi_\omega \otimes \psi_\omega d\pi(\omega) \right) g \right\rangle_{\rho_X}.$$

$\square$

Now we are ready to prove that the second term of the expansion in Eq. 17 is zero. We obtain this result by proving that $\|(I - P)\psi_\omega\| = 0$ almost everywhere.

Lemma 2. *Under Asm. 3, the following holds for any $\lambda > 0, M, n \in \mathbb{N}$,*

$$\|S_M \widehat{C}_{M,\lambda}^{-1} S_M^* (I - P) f_\rho\|_{\rho_X} = 0 \ a. \ s.$$

Proof. Note that, since P is the projection operator on the closure of the range of L and L is trace class, then $(I - P)L = 0$, this implies that $\text{Tr}((I - P)L(I - P)) = 0$. By the characterization of L given in Lemma 1, the linearity of the bounded operator $I - P$ and of the trace, we have that

$$0 = \text{Tr}\left((I - P)L(I - P)\right) = \text{Tr}\left((I - P)\left(\int \psi_\omega \otimes \psi_\omega d\pi(\omega)\right)(I - P)\right)$$

$$= \int \text{Tr}\left((I - P)(\psi_\omega \otimes \psi_\omega)(I - P)\right) d\pi(\omega)$$

$$= \int \|(I - P)\psi_\omega\|_{\rho_X}^2 d\pi(\omega),$$

where the last step is due to the fact that $\text{Tr}(A(v \otimes v)A) = \text{Tr}(Av \otimes Av) = \|Av\|_{\rho_X}^2$ for any bounded self adjoint operator A and any function $v \in L^2(X, \rho_X)$. The equation above implies that $(I - P)\psi_\omega = 0$ almost surely on the support of π. Now we study $S_M^*(I - P)$, for any $\beta \in \mathbb{R}^M$ and any $f \in L^2(X, \rho_X)$ we have

$$\langle \beta, S_M^*(I - P)f \rangle_{\mathbb{R}^M} = \frac{1}{\sqrt{M}} \sum_{i=1}^{M} \beta_i \langle (I - P)\psi_{\omega_i}, f \rangle_{\rho_X} = 0 \ a. \ s.,$$

where the last step is due to the fact that $\langle 0, v \rangle = 0$ for any v and $(I - P)\psi_{\omega_i} = 0$ almost surely, since ω_i are distributed according to π and $(I - P)\psi_\omega = 0$ almost surely on the support of π. Now

$$\|S_M \widehat{C}_{M,\lambda}^{-1} S_M^* (I - P) f_\rho\|_{\rho_X} \leq \|S_M \widehat{C}_{M,\lambda}^{-1}\| \|S_M^*(I - P)\| \|f_\rho\|_{\rho_X} = 0 \ a. \ s.$$

$\square$

Lemma 3. *Under Asm. 3, and Eq. (15) the following holds for any $\lambda > 0, M, n \in \mathbb{N}$*

$$\|S_M \widehat{C}_{M,\lambda}^{-1} S_M^* P f_\rho - L_M L_{M,\lambda}^{-1} P f_\rho\| \leq R\kappa^{2r-1} \|L_{M,\lambda}^{-1/2} L^{1/2}\| \|S_M \widehat{C}_{M,\lambda}^{-1} C_{M,\lambda}^{-1/2}\| \|C_{M,\lambda}^{-1/2}(C_M - \widehat{C}_M)\|.$$

Proof. First of all we recall that $Z^* f(ZZ^*) = f(Z^*Z)Z^*$ for any continuous spectral function and any compact operator Z. By the characterization of L_M in Rem. 8 under Asm. 3, we have

$$L_M L_{M,\lambda}^{-1} = S_M S_M^* (S_M S_M^* + \lambda I)^{-1} = S_M (S_M^* S_M + \lambda I)^{-1} S_M^* = S_M C_{M,\lambda}^{-1} S_M^*,$$

since $(\cdot + \lambda I)^{-1}$ is a continuos spectral function on $[0, \infty)$, which contains the spectrum of L that is in $[0, \kappa^2]$. Equivalently, the equation above could be proven algebraically via the Woodbury identity. Now we have

$$(S_M \widehat{C}_{M,\lambda}^{-1} S_M^* - L_M L_{M,\lambda}^{-1}) P f_\rho = S_M (\widehat{C}_{M,\lambda}^{-1} - C_{M,\lambda}^{-1}) S_M^* P f_\rho = S_M \widehat{C}_{M,\lambda}^{-1} (C_M - \widehat{C}_M) C_{M,\lambda}^{-1} S_M^* P f_\rho,$$

where the last step is due to the identity $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ valid for any bounded invertible linear operator A, B. In particular by multiplying and dividing by $C_{M,\lambda}^{1/2}$ we have the following decomposition

$$S_M \widehat{C}_{M,\lambda}^{-1}(C_M - \widehat{C}_M)C_{M,\lambda}^{-1}S_M^* Pf_\rho = (S_M \widehat{C}_{M,\lambda}^{-1}C_{M,\lambda}^{1/2}) \ (C_{M,\lambda}^{-1/2}(C_M - \widehat{C}_M)) \ (C_{M,\lambda}^{-1}S_M^* Pf_\rho).$$

The result is given by bounding the norm of the lhs of the identity above, by the product of the norms of the parentheses on the rhs (see Rem. 5). Note that by applying Eq. (15), we have that there exists $g \in L^2(X, \rho_X)$, such that $Pf_\rho = L^r g$ and by dividing and multiplying for $L_{M,\lambda}^{1/2}$, we have

$$C_{M,\lambda}^{-1}S_M^* Pf_\rho = (C_{M,\lambda}^{-1}S_M^* L_{M,\lambda}^{1/2}) \ (L_{M,\lambda}^{-1/2}L^{1/2}) \ L^{r-1/2} \ g.$$

Now note that, by Asm. 6, we have $r \geq 1/2$, $\|g\|_{\rho_X} \leq R$ and $\|L^{r-1/2}\| \leq \kappa^{2r-1}$ since L is compact with the spectrum in $[0, \kappa^2]$ and $2r - 1 \geq 0$. By the fact that $(\cdot + \lambda I)^{-2}$ is a continuous spectral function on $[0, \infty)$ containing the spectrum of C_M, we have that $S_M C_{M,\lambda}^{-2} S_M^* = L_{M,\lambda}^{-2} L_M$ and so for any $\lambda > 0$

$$\|C_{M,\lambda}^{-1}S_M^* L_M^{1/2}\|^2 = \|L_M^{1/2} S_M C_{M,\lambda}^{-2} S_M^* L_M^{1/2}\| = \|L_{M,\lambda}^{-2} L_M^2\| \leq 1.$$

$\square$

Lemma 4. *Under Asm. 3, and Eq. (15) the following holds for any $\lambda > 0, M \in \mathbb{N}$*

$$\|(LL_\lambda^{-1} - L_M L_{M,\lambda}^{-1})Pf_\rho\| \leq R\sqrt{\lambda}\|L_{M,\lambda}^{-1/2}L_\lambda^{1/2}\|\|L_\lambda^{-1/2}(L - L_M)\|^{2r-1}\|L_\lambda^{-1/2}(L - L_M)L_\lambda^{-1/2}\|^{2-2r}$$

Proof. By the algebraic identities $A(A + \lambda I) = I - \lambda(A + \lambda I)^{-1}$ valid for any bounded positive operator and $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ valid for any invertible bounded operators, we have

$$(LL_\lambda^{-1} - L_M L_{M,\lambda}^{-1})Pf_\rho = \lambda(L_{M,\lambda}^{-1} - L_\lambda^{-1})Pf_\rho = \lambda L_{M,\lambda}^{-1}(L - L_M)L_\lambda^{-1}Pf_\rho.$$

By applying Eq. (15), we have that there exists $g \in L^2(X, \rho_X)$, such that $Pf_\rho = L^r g$, so by multiplying and dividing by $L_M^{1/2}$, we preform the following decomposition

$$(LL_\lambda^{-1} - L_M L_{M,\lambda}^{-1})Pf_\rho = \sqrt{\lambda} \ (\sqrt{\lambda}L_{M,\lambda}^{-1/2}) \ (L_{M,\lambda}^{-1/2}L_\lambda^{1/2}) \ (L_\lambda^{-1/2}(L - L_M)L_\lambda^{-(1-r)}) \ (L_\lambda^{-r}L^r) \ g.$$

The result is given by bounding the norm of the lhs of the identity above, by the product of the norms of the parentheses on the rhs (see Rem. 5). Note that $\|\sqrt{\lambda}L_{M,\lambda}^{-1/2}\| \leq 1$ and $\|L_\lambda^{-r}L^r\| \leq 1$ for any $\lambda > 0$ and $\|g\|_{\rho_X} \leq R$. Now we apply Proposition 9 on $\|L_\lambda^{-1/2}(L - L_M)L_\lambda^{-(1-r)}\|$, indeed note that $0 \leq 1 - r \leq 1/2$, so by setting $\sigma = 2 - 2r$, $X = L_\lambda^{-1/2}(L - L_M)$, $A = L_\lambda^{-1/2}$ and applying the proposition, we have

$$\|L_\lambda^{-1/2}(L - L_M)L_\lambda^{-\sigma/2}\| \leq \|L_\lambda^{-1/2}(L - L_M)\|^{2r-1}\|L_\lambda^{-1/2}(L - L_M)L_\lambda^{-1/2}\|^{2-2r}.$$

$\square$

Lemma 5. *Under Asm. 3, and Eq. (15) the following holds for any $\lambda > 0$,*

$$\|LL_\lambda^{-1}Pf_\rho - Pf_\rho\| \leq R\lambda^r.$$

Proof. By the identity $A(A + \lambda I)^{-1} = I - \lambda(A + \lambda)^{-1}$ valid for $\lambda > 0$ and any bounded self-adjoint positive operator and by Eq. (15) for which there exists $g \in L^2(X, \rho_X)$ such that $Pf_\rho = L^r g$, we have

$$(I - LL_\lambda^{-1})Pf_\rho = \lambda L_\lambda^{-1}Pf_\rho = \lambda L_\lambda^{-1}L^r g = \lambda^r \ (\lambda^{1-r}L_\lambda^{-(1-r)}) \ (L_\lambda^{-r}L^r) \ g. \quad (22)$$

The result is given by bounding the norm of the lhs of Eq. 22, by the product of the norms of the parentheses on the rhs (see Rem. 5). Note that $\|\lambda^{1-r}L_\lambda^{-(1-r)}\| \leq 1$ and $\|L_\lambda^{-r}L^r\| \leq 1$, while $R := \|g\|_{\rho_X}$ according to Eq. (15). $\square$

Theorem 4 (Analytic Decomposition). *Under Assumptions 3 and Eq. (15) let $\widehat{f}_{\lambda,M}$ as in Eq. 7. For any $\lambda > 0$ and $M \in \mathbb{N}$, the following holds*

$$|\mathcal{E}(\widehat{f}_{\lambda,M}) - \inf_{f \in \mathcal{H}} \mathcal{E}(f)|^{1/2} \;\leq\; \beta \, (\; \underbrace{\mathcal{S}(\lambda, M, n)}_{\text{Sample Error}} \;+\; \underbrace{\mathcal{C}(\lambda, M)}_{\text{Computational Error}} \;+\; \underbrace{R\lambda^v}_{\text{Approximation Error}} \;) \qquad (23)$$

where $v = \min(r, 1)$,

1. $\mathcal{S}(\lambda, M, n) := \|C_{M,\lambda}^{-1/2}(\widehat{S}_M^* \widehat{y} - S_M^* f_\rho)\|_{\rho_X} + R\kappa^{2r-1}\|C_{M,\lambda}^{-1/2}(C_M - \widehat{C}_M)\|$,

2. $\mathcal{C}(\lambda, M) := R\sqrt{\lambda}\|L_\lambda^{-1/2}(L - L_M)\|^{2v-1}\|L_\lambda^{-1/2}(L - L_M)L_\lambda^{-1/2}\|^{2-2v}$,

3. $\beta := \max(1, (1 - \beta_1)^{-1}) \max(1, (1 - \beta_2)^{-1/2})$, *with* $\beta_1 := \lambda_{\max}(C_{M,\lambda}^{-1/2}(C_M - \widehat{C}_M)C_{M,\lambda}^{-1/2})$ *and* $\beta_2 := \lambda_{\max}(L_\lambda^{-1/2}(L - L_M)L_\lambda^{-1/2})$.

Proof. Under Asm. 3, the excess risk is characterized by Eq. 16 as we recalled at the beginning of this subsection. We decomposed the quantity $\widehat{f}_{\lambda,M} - Pf_\rho$ according to the terms in Eq. 17-21. The $L^2(X, \rho_X)$ norm of $\widehat{f}_{\lambda,M} - Pf_\rho$ is bounded by the sum of the $L^2(X, \rho_X)$ norms of the terms, that are further bounded in Lemma. 2, 3, 4, 5. In particular, for the first term, by writing $\widehat{f}_{\lambda,M}$ in terms of the linear operators in Def. 2 (see Rem. 9) and by multipling and dividing by $C_{M,\lambda}^{1/2}$ we have

$$\widehat{f}_{\lambda,M} - S_M\widehat{C}_{M,\lambda}^{-1}S_M^* f_\rho = S_M\widehat{C}_{M,\lambda}^{-1}(\widehat{S}_M^* \widehat{y} - S_M^* f_\rho) \;=\; (S_M\widehat{C}_{M,\lambda}^{-1}C_{M,\lambda}^{1/2})\,(C_{M,\lambda}^{-1/2}(\widehat{S}_M^* \widehat{y} - S_M^* f_\rho)),$$

then we bound the norm of the term with the norm of the parenthesis in the decomposition above (see Rem. 5). By collecting the result above with the bounds in Lemma. 2, 3, 4, 5, we have

$$|\mathcal{E}(\widehat{f}_{\lambda,M}) - \inf_{f \in \mathcal{H}} \mathcal{E}(f)|^{1/2} \leq b_1 A + b_1 b_2 B + b_3 \mathcal{C}(\lambda, M) + D,$$

where $b_1 := \|S_M\widehat{C}_{M,\lambda}^{-1}C_{M,\lambda}^{1/2}\|$, $b_2 := \|L_{M,\lambda}^{-1/2}L^{1/2}\|$, $b_3 := \|L_{M,\lambda}^{-1/2}L_\lambda^{1/2}\|$,
$A := \|C_{M,\lambda}^{-1/2}(\widehat{S}_M^* \widehat{y} - S_M^* f_\rho)\|_{\rho_X}$, $B := R\kappa^{2r-1}\|C_{M,\lambda}^{-1/2}(C_M - \widehat{C}_M)\|$ and $D := R\lambda^r$. Now note that $b_2 \leq b_3$ for any $\lambda > 0$ since, for any X, T, bounded linear operators, with T positive, by multiplying and dividing for T_λ the following holds

$$\|XT\| \leq \|XT_\lambda\|\|T_\lambda^{-1}T\|, \qquad (24)$$

and $\|T_\lambda^{-1}T\| \leq 1$, for any $\lambda > 0$. Since $A, B, \mathcal{C}(\lambda, M), D$ will contribute to the rates of the bound, while b_1, b_3 are responsible for the numerical constants, we are going to bound the excess risk in order to collect b_1, b_3 in a multiplicative term as follows

$$|\mathcal{E}(\widehat{f}_{\lambda,M}) - \inf_{f \in \mathcal{H}} \mathcal{E}(f)|^{1/2} \leq \max(1, b_1) \max(1, b_3)(A + B + \mathcal{C}(\lambda, M) + D).$$

Finally, we further simplify b_1, b_3, in particular we apply Prop. 8 in the appendix, obtaining $b_3 \leq (1 - \beta_2)^{-1/2}$. For b_1 note that,

$$\|S_M\widehat{C}_{M,\lambda}^{-1}C_{M,\lambda}^{1/2}\| \leq \|S_M\widehat{C}_{M,\lambda}^{-1/2}\|\|\widehat{C}_{M,\lambda}^{-1/2}C_{M,\lambda}^{1/2}\| \leq \|\widehat{C}_{M,\lambda}^{-1/2}C_{M,\lambda}^{1/2}\|^2,$$

since, $\|S_M\widehat{C}_{M,\lambda}^{-1/2}\| \leq \|C_{M,\lambda}^{1/2}\widehat{C}_{M,\lambda}^{-1/2}\|$ for the same reasoning in Eq. (24). Then we apply Prop. 8 in the appendix, obtaining $b_1 \leq (1 - \beta_1)^{-1}$. $\qquad\square$

In the next subsection, we are going to find probabilistic estimates for terms in the analytic decomposition of the excess risk in Thm. 4.

A.3 Probabilistic Estimates

In this section we provide bounds in probability for the quantities $\beta, \mathcal{S}, \mathcal{C}$ of Thm. 4 and for the empirical effective dimension. The notation is introduced in Sect. A.1. First, we fix the notation on the random variables used in the rest of the subsection. Recall that $\beta, \mathcal{S}, \mathcal{C}$ are expressed with respect

to the random variables $\boldsymbol{z} := ((x_1, y_1), \ldots, (x_n, y_n))$, and $\boldsymbol{\omega} := (\omega_1, \ldots, \omega_M)$. The associated sample space is $W := Z \times \Omega^M$ and $Z := (X \times \mathbb{R})^n$, with probability measure $\mathbb{P} := \rho^{\otimes n} \otimes \pi^{\otimes M}$. In particular let $Q \subseteq W$ be an event, we denote with $Q|\boldsymbol{\omega}$ the subset of Z associated to the event Q given $\boldsymbol{\omega}$, that is $Q|\boldsymbol{\omega} := \{\boldsymbol{z} \mid (\boldsymbol{z}, \boldsymbol{\omega}) \in Q\}$. By denoting $\rho^{\otimes n}$ with $\mathbb{P}_Z$ and $\pi^{\otimes M}$ by $\mathbb{P}_\Omega$, we recall that

$$\mathbb{P}(Q) = \int_{\Omega^M} \mathbb{P}_Z(Q|\boldsymbol{\omega}) d\mathbb{P}_\Omega(\boldsymbol{\omega}). \tag{25}$$

Moreover, we recall the following basic facts about $\mathcal{F}_\infty(\lambda)$ and $\mathcal{N}(\lambda)$. We can characterize the upper and lower bounds for $\mathcal{F}_\infty(\lambda)$, in particular we have that $\mathcal{F}_\infty(\lambda) \leq \kappa^2 \lambda^{-1}$ when ψ is uniformly bounded by κ (see Asm. 3), moreover $\mathcal{F}_\infty(\lambda) \geq \mathcal{N}(\lambda)$ indeed $\mathcal{N}(\lambda)$ is characterized by $\mathcal{N}(\lambda) = \mathbb{E}_\omega \|(L + \lambda I)^{-1/2} \psi_\omega\|_{\rho_X}^2$ (see Eq. 30), so

$$\mathcal{N}(\lambda) = \mathbb{E}_\omega \|(L + \lambda I)^{-1/2} \psi_\omega\|_{\rho_X}^2 \leq \sup_{\omega \in \Omega} \|(L + \lambda I)^{-1/2} \psi_\omega\|_{\rho_X}^2 = \mathcal{F}_\infty(\lambda).$$

A.3.1 Estimates for $\mathcal{S}(\lambda, M, n)$

The next lemma bounds the first term of $\mathcal{S}(\lambda, M, n)$ and use a similar technique to the one in [10], while Lemma 7 bounds the whole $\mathcal{S}(\lambda, M, n)$. First we need to introduce $\mathcal{N}_M(\lambda)$ that is the effective dimension induced by the kernel K_M. For any $\lambda > 0$ define $\mathcal{N}_M(\lambda)$ as follows,

$$\mathcal{N}_M(\lambda) := \mathrm{Tr}((L_M + \lambda I)^{-1} L_M).$$

In Prop. 10 in the appendix, we bound $\mathcal{N}_M(\lambda)$ in terms of the effective dimension $\mathcal{N}(\lambda)$ that is the one associated to the kernel K. Prop. 10 refines the result of Prop. 1 of [16], with simpler proof and slightly improved constants.

Lemma 6. *Let $\delta \in (0, 1]$, $n, M \in \mathbb{N}$ and $\lambda > 0$. Given $\omega_1, \ldots, \omega_M \in \Omega$, under Assumptions 3, 4, the following holds with probability at least $1 - \delta$*

$$\|C_{M,\lambda}^{-1/2}(\widehat{S}_M^* \widehat{y} - S_M^* f_\rho)\| \leq 2 \left(\frac{B\kappa}{\sqrt{\lambda}n} + \sqrt{\frac{\sigma^2 \mathcal{N}_M(\lambda)}{n}} \right) \log \frac{2}{\delta}.$$

Proof. In this proof we bound the quantity under study, by using the Bernstein inequality for sum of zero-mean random vectors (see Prop. 2 in the appendix). Since $\widehat{S}_M^* \widehat{y} = n^{-1} \sum_{i=1}^n \phi_M(x_i) y_i$ (see Def. 2) we have

$$C_{M,\lambda}^{-1/2}(\widehat{S}_M^* \widehat{y} - S_M^* f_\rho) = \frac{1}{n} \sum_{i=1}^n \zeta_i,$$

where $\zeta_1, \ldots, \zeta_n$ are defined as $\zeta_i = z_i - \mu$ with $z_i := C_{M,\lambda}^{-1/2} \phi_M(x_i) y_i$, and $\mu \in \mathbb{R}^M$ defined as $\mu := C_{M,\lambda}^{-1/2} S_M^* f_\rho$, for $1 \leq i \leq n$. Note that $\zeta_1, \ldots, \zeta_n$ are independent and identically distributed random vectors given $\omega_1, \ldots, \omega_M$, since $(x_1, y_1), \ldots, (x_n, y_n)$ are assumed i.i.d. with respect to ρ. Moreover note that, by definition of f_ρ,

$$\int \phi_M(x) y \, d\rho(x, y) = \int y \, d\rho(y|x) d\rho_X(x) = \int \phi_M(x) f_\rho(x) d\rho_X(x) = S_M^* f_\rho.$$

So, by linearity of the expectation the ζ_i,

$$\mathbb{E} z_i = C_{M,\lambda}^{-1/2} \int \phi_M(x) y \, d\rho(x, y) = C_{M,\lambda}^{-1/2} S_M^* f_\rho = \mu,$$

which implies that $\zeta_i = z_i - \mu$ is a zero-mean random variable for $1 \leq i \leq n$. Let z be another random variable independent and identically distributed as the z_i's. To apply the Bernstein inequality for random vectors, we need to bound their moments. First of all note that for any $p \geq 1$

$$\begin{aligned}
\mathbb{E} \|\zeta_i\|^p &= \mathbb{E} \|z_i - \mu\|^p = \mathbb{E} \|z_i - \mathbb{E} z\|^p \\
&\leq \mathbb{E}_{z_i} \mathbb{E}_z \|z_i - z\|^p \leq 2^{p-1} \mathbb{E}_{z_i} \mathbb{E}_z (\|z_i\|^p + \|z\|^p) = 2^p \mathbb{E}_z \|z\|^p.
\end{aligned}$$

18

In particular, by applying Asm. 3 and Asm. 4, we have

$$\mathbb{E}_z \|z\|^p = \int_{X \times \mathbb{R}} \|C_{M,\lambda}^{-1/2} \phi_M(x) y\|^p d\rho(x,y) = \int_X \|C_{M,\lambda}^{-1/2} \phi_M(x)\|^p \int |y|^p d\rho(y|x) \, d\rho_X(x)$$

$$\leq \frac{1}{2} p! \sigma^2 B^{p-2} \int_X \|C_{M,\lambda}^{-1/2} \phi_M(x)\|^p d\rho_X(x)$$

$$\leq \frac{1}{2} p! \sigma^2 B^{p-2} \left(\sup_{x \in X} \|C_{M,\lambda}^{-1/2} \phi_M(x)\|^{p-2} \right) \int_X \|C_{M,\lambda}^{-1/2} \phi_M(x)\|^2 d\rho_X$$

$$= \frac{1}{2} p! \sqrt{J(\lambda)\sigma^2}^2 \left(\frac{B\kappa}{\sqrt{\lambda}} \right)^{p-2}.$$

where $J(\lambda) = \int_X \|C_{M,\lambda}^{-1/2} \phi_M(x)\|^2 d\rho_X(x)$, while $\|C_{M,\lambda}^{-1/2} \phi_M(x)\| \leq \kappa/\sqrt{\lambda}$ a.s. is given by

$$\|C_{M,\lambda}^{-1/2} \phi_M(x)\|^2 \leq \frac{1}{\lambda} \sup_{x \in X} \|\phi_M(x)\|^2 = \frac{1}{\lambda M} \sup_{x \in X} \sum_{i=1}^{M} |\psi_{\omega_i}(x)|^2$$

$$\leq \frac{1}{\lambda M} \sum_{i=1}^{M} \sup_{x \in X} |\psi_{\omega_i}(x)|^2 \leq \frac{1}{\lambda M} \sum_{i=1}^{M} \sup_{\omega \in \Omega, x \in X} |\psi_\omega(x)|^2 \leq \left(\frac{\kappa}{\sqrt{\lambda}} \right)^2,$$

where the last step is due to Asm. 3. Finally, to concentrate the sum of random vectors, we apply Prop. 2. To conclude the proof we need to prove that $J(\lambda) = \mathcal{N}_M(\lambda)$. Note that, by Rem. 8, we have that $L_M = S_M S_M^*$ and $C_M = S_M^* S_M$, so

$$\mathcal{N}_M(\lambda) = \operatorname{Tr} L_M L_{M,\lambda}^{-1} = \operatorname{Tr} S_M^* L_{M,\lambda}^{-1} S_M = \operatorname{Tr} C_M C_{M,\lambda}^{-1},$$

since $L_M = S_M S_M^*$ and $S_M^* L_{M,\lambda}^{-1} S_M = C_M C_{M,\lambda}^{-1}$. By the the ciclicity of the trace and the definition of C_M in Def. 2, we have

$$\operatorname{Tr} C_M C_{M,\lambda}^{-1} = \int_X \operatorname{Tr}(\phi_M(x) \phi_M(x)^\top C_{M,\lambda}^{-1}) d\rho_X(x) = \int_X \|C_{M,\lambda}^{-1/2} \phi_M(x)\|^2 d\rho_X(x) = J(\lambda).$$

$\square$

Lemma 7 (Bounding $\mathcal{S}(\lambda, M, n)$). *Let $\delta \in (0, 1/3]$, $n \in \mathbb{N}$ and let $\mathcal{S}(\lambda, M, n)$ be as in Thm. 4, point 1. Let $\bar{B} = B + 2R\kappa^{2r}$, $\bar{\sigma} = \sigma + \sqrt{R}\kappa^r$. Under Asm. 3, 4 the following holds with probability at least $1 - 3\delta$*

$$\mathcal{S}(\lambda, m, n) \leq 4 \left(\frac{\bar{B}\kappa}{\sqrt{\lambda}n} + \sqrt{\frac{\bar{\sigma}^2 \mathcal{N}(\lambda)}{n}} \right) \log \frac{2}{\delta}, \tag{26}$$

when $0 < \lambda < \|L\|$ and $M \geq (4 + 18\mathcal{F}_\infty(\lambda)) \log \frac{12\kappa^2}{\lambda \delta}$.

Proof. Let $0 < \lambda < \|L\|$ and $M \geq (4 + 18\mathcal{F}_\infty(\lambda)) \log \frac{12\kappa^2}{\lambda \delta}$ (the assumption on λ, M are necessary for the application of Prop. 10). Let $Q \subseteq W$ be the event satisfying Eq. 26. The goal is to prove that the probability associated to the event Q is $\mathbb{P}(Q) \geq 1 - 3\delta$. Since the quantity $\mathcal{S}(\lambda, M, n)$ is defined in Thm. 4 as

$$\mathcal{S}(\lambda, M, n) = \|C_{M,\lambda}^{-1/2}(\widehat{S}_M^* y - S_M^* f_\rho)\| + R\kappa^{2r-1}\|C_{M,\lambda}^{-1/2}(C_M - \widehat{C}_M)\|,$$

we are first going to bound in probability the single terms on the rhs, given $\omega_1, \ldots, \omega_M$, then we take the union event, and we use this to prove that $\mathbb{P}(Q) \geq 1 - 3\delta$.

First of all we need to define four other events. Define the event $E_{\boldsymbol{\omega}}^1 \subseteq Z$, as the event satisfying

$$\|C_{M,\lambda}^{-1/2}(\widehat{S}_M^* \widehat{y} - S_M^* f_\rho)\| \leq 2 \left(\frac{B\kappa}{\sqrt{\lambda}n} + \sqrt{\frac{\sigma^2 \mathcal{N}_M(\lambda)}{n}} \right) \log \frac{2}{\delta}. \tag{27}$$

By Lemma 6 we know that $E_{\boldsymbol{\omega}}^1$ holds with probability $\mathbb{P}_Z(E_{\boldsymbol{\omega}}^1) \geq 1 - \delta$ almost everywhere for $\boldsymbol{\omega}$. Then, define the event $E_{\boldsymbol{\omega}}^2 \subset Z$, as the event satisfying

$$\|C_{M,\lambda}^{-1/2}(C_M - \widehat{C}_M)\| \leq \frac{4\kappa^2 \log \frac{2}{\delta}}{\sqrt{\lambda}n} + \sqrt{\frac{4\kappa^2 \mathcal{N}_M(\lambda) \log \frac{2}{\delta}}{n}}. \tag{28}$$

19

By applying Prop. 5, with $v_i = z_i = \phi_M(x_i)$ for $1 \le i \le n$, we have that $E_{\boldsymbol{\omega}}^2$ holds with probability $\mathbb{P}_Z(E_{\boldsymbol{\omega}}^2) \ge 1 - \delta$ almost everywhere for $\boldsymbol{\omega}$. Define $E_{\boldsymbol{\omega}} \subseteq Z$ as the event satisfying

$$\mathcal{S}(\lambda, M, n) \le 2 \left(\frac{B\kappa + 2R\kappa^{2r+1}}{\sqrt{\lambda}\, n} + (\sigma + R\kappa^{2r})\sqrt{\frac{\mathcal{N}_M(\lambda)}{n}} \right) \log \frac{2}{\delta}. \tag{29}$$

Denote with t the right hand side of the equation above and with s_1, t_1, s_2, t_2 respectively the lhs and the rhs of Eq. (27), (28). We have that $s_1 \le t_1$ and $s_2 \le t_2$ implies $\mathcal{S}(\lambda, M, n) \le t$, indeed $\mathcal{S}(\lambda, M, n) = s_1 + R\kappa^{2r-1}s_2$ and $t_1 + R\kappa^{2r-1}t_2 \le t$, since $\log(2/\delta) > 1$. In set terms $(E_{\boldsymbol{\omega}}^1 \cap E_{\boldsymbol{\omega}}^2) \subseteq E_{\boldsymbol{\omega}}$, that implies $\mathbb{P}_Z(E_{\boldsymbol{\omega}}) \ge \mathbb{P}_Z(E_{\boldsymbol{\omega}}^1 \cap E_{\boldsymbol{\omega}}^2)$, in particular

$$\mathbb{P}_Z(E_{\boldsymbol{\omega}}) \ge \mathbb{P}_Z(E_{\boldsymbol{\omega}}^1 \cap E_{\boldsymbol{\omega}}^2) \ge \mathbb{P}_Z(E_{\boldsymbol{\omega}}^1) + \mathbb{P}_Z(E_{\boldsymbol{\omega}}^2) - 1 \ge 1 - 2\delta,$$

where we used the fact that for any probability measure P and two events A, B, we have $P(A \cap B) = P(A) + P(B) - P(A \cup B) \ge P(A) + P(B) - 1$. The last event that we need to define is $A \subseteq \Omega^M$ satisfying $\mathcal{N}_M(\lambda) \le 1.5\mathcal{N}(\lambda)$. By Prop. 10, we know that A holds with probability $\mathbb{P}_\Omega(A) \ge 1 - \delta$.

Now we characterize the probability of $Q|\boldsymbol{\omega}$ when $\boldsymbol{\omega} \in A$. Denote with t_E the rhs of Eq.29 defining $E_{\boldsymbol{\omega}}$ and t_Q the rhs of Eq. 26 defining Q (and so $Q|\boldsymbol{\omega}$). When $\boldsymbol{\omega} \in A$, we have that $\mathcal{N}_M(\lambda) \le 1.5\mathcal{N}(\lambda)$ and so $t_E \le t_Q$. Then, when $\boldsymbol{\omega} \in A$, we have that $\mathcal{S}(\lambda, M, n) \le t_E$ implies $\mathcal{S}(\lambda, M, n) \le t_Q$, that is $E_{\boldsymbol{\omega}} \subseteq Q|\boldsymbol{\omega}$, implying that $\mathbb{P}_Z(E_{\boldsymbol{\omega}}) \le \mathbb{P}_Z(Q|\boldsymbol{\omega})$. By using the expansion of $\mathbb{P}(Q)$ in Eq. 25 we have

$$\mathbb{P}(Q) = \int_A \mathbb{P}_Z(Q|\boldsymbol{\omega})d\mathbb{P}_\Omega(\boldsymbol{\omega}) + \int_{\Omega^M \setminus A} \mathbb{P}_Z(Q|\boldsymbol{\omega})d\mathbb{P}_\Omega(\boldsymbol{\omega}) \ge \int_A \mathbb{P}_Z(Q|\boldsymbol{\omega})d\mathbb{P}_\Omega(\boldsymbol{\omega})$$

$$\ge \int_A \mathbb{P}_Z(E_\omega)d\mathbb{P}_\Omega(\boldsymbol{\omega}) \ge (1 - 2\delta)\int_A d\mathbb{P}_\Omega(\boldsymbol{\omega}) \ge (1 - 2\delta)(1 - \delta) \ge 1 - 3\delta.$$

$\square$

A.3.2 Estimates for $\mathcal{C}(\lambda, M)$

Lemma 8 (Bounding $\mathcal{C}(\lambda, m)$). *Let $\mathcal{C}(\lambda, M)$ as in Thm. 4, point 2. Let $\delta \in (0, 1/2]$ and $\lambda > 0$. Under Asm. 3, following holds with probability at least $1 - 2\delta$*

$$\mathcal{C}(\lambda, m) \le 4R\kappa^{2r-1} \left(\frac{\sqrt{\lambda \mathcal{F}_\infty(\lambda)} \log \frac{2}{\delta}}{M^r} + \sqrt{\frac{\lambda \mathcal{N}(\lambda)^{2r-1}\mathcal{F}_\infty(\lambda)^{2-2r} \log \frac{2}{\delta}}{M}} \right) t^{1-r},$$

when $M \ge (4 + 18\mathcal{F}_\infty(\lambda)) \log \frac{8\kappa^2}{\lambda\delta}$ and $t := \log \frac{11\kappa^2}{\lambda}$.

Proof. We now study $\mathcal{C}(\lambda, M)$ that is

$$\mathcal{C}(\lambda, M) = R\lambda^{1/2}\|L_\lambda^{-1/2}(L - L_M)\|^{2r-1}\|L_\lambda^{-1/2}(L - L_M)L_\lambda^{-1/2}\|^{2-2r}.$$

We are going to bound the two terms in probability, via Prop. 5 and Prop. 6. First of all, we recall that $\mathcal{F}_\infty(\lambda) := \sup_{\omega \in \Omega}\|L_\lambda^{-1/2}\psi_\omega\|_{\rho_X}^2$ and that $\mathcal{N}(\lambda) = \mathbb{E}\|L_\lambda^{-1/2}\psi_\omega\|_{\rho_X}^2$, indeed by Lemma. 1, characterizing L in terms of ψ_ω, the linearity and the ciclicity of the trace, we have,

$$\mathcal{N}(\lambda) := \mathrm{Tr}(L_\lambda^{-1}L) = \mathrm{Tr}\left(L_\lambda^{-1}\int \psi_\omega \otimes \psi_\omega d\pi(\omega) \right) = \int \mathrm{Tr}(L_\lambda^{-1}(\psi_\omega \otimes \psi_\omega))d\pi(\omega) \tag{30}$$

$$= \int \langle \psi_\omega, L_\lambda^{-1}\psi_\omega \rangle_{\rho_X} d\pi(\omega) = \int \|L_\lambda^{-1/2}\psi_\omega\|_{\rho_X}^2 d\pi(\omega), \tag{31}$$

where the last steps are due to the identity $\mathrm{Tr}(A(v \otimes v)) = \langle v, Av \rangle = \|A^{1/2}v\|^2$ valid for any vector v and any bounded self-adjoint positive operator A on a Hilbert space.

Define $A \subseteq \Omega^M$ the event satisfying

$$\|L_\lambda^{-1/2}(L - L_M)\| \le \frac{4\sqrt{\mathcal{F}_\infty(\lambda)\kappa^2} \log \frac{2}{\delta}}{M} + \sqrt{\frac{4\kappa^2 \mathcal{N}(\lambda) \log \frac{2}{\delta}}{M}}.$$

By the fact that $\|\cdot\| \le \|\cdot\|_{HS}$ and by Prop. 5, with $v_i = L_\lambda^{-1/2}\psi_{\omega_i}$ and $z_i = \psi_{\omega_i}$ for $i \in \{1, \ldots, M\}$ and $Q = T = L$, $T_n = L_M$, we know that the event A has probability $\mathbb{P}_\Omega(A) \ge 1 - \delta$.

Define $B \subseteq \Omega^M$ the event satisfying

$$\|L_\lambda^{-1/2}(L - L_M)L_\lambda^{-1/2}\| \le \frac{2\eta(1 + \mathcal{F}_\infty(\lambda))}{3M} + \sqrt{\frac{2\eta\mathcal{F}_\infty(\lambda)}{M}}, \tag{32}$$

with $\eta := \log\frac{8\kappa^2}{\lambda\delta}$. By Prop. 6, with $Q = L$ and $v_i = \psi_{\omega_i}$ for $i \in \{1, \ldots, M\}$, we know that B has probability $\mathbb{P}_\Omega(B) \ge 1 - \delta$.

Now set $E = A \cap B$. When E holds and under the assumption that $M \ge (4 + 18\mathcal{F}_\infty(\lambda))\log\frac{8\kappa^2}{\lambda\delta}$, we have that the right hand side of Eq. (32) is smaller than $\sqrt{\frac{4\eta\mathcal{F}_\infty(\lambda)}{M}}$, and $\eta = \log\frac{2}{\delta} + \log\frac{4\kappa^2}{\lambda}$, so

$$\mathcal{C}(\lambda, M) \le R\lambda^{1/2}\left(\frac{4\sqrt{\mathcal{F}_\infty(\lambda)\kappa^2}\log\frac{2}{\delta}}{M} + \sqrt{\frac{4\kappa^2\mathcal{N}(\lambda)\log\frac{2}{\delta}}{M}}\right)^{2r-1}\sqrt{\frac{4\eta\mathcal{F}_\infty(\lambda)}{M}}^{\,2-2r} \tag{33}$$

$$\le 4R\kappa^{2r-1}\left(\frac{\sqrt{\lambda\mathcal{F}_\infty(\lambda)}(\log\frac{2}{\delta})^r}{M^r} + \sqrt{\frac{\lambda\mathcal{N}(\lambda)^{2r-1}\mathcal{F}_\infty(\lambda)^{2-2r}\log\frac{2}{\delta}}{M}}\right)\left(1 + \frac{\log\frac{4\kappa^2}{\lambda}}{\log\frac{2}{\delta}}\right)^{1-r}. \tag{34}$$

The event E holds with probability $\mathbb{P}_\Omega(E) \ge \mathbb{P}_\Omega(A) + \mathbb{P}_\Omega(B) - 1 \ge 1 - 2\delta$. We recall that since E does not depend on z, the probability of E in W is given by the canonical extension $E' = Z \times E$ whose probability is again $\mathbb{P}(E') = \mathbb{P}_\Omega(E) \ge 1 - 2\delta$. Finally, we further upper bound in Eq.(33) the term $(\log\frac{2}{\delta})^r$ with $\log\frac{2}{\delta}$ since $\log\frac{2}{\delta} > 1$, $r \in [1/2, 1]$, and $1 + (\log\frac{4\kappa^2}{\lambda})/(\log\frac{2}{\delta})$ with $\log\frac{11\kappa^2}{\lambda}$, for the same reasons and the fact that $1 + \log 4 = \log 4e \le \log 11$. $\qquad\square$

A.3.3 Estimates for β

Lemma 9 (Bounding the norm of C_M). *Let $\delta \in (0, 1]$. Under Asm. 3, the following holds with probability at least $1 - \delta$,*

$$\|C_M\| \ge \frac{3}{4}\|L\|,$$

when $M \ge 32\left(\frac{\kappa^2}{\|L\|} + \kappa^2\right)\log\frac{2}{\delta}$.

Proof. Define the event $A \in \Omega^M$ as the one satisfying

$$\|L - L_M\|_{HS} \le \frac{4\kappa^2}{M}\log\frac{2}{\delta} - \sqrt{\frac{2\kappa^2}{M}\log\frac{2}{\delta}}.$$

Note that when $\omega \in A$ and $M \ge 32\left(\frac{\kappa^2}{\|L\|} + \kappa^2\right)\log\frac{2}{\delta}$, we have $\|L - L_M\|_{HS} \le \frac{1}{4}\|L\|$ and, by using the characterization of C_M, L_M in Rem. 8 and the fact that $\|\cdot\|_{HS} \ge \|\cdot\|$, we have

$$\|C_M\| = \|S_M^*S_M\| = \|S_MS_M^*\| = \|L_M\| \ge \|\,\|L\| - \|L - L_M\|\,\|$$

$$\ge \|L\| - \|L - L_M\| \ge \|L\| - \|L - L_M\|_{HS} \ge \|L\| - \frac{1}{4}\|L\| \ge \frac{3}{4}\|L\|.$$

Now we find a lower bound for the probability of A. Let $\zeta_i = L - \psi_{\omega_i} \otimes \psi_{\omega_i}$ be a random operator with ω_i independently and identically distributed w.r.t π and $i \in \{1, \ldots, M\}$. We have that $L - L_M = \frac{1}{M}\sum_{i=1}^M \zeta_i$ and $\mathbb{E}\zeta_i = 0$, by the characterization of L in Lemma 1. Denote with $\mathcal{L}$ the Hilbert space of Hilbert-Schmidt operators on $L^2(X, \rho_X)$. Now note that, since it is trace class, ζ_i is a random vector belonging to $\mathcal{L}$, so we can apply Prop. 2, with $T = \sup_{\omega \in \Omega}\|L - \psi_\omega \otimes \psi_\omega\|_{HS} \le 2\kappa^2$ and $S = \mathbb{E}\|\zeta_1\|_{HS}^2 \le \kappa^2$, obtaining that $\mathbb{P}_\Omega(A) \ge 1 - \delta$. $\qquad\square$

Lemma 10 (Bounding β). *Let $\delta \in (0, 1/3]$, and β be as in Thm. 4, point 3. Under Asm. 3, the following holds with probability at least $1 - 3\delta$,*

$$\beta < 2,$$

when $0 < \lambda \le \frac{3}{4}\|L\|$, and

$$n \ge 18\left(2 + \frac{\kappa}{\lambda}\right)\log\frac{4\kappa^2}{\lambda\delta}, \qquad M \ge 18\left(2 + \mathcal{F}_\infty(\lambda)\right)\log\frac{4\kappa^2}{\lambda\delta} \vee 32\left(\frac{\kappa^2}{\|L\|} + \kappa^2\right)\log\frac{2}{\delta}.$$

Proof. Let β, β_1, β_2 be defined as in Thm. 4, point 3. To bound β in probability, we first bound β_1 and β_2 in probability, and then, under the intersection of the events, we control β. First of all, denote with (a) the condition on λ, with (b) the condition on n and with $(c.1)$ the condition $M \ge 32\left(\frac{\kappa^2}{\|L\|} + \kappa^2\right)\log\frac{2}{\delta}$, while with $(c.2)$ the condition $M \ge 18\left(2 + \mathcal{F}_\infty(\lambda)\right)\log\frac{4\kappa^2}{\lambda\delta}$. Define the event $E \subseteq W$ as the one satisfying $\beta_1 \le \frac{1}{3}$. To bound the probability of E we need an auxiliary event $A \in \Omega^M$ that is the one satisfying $\frac{3}{4}\|L\| \le \|C_M\|$. The specific choice of A will be made clear later. We have

$$\mathbb{P}(E) = \int_A \mathbb{P}_Z(E|\boldsymbol{\omega})d\mathbb{P}_\Omega(\boldsymbol{\omega}) + \int_{\Omega^M \setminus A} \mathbb{P}_Z(E|\boldsymbol{\omega})d\mathbb{P}_\Omega(\boldsymbol{\omega}) \ge \int_A \mathbb{P}_Z(E|\boldsymbol{\omega})d\mathbb{P}_\Omega(\boldsymbol{\omega}).$$

By Prop. 6 and Rem. 10 point 3, we know that $\mathbb{P}_Z(E|\boldsymbol{\omega}) \ge 1 - \delta$, for any $\boldsymbol{\omega}$, when $\lambda \le \|C_M\|$ and condition (b) hold. Note that, when $\boldsymbol{\omega}$ is in A then $\frac{3}{4}\|L\| \le \|C_M\|$ and so the condition $\lambda \le \|C_M\|$ is always satisfied by assuming (a). Then under $(a), (b)$, we have $\mathbb{P}_Z(E|\boldsymbol{\omega}) \ge 1 - \delta$ when $\boldsymbol{\omega} \in A$, and so under the same conditions

$$\mathbb{P}(E) \ge \int_A \mathbb{P}_Z(E|\boldsymbol{\omega})d\mathbb{P}_\Omega(\boldsymbol{\omega}) \ge (1 - \delta)\int_A d\mathbb{P}_\Omega(\boldsymbol{\omega}) = (1 - \delta)\mathbb{P}_\Omega(A).$$

By Lemma 9, $\mathbb{P}_\Omega(A) \ge 1 - \delta$, when $(c.1)$ holds, so $\mathbb{P}(E) \ge (1 - \delta)\mathbb{P}_\Omega(A) \ge (1 - \delta)^2 \ge 1 - 2\delta$ when $(a), (b), (c.1)$ hold.

Define $D_0 \subseteq \Omega^M$ as the event satisfying $\beta_2 \le \frac{1}{3}$. By Prop. 6 and Rem. 10 point 3, we know that $\mathbb{P}_\Omega(D_0) \ge 1 - \delta$, when the condition $(c.2)$ and (a) hold. So the event $D := Z \times D_0$ has probability $\mathbb{P}(D) = \mathbb{P}_\Omega(D_0) = 1 - \delta$, when $(c.2)$ holds. Finally, note that under the conditions $(a), (b), (c.1), (c.2)$, when the event $D \cap E$ hold, we have $\beta \le (2/3)^{-3/2} < 2$. The probability of $D \cap E$ under the conditions $(a), (b), (c.1), (c.2)$ has probability

$$\mathbb{P}(D \cap E) = \mathbb{P}(D) + \mathbb{P}(E) - \mathbb{P}(D \cup E) \ge \mathbb{P}(D) + \mathbb{P}(E) - 1 \ge 1 - 3\delta.$$

$\square$

A.4   Proof of the Main Result

Here we prove Thm. 1, 2, 3 that are the main results of the paper. In particular the following Thm. 5 is a general version of the three theorem above, without the need of Assumptions 5, 2 and valid for a wide range of λ, M. In Thm. 6, we specialize the result of Thm. 5, selecting M in terms of λ such that the upper bound of the excess risk depends only on λ and is proportional to the same upper bound for kernel ridge regression that leads to optimal generalization bounds. Note that Thm. 6 is again independent of Asm. 5, 2. Finally, Thm. 7 is obtained by Thm. 6, by adding Asm. 5, 2 and has all the constant explicit. Then Thm. 1 is a specification of Thm. 7, for the simple scenario where it is only required the existence of $f_{\mathcal{H}}$ (that is Asm. 5 satisfied with $\gamma = 1$, Asm. 6 satisfied with $r = 1/2$ and Asm. 2 with $\alpha = 1$, see discussion after the introduction of the assumptions). Thm. 2 is a specification of Thm. 7 for the fast rates (Asm. 2 is satisfied with $\alpha = 1$), while Thm. 3 is a simplified version of Thm. 7 where the constants have been hidden.

Theorem 5 (Generalization Bound for RF-KRLS). *Let $\delta \in (0, 1]$. Let $\widehat{f}_{\lambda,M}$ be as in Eq. (7). Under Asm. 3, 6, 4 when $0 < \lambda \le \frac{3}{4}\|L\|$ and*

$$n \ge 18\left(2 + \frac{\kappa^2}{\lambda}\right)\log\frac{36\kappa^2}{\lambda\delta},$$

$$M \ge 18\left(q_0 + \mathcal{F}_\infty(\lambda)\right)\log\frac{108\kappa^2}{\lambda\delta},$$

with $q_0 = 2(2 + \frac{\kappa}{\|L\|} + \kappa^2)$, then the following holds with probability at least $1 - \delta$,

$$\sqrt{\mathcal{E}(\widehat{f}_{\lambda,M}) - \inf_{f \in \mathcal{H}}\mathcal{E}(f)} \le 2\left(\frac{4\bar{B}\kappa}{\sqrt{\lambda}n} + \sqrt{\frac{16\bar{\sigma}^2\mathcal{N}(\lambda)}{n}} + \mathfrak{C}(\lambda, M) + R\lambda^r\right)\log\frac{18}{\delta}. \qquad (35)$$

22

where $\bar{B} = B + 2R\kappa$, $\bar{\sigma} = \sigma + 2\sqrt{R}\kappa$,

$$\mathfrak{C}(\lambda, M) := R\kappa^{2r-1} \left(\frac{\sqrt{\lambda \mathcal{F}_\infty(\lambda)}}{M^r} + \sqrt{\frac{\lambda \mathcal{N}(\lambda)^{2r-1} \mathcal{F}_\infty(\lambda)^{2-2r}}{M}} \right) t^{1-r}. \tag{36}$$

and $t := \log \frac{11\kappa^2}{\lambda}$.

Proof. Under Asm. 3, the existence of $f_\mathcal{H}$ in Asm. 4 and Asm. 6, plus Rem. 6, we have the following analytical decomposition of the excess risk, by Thm, 4

$$|\mathcal{E}(\widehat{f}_{\lambda,M}) - \inf_{f \in \mathcal{H}} \mathcal{E}(f)|^{1/2} \le \beta \left(\mathcal{S}(\lambda, M, n) + \mathcal{C}(\lambda, M) + R\lambda^r \right), \tag{37}$$

where the quantities β, $\mathcal{C}(\lambda, M)$ and $\mathcal{S}(\lambda, M, n)$ are defined in the statement of Thm. 4. Under the same assumptions, Lemma 7, 8 and 10 are devoted to bound in probability the three quantities, with the help of the concentration inequalities recalled in Section B, plus some auxiliary results in Section D, of the appendixes.

Let $\tau := \delta/9$. Define the event $D \subseteq W$ as the one satisfying $\beta < 2$ (see Subsection A.3 for the definition of the sample space W for learning with random features, and the associated probability measure $\mathbb{P}$). By Lemma 10 we know that the event D has probability $\mathbb{P}(D) \ge 1 - 3\tau$, when the following conditions hold

$$(d_1) \ 0 \le \lambda \le \frac{3}{4}\|L\|, \qquad (d_2) \ n \ge 18 \left(2 + \kappa/\lambda\right) \log \frac{4\kappa^2}{\lambda\tau},$$

$$(d_3) \ M \ge 18 \left(2 + \mathcal{F}_\infty(\lambda)\right) \log \frac{4\kappa^2}{\lambda\tau} \ \vee \ 32 \left(\frac{\kappa^2}{\|L\|} + \kappa^2 \right) \log \frac{2}{\tau}.$$

Define the event $E \subseteq W$ as the one satisfying

$$\mathcal{S}(\lambda, M, n) \le 4 \left(\frac{\bar{B}\kappa}{\sqrt{\lambda}n} + \sqrt{\frac{\bar{\sigma}^2 \mathcal{N}(\lambda)}{n}} \right) \log \frac{2}{\tau}. \tag{38}$$

By Lemma 7 we know that the probability of E is $\mathbb{P}(E) \ge 1 - 4\tau$, when the following conditions hold

$$(e_1) \ 0 < \lambda < \|L\|, \qquad (e_2) \ M \ge (4 + 18\mathcal{F}_\infty(\lambda)) \log \frac{12\kappa^2}{\lambda\tau}.$$

Define the event $G \subseteq W$ as the one satisfying

$$\mathcal{C}(\lambda, M) \le \mathfrak{C}(\lambda, M),$$

By Lemma 8 we know that G holds with probability $\mathbb{P}(G) \ge 1 - 2\tau$, when the (d_1) and the following condition holds

$$(g_1) \ M \ge (4 + 18\mathcal{F}_\infty(\lambda)) \log \frac{8\kappa^2}{\lambda\tau}.$$

Finally Eq. (35) is obtained from Eq. (37), by bounding β with 2, $\mathcal{S}(\lambda, M, n)$ with Eq. (38) and $\mathcal{C}(\lambda, M)$ by $\mathfrak{C}(\lambda, M)$. So by definition, Eq. (35) holds under the event $D \cap E \cap G$ and the conditions $(d_1), (e_1)$ on λ, (d_2) on n and $(d_3), (e_2), (g_1)$ on M. The event $D \cap E \cap G$ has probability

$$\begin{aligned}
\mathbb{P}(D \cap E \cap G) &= \mathbb{P}(W \setminus ((W \setminus D) \cup (W \setminus E) \cup (W \setminus G))) \\
&\ge 1 - \left[(1 - \mathbb{P}(D)) + (1 - \mathbb{P}(E)) + (1 - \mathbb{P}(G)) \right] \\
&= \mathbb{P}(D) + \mathbb{P}(E) + \mathbb{P}(G) - 2 \ge 1 - 9\tau.
\end{aligned}$$

Finally note that the conditions on λ, n, M in the statement of this theorem imply, respectively, conditions $(d_1), (e_1)$ on λ, (d_2) on n, and $(d_3), (e_2), (g_1)$ on M. $\qquad\square$

Theorem 6 (Generalization Bound for RF-KRLS). *Let* $\delta \in (0, 1]$. *Let* $\widehat{f}_{\lambda,M}$ *be as in Eq. (7). Under Asm. 3, 6, 4, when* $0 < \lambda \le \frac{3}{4}\|L\|$ *and*

$$n \ge 18 \left(2 + \frac{\kappa^2}{\lambda} \right) \log \frac{36\kappa^2}{\lambda\delta},$$

$$M \ge 4\kappa^2 \left(\frac{\mathcal{N}(\lambda)}{\lambda} \right)^{2r-1} \left(\mathcal{F}_\infty(\lambda) \log \frac{11\kappa^2}{\lambda} \right)^{2-2r} \ \vee \ 18 \left(q_0 + \mathcal{F}_\infty(\lambda) \right) \log \frac{108\kappa^2}{\lambda\delta},$$

with $q_0 = 2(2 + \frac{\kappa}{\|L\|} + \kappa^2)$, then the following holds with probability at least $1 - \delta$,

$$\sqrt{\mathcal{E}(\widehat{f}_{\lambda,M}) - \inf_{f \in \mathcal{H}} \mathcal{E}(f)} \leq 8 \left(\frac{\bar{B}\kappa}{\sqrt{\lambda}n} + \sqrt{\frac{\bar{\sigma}^2 \mathcal{N}(\lambda)}{n}} + R\lambda^r \right) \log \frac{18}{\delta}. \tag{39}$$

Here $\bar{B} = B + 2R\kappa$, $\bar{\sigma} = \sigma + 2\sqrt{R}\kappa$.

Proof. First we apply Thm. 5, then we add a condition on M with respect to λ such that we can bound $\mathfrak{C}(\lambda, M)$ with $R\lambda^r$. The condition we will consider is the following

$$(g_2) \quad M \geq 4\kappa^2 \lambda^{1-2v} \mathcal{N}(\lambda)^{2v-1} \mathcal{F}_\infty(\lambda)^{2-2v} t^{2-2r}.$$

Indeed lower bounding with (g_2) the occurrences of M in $\mathfrak{C}(\lambda, M)$, we have

$$\mathfrak{C}(\lambda, M) \leq R\kappa^{2r-1} \left(\sqrt{\frac{\lambda^{1+4r^2-2r} \mathcal{F}_\infty(\lambda)^{1+4r^2-4r}}{4^{2r}\kappa^{4r-2}\mathcal{N}(\lambda)^{4r^2-2r}t^{6r-4r^2-2}}} + \sqrt{\frac{\lambda^{2r}}{4\kappa^2}} \right)$$

$$\leq R \left(\sqrt{\frac{\lambda^{2r}}{4^{2r}\kappa^{12r-8r^2-4}\mathcal{N}(\lambda)^{4r^2-2r}t^{6r-4r^2-2}}} + \sqrt{\frac{\lambda^{2r}}{4\kappa^{4-4r}}} \right) \leq R\lambda^r,$$

where the second step is due to $\mathcal{F}_\infty(\lambda) \leq \kappa^2/\lambda$ and $1 + 4r^2 - 4r \geq 0$ for $r \geq 1/2$, while the last step is due to the following three facts. First, that $4^{2r}\mathcal{N}(\lambda)^{4r^2-2r} \geq 4$, since $4r^2 - 2r \geq 0$ on $r \in [1/2, 1]$ and, by denoting with $(\lambda_i(L))_{i \geq 1}$ the eigenvalues of L, with $\|L\| := \lambda_1(L) \geq \lambda_2(L) \geq \cdots \geq 0$, and recalling that $0 \leq \lambda \leq \frac{3}{4}\|L\|$, we have

$$\mathcal{N}(\lambda) := \mathrm{Tr}(LL_\lambda^{-1}) = \sum_{i \geq 1} \frac{\lambda_i(L)}{\lambda_i(L) + \lambda} \geq \frac{\lambda_1(L)}{\lambda_1(L) + \lambda} := \frac{\|L\|}{\|L\| + \lambda} > 1/2.$$

Second, that $t^{6r-4r^2-2} \geq 1$, since $6r - 4r^2 - 2 \geq 0$ on $r \in [1/2, 1]$ and $t \geq 1$, since $0 \leq \lambda \leq \frac{3}{4}\|L\| \leq \frac{3}{4}\kappa^2$. Third, that $\kappa^{12r-8r^2-4} \geq 1$ and $\kappa^{4-4r} \geq 1$, since $12r - 8r^2 - 4 \geq 0$ and $4 - 4r \geq 0$ on $r \in [1/2, 1]$, $\kappa \geq 1$. $\square$

The following theorem is a specialization of the previous one, under 5, 2 and an explicit relation of λ with respect to n.

Theorem 7. *Let $\delta \in (0, 1]$. Under Asm. 3 and 5, 6, 2, 4, let $p := (2r + \gamma - 1)^{-1}$, and*

$$n \geq (2/\|L\|)^{\frac{p+1}{p}} \vee \left(264\kappa^2 p \log(556\kappa^2\delta^{-1}\sqrt{p\kappa^2}) \right)^{1+p}$$

$$\lambda_n = n^{-\frac{1}{2r+\gamma}},$$

$$M_n \geq c_0 \, n^{\frac{\alpha+(2r-1)(1+\gamma-\alpha)}{2r+\gamma}} \log \frac{108\kappa^2}{\lambda\delta},$$

with $c_0 = 9(3 + 4\kappa^2 + \frac{4\kappa^2}{\|L\|} + \frac{\kappa^2}{4}Q^{2r-1}F^{2-2r})$, then the following holds with probability at least $1 - \delta$,

$$\mathcal{E}(\widehat{f}_{\lambda_n, M_n}) - \inf_{f \in \mathcal{H}} \mathcal{E}(f) \leq c_1 \log^2 \frac{18}{\delta} \, n^{-\frac{2r}{2r+\gamma}}, \tag{40}$$

and $c_1 = 64(\bar{B}\kappa + \bar{\sigma}Q + R)^2$.

Proof. Let $\lambda = n^{-\frac{1}{2r+\gamma}}$ in Thm. 6 and substitute $\mathcal{F}_\infty(\lambda)$ and $\mathcal{N}(\lambda)$ by their bounds given in Asm. 5, 2. Note that to guarantee that n satisfies the associated constraint with respect to λ, in Thm. 6, and that λ is in $(0, \frac{3}{4}\|L\|]$ we need that $n \geq \left(\frac{4}{3\|L\|}\right)^{\frac{p+1}{p}}$ and

$$n \geq \left(264\kappa^2 p \log \frac{556\kappa^2\sqrt{p\kappa^2}}{\delta} \right)^{1+p},$$

with $p = \frac{1}{2r+\gamma-1}$. $\square$

24

Note that the theorems in Sect 3 are corollaries of the theorem above. For the sake of readability, in contrast to Thm. 7, the results in Thm. 1, 2, 3 are expressed with respect to $\tau := \log \frac{1}{\delta}$. Moreover in the statement of Thm. 1, 2, 3 the constants and the logarithmic terms are omitted. They can be recovered by plugging the coefficients detailed in the following proofs in the statement of Thm. 7.

Proof of Theorem 1. This is an application of Thm. 7, with minimum number of assumptions. Indeed the existence of $f_{\mathcal{H}}$ and the fact that $|y| \leq b$ a. s. satisfies Asm. 4 with $\sigma = B = 2b$. The fact that X is a Polish space and that ψ is bounded continuous satisfy Asm. 3, and so the kernel is bounded by κ^2. Since the kernel is bounded, we have that Asm. 5 is always satisfied with $\gamma = 1, Q = \kappa$; Asm. 6 is always satisfied with $r = 1/2, R = 1 \vee \|f_{\mathcal{H}}\|_{\mathcal{H}}$; Asm. 2 is always satisfied with $\alpha = 1, F = \kappa^2$.

In particular we have the following constants $n_0 := 4\|L\|^{-2} \vee \left(264\kappa^2 \log \frac{556\kappa^3}{\delta} \right)^2$,

$$c_0 := 9 \left(3 + 4\kappa^2 + \frac{4\kappa^2}{\|L\|} + \kappa^4/4 \right), \qquad c_1 := 8(\bar{B}\kappa + \bar{\sigma}\kappa + 1 \vee \|f_{\mathcal{H}}\|_{\mathcal{H}}),$$

with $\bar{B} := 2b + 2\kappa(1 \vee \|f_{\mathcal{H}}\|_{\mathcal{H}})$ and $\bar{\sigma} := 2b + 2\kappa\sqrt{1 \vee \|f_{\mathcal{H}}\|_{\mathcal{H}}}$. $\qquad\square$

Corollary 1. *Under the same assumptions of Thm. 1, if $n \geq \|L\|^{-2} \vee \left(1056 \log \frac{1056\sqrt{278\kappa^5 b}}{\sqrt{c_1}} \right)^2$ and $\lambda_n = n^{-1/2}$, then a number of random features M_n equal to*

$$M_n = 2c_0 \sqrt{n} \log (c_2 n),$$

is enough to guarantee that

$$\mathbb{E} \; \mathcal{E}(\widehat{f}_{\lambda_n, M_n}) - \mathcal{E}(f_{\mathcal{H}}) \leq \frac{40c_1}{\sqrt{n}}.$$

In particular the constants c_0, c_1 are as in Thm. 1 and $c_2 = \frac{8\kappa^2\sqrt{b}}{\sqrt{c_1}}$.

Proof. In the rest we will denote $\mathcal{E}(\widehat{f}_{\lambda_n, M_n}) - \mathcal{E}(f_{\mathcal{H}})$, with $\mathcal{R}(\widehat{f}_{\lambda_n, M_n})$ and will use the notation of Sect. A.3. Fix $\delta_0 = \frac{2c_1}{\kappa^2 b} n^{-1}$. Denote with E, the event satisfying $\mathcal{R}(\widehat{f}_{\lambda_n, M_n}) > t_0$, with $t_0 := c_1 \log^2 \frac{18}{\delta_0} n^{-1/2}$.

First, note that $M_n \geq 2c_0\sqrt{n}\log \left(\frac{8\kappa^2\sqrt{b}}{\sqrt{c_1}} n \right)$, satisfies $M_n \geq c_0 \sqrt{n}\log \frac{108\kappa^2\sqrt{n}}{\delta_0}$ and any $n \geq \|L\|^{-2} \vee \left(1056 \log \frac{1056\sqrt{278\kappa^5 b}}{\sqrt{c_1}} \right)^2$ satisfies $n \geq n_0(\delta_0)$ with $n_0(\delta)$ as in Thm. 1. So, we can apply Thm. 1, from which we know that E holds with probability smaller than δ_0.

Second, by Rem. 6 and Rem. 9, we have that

$$\mathcal{R}(\widehat{f}_{\lambda_n, M_n}) = \|S_M \widehat{C}_{M,\lambda}^{-1} \widehat{S}_M^* \widehat{y} - P f_\rho\|_{\rho_X} \tag{41}$$

$$\leq \|S_M\|\|\widehat{C}_{M,\lambda}^{-1}\|\|\widehat{S}_M^*\|\|\widehat{y}\|_{\mathbb{R}^n} + \|P\|\|f_\rho\|_{\rho_X} \tag{42}$$

$$\leq \frac{\kappa^2 b}{\lambda} + b \leq \frac{2\kappa^2 b}{\lambda} =: R_0, \tag{43}$$

where we used the fact that $\|S_M\|, \|\widehat{S}_M^*\| \leq \kappa$ (see Def. 2), that $\|\widehat{C}_{M,\lambda}^{-1}\| \leq \lambda^{-1}$, that $\|\widehat{y}\|^2 = \frac{1}{n}\sum_{y_i}^2$, that $f_\rho(x) = \mathbb{E}[y|x]$, that the y's are bounded in $[-b, b]$, and the fact that $\kappa^2/\lambda \geq 1$, by definition of κ, λ.

Now, by denoting with $\mathbf{1}_E$ the indicator function for E, we have

$$\mathbb{E} \; \mathcal{R}(\widehat{f}_{\lambda_n, M_n}) = \mathbb{E} \; \mathbf{1}_E \mathcal{R}(\widehat{f}_{\lambda_n, M_n}) \; + \; \mathbb{E} \; \mathbf{1}_{Z \setminus E} \mathcal{R}(\widehat{f}_{\lambda_n, M_n}).$$

In particular

$$\mathbb{E} \; \mathbf{1}_E \mathcal{R}(\widehat{f}_{\lambda_n, M_n}) \leq R_0 \; \mathbb{E} \; \mathbf{1}_E = R_0 \mathbb{P}(E) \leq R_0 \delta_0.$$

For the second term we have

$$\mathbb{E} \; \mathbf{1}_{Z \setminus E} \mathcal{R}(\widehat{f}_{\lambda_n, M_n}) \quad = \quad \mathbb{E} \; \mathbf{1}_{\{\mathcal{R}(\widehat{f}_{\lambda_n, M_n}) \leq t_0\}} \mathcal{R}(\widehat{f}_{\lambda_n, M_n}) \quad = \quad \int_0^{t_0} \mathbb{P}(\mathcal{R}(\widehat{f}_{\lambda_n, M_n}) > t)dt.$$

By changing variable, in the integral above, via $t = \frac{c_1}{\sqrt{n}} \log^2 \frac{18}{\delta}$, and using the fact that $\mathbb{P}(\mathcal{R}(\widehat{f}_{\lambda_n, M_n}) > \frac{c_1}{\sqrt{n}} \log^2 \frac{18}{\delta}) \leq \delta$, we have

$$\int_0^{t_0} \mathbb{P}(\mathcal{R}(\widehat{f}_{\lambda_n, M_n}) > t)dt = \frac{2c_1}{\sqrt{n}} \int_{\delta_0}^{18} \frac{\log \frac{18}{\delta}}{\delta} \mathbb{P}\left(\mathcal{R}(\widehat{f}_{\lambda_n, M_n}) > \frac{c_1}{\sqrt{n}} \log^2 \frac{18}{\delta} \right) d\delta$$

$$\leq \frac{2c_1}{\sqrt{n}} \int_{\delta_0}^{18} \log \frac{18}{\delta} d\delta$$

$$= \frac{2c_1}{\sqrt{n}} \left(18 - \delta_0 \left(1 + \log \frac{18}{\delta_0} \right) \right) \quad \leq \quad \frac{36c_1}{\sqrt{n}}.$$

So finally we have

$$\mathbb{E}\, \mathcal{R}(\widehat{f}_{\lambda_n, M_n}) \leq R_0 \delta_0 + \frac{36c_1}{\sqrt{n}} = \frac{40c_1}{\sqrt{n}}.$$

$\square$

Proof of Theorem 2. This is an application of Thm. 7, where assumption Asm. 2 is satisfied with $F = \kappa^2$ and $\alpha = 1$. Indeed the existence of $f_{\mathcal{H}}$ and the fact that $|y| \leq b$ a. s. satisfies Asm. 4 with $\sigma = B = 2b$. The fact that X is a Polish space and that ψ is bounded continuous satisfy, Asm. 3, and so the kernel is bounded by κ^2. Since the kernel is bounded, we have that Asm. 2 is always satisfied with $\alpha = 1, F = \kappa^2$. Asm. 4 and Asm. 6 are directly satisfied by Asm. 1. In particular we obtain

$$n_0 := (2/\|L\|)^{\frac{p+1}{p}} \vee \left(264\kappa^2 p \, \log(556\kappa^2 \delta^{-1} \sqrt{p\kappa^2}) \right)^{1+p},$$

$$c_0 := 9 \left(3 + 4\kappa^2 + \frac{4\kappa^2}{\|L\|} + \frac{\kappa^{4-2r}}{4} Q^{2r-1} \right), \quad c_1 := 64 \left(\bar{B}\kappa + \bar{\sigma}Q + R \right)^2,$$

with $\bar{B} := 2b + 2\kappa R$ and $\bar{\sigma} := 2b + 2\kappa\sqrt{R}$.

$\square$

Proof of Example 2. By definition of ψ_s, π_s we have

$$\int \psi_s(x, \omega)\psi_s(x', \omega)d\pi_s(\omega) = \int \psi(x, \omega)\sqrt{C_s s(\omega)}\psi_s(x', \omega)\sqrt{C_s s(\omega)}\frac{1}{C_s s(\omega)}d\pi(\omega)$$

$$= \int \psi(x, \omega)\psi(x', \omega)d\pi(\omega) = K(x, x').$$

Now we show that ψ_s, π_s achieves $\mathcal{F}_\infty(\lambda) = \mathcal{N}(\lambda)$. By recalling that $s(\omega) = \|(L + \lambda I)^{-1/2}\psi(\cdot, \omega)\|_{\rho_X}^{-2}$, we have

$$\mathcal{F}_\infty(\lambda) = \sup_{\omega \in \Omega}\|(L + \lambda I)^{-1/2}\psi_s(\cdot, \omega)\|_{\rho_X}^2 = C_s \sup_{\omega \in \Omega} s(\omega)\|(L + \lambda I)^{-1/2}\psi(\cdot, \omega)\|_{\rho_X}^2$$

$$= C_s \sup_{\omega \in \Omega}\|(L + \lambda I)^{-1/2}\psi(\cdot, \omega)\|_{\rho_X}^{-2}\|(L + \lambda I)^{-1/2}\psi(\cdot, \omega)\|_{\rho_X}^2 = C_s.$$

We recall that $C_s = \int \frac{1}{s(\omega)}d\pi$. Denoting with ψ_ω the function $\psi(\cdot, \omega)$ and considering that $\|Ax\|_{\rho_X} = \text{Tr}(A^2(x \otimes x))$ for any bounded symmetrix linear operator A and vector v, and that the trace is linear,

$$\mathcal{F}_\infty(\lambda) = C_s = \int \|(L + \lambda I)^{-1/2}\psi_\omega\|_{\rho_X}^2 d\pi(\omega) = \int \text{Tr}((L + \lambda I)^{-1}(\psi_\omega \otimes \psi_\omega))d\pi(\omega)$$

$$= \text{Tr}\left((L + \lambda I)^{-1} \int (\psi_\omega \otimes \psi_\omega)d\pi(\omega) \right) = \text{Tr}((L + \lambda I)^{-1}L) = \mathcal{N}(\lambda).$$

where the fact that $L = \int \psi_\omega \otimes \psi_\omega d\pi(\omega)$ is due to Lemma 1.

$\square$

Proof of Theorem 3. This is a version of Thm. 7, with simplified set of assumptions. Indeed the existence of $f_{\mathcal{H}}$ and the fact that $|y| \leq b$ a. s. satisfies Asm. 4 with $\sigma = B = 2b$. The fact that X is a Polish space and that ψ is bounded continuous, satisfy Asm. 3, and so the kernel

is bounded by κ^2. Asm. 4 and Asm. 6 are directly satisfied by Asm. 1.In particular we obtain

$$n_0 := (2/\|L\|)^{\frac{p+1}{p}} \ \vee \ \left(264\kappa^2 p \ \log(556\kappa^2\delta^{-1}\sqrt{p\kappa^2})\right)^{1+p},$$

$$c_0 := 9\left(3 + 4\kappa^2 + \frac{4\kappa^2}{\|L\|} + \frac{\kappa^2}{4}Q^{2r-1}F^{2-2r}\right), \quad c_1 := 64\left(\bar{B}\kappa + \bar{\sigma}Q + R\right)^2,$$

with $\bar{B} := 2b + 2\kappa R$ and $\bar{\sigma} := 2b + 2\kappa\sqrt{R}$. $\qquad\square$

B Concentration Inequalities

Here we recall some standard concentration inequalities that will be used in Sect. A.3. The following inequality is from Thm.3 of [31] and will be used in Lemma 10, together with other inequalities, to concentrate the empirical effective dimension to the true effective dimension.

Proposition 1 (Bernstein's inequality for sum of random variables)**.** *Let $x_1, \ldots, x_n$ be a sequence of independent and identically distributed random variables on $\mathbb{R}$ with zero mean. If there exists an $T, S \in \mathbb{R}$ such that $x_i \leq T$ almost everywhere and $\mathbb{E}x_i^2 \leq S$, for $i \in \{1, \ldots, n\}$. For any $\delta > 0$ the following holds with probability at least $1 - \delta$:*

$$\frac{1}{n}\sum_{i=1}^n x_i \leq \frac{2T\log\frac{1}{\delta}}{3n} + \sqrt{\frac{2S\log\frac{1}{\delta}}{n}}.$$

If there exists $T' \geq \max_i |x_i|$ almost everywhere, then the same bound, with T' instead of T, holds for the for the absolute value of the left hand side, with probability at least $1 - 2\delta$.

Proof. It is a restatement of Theorem 3 of [31]. $\qquad\square$

The following inequality is and adaptation of Thm. 3.3.4 in [32] and is a generalization of the previous one to random vectors. It is used primarily in Lemma 6, to control the sample error. Moreover it is used in Prop. 10, Lemma 10, to control the empirical effective dimension and to bound the term β of Thm. 4, in the main theorem. Finally it is used to prove the inequality in Prop. 5.

Proposition 2 (Bernstein's inequality for sum of random vectors)**.** *Let $z_1, \ldots, z_n$ be a sequence of independent identically distributed random vectors on a separable Hilbert space $\mathcal{H}$. Assume $\mu = \mathbb{E}z_i$ exists and let $\sigma, M \geq 0$ such that*

$$\mathbb{E}\|z_i - \mu\|_{\mathcal{H}}^p \leq \frac{1}{2}p!\sigma^2 M^{p-2}, \quad \forall p \geq 2,$$

for any $i \in \{1, \ldots, n\}$. Then for any $\delta \in (0, 1]$:

$$\left\|\frac{1}{n}\sum_{i=1}^n z_i - \mu\right\|_{\mathcal{H}} \leq \frac{2M\log\frac{2}{\delta}}{n} + \sqrt{\frac{2\sigma^2\log\frac{2}{\delta}}{n}}$$

with probability at least $1 - \delta$.

Proof. restatement of Theorem 3.3.4 of [32]. $\qquad\square$

The following inequality is essentially Thm. 7.3.1 in [33] (generalized to separable Hilbert spaces by the technique in Section 4 of [34]). It is a generalization of the Bernstein inequality to random operators. It is mainly used to prove the inequality in Prop. 6.

Proposition 3 (Bernstein's inequality for sum of random operators)**.** *Let $\mathcal{H}$ be a separable Hilbert space and let $X_1, \ldots, X_n$ be a sequence of independent and identically distributed self-adjoint positive random operators on $\mathcal{H}$. Assume that there exists $\mathbb{E}X_i = 0$ and $\lambda_{\max}(X_i) \leq T$ almost surely for some $T > 0$, for any $i \in \{1, \ldots, n\}$. Let S be a positive operator such that $\mathbb{E}(X_i)^2 \leq S$. Then for any $\delta \in (0, 1]$ the following holds*

$$\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) \leq \frac{2T\beta}{3n} + \sqrt{\frac{2\|S\|\beta}{n}}$$

with probability at least $1 - \delta$. *Here* $\beta = \log \frac{2 \operatorname{Tr} S}{\|S\|\delta}$.

If there exists L' *such that* $L' \geq \max_i \|X_i\|$ *almost everywhere, then the same bound holds with* L' *instead of* L *for the operator norm, with probability at least* $1 - 2\delta$.

Proof. The theorem is a restatement of Theorem 7.3.1 of [33] generalized to the separable Hilbert space case by means of the technique in Section 4 of [34]. □

C   Operator Inequalities

Let $\mathcal{H}, \mathcal{K}$ be separable Hilbert spaces and $A, B : \mathcal{H} \to \mathcal{H}$ bounded linear operators.

The following inequality is needed to prove the interpolation inequality in Prop. 9, that is needed to perform a fine split of the computational error.

Proposition 4 (Cordes Inequality [35]). *If* A, B *are self-adjoint and positive, then*

$$\|A^s B^s\| \leq \|AB\|^s \quad \text{when } 0 \leq s \leq 1$$

D   Auxiliary Results

The next proposition is used in Lemma 7 to control the sample error. It is based on the Bernstein inequality for random vectors, Prop. 2.

Proposition 5. *Let* $\mathcal{H}, \mathcal{K}$ *be two separable Hilbert spaces and* $(v_1, z_1), \ldots, (v_n, z_n) \in \mathcal{H} \times \mathcal{K}$, *with* $n \geq 1$, *be independent and identically distributed random pairs of vectors, such that there exists a constant* $\kappa > 0$ *for which* $\|v\|_{\mathcal{H}} \leq \kappa$ *and* $\|z\|_{\mathcal{H}} \leq \kappa$ *almost everywhere. Let* $Q = \mathbb{E}\, v \otimes v$, *let* $T = \mathbb{E}\, v \otimes z$ *and* $T_n = \frac{1}{n} \sum_{i=1}^n v_i \otimes z_i$. *For any* $0 < \lambda \leq \|Q\|$ *and any* $\tau \geq 0$, *the following holds*

$$\|(Q + \lambda I)^{-1/2}(T - T_n)\|_{HS} \leq \frac{4\sqrt{\tilde{\mathcal{F}}_\infty(\lambda)}\kappa \log \frac{2}{\tau}}{n} + \sqrt{\frac{4\kappa^2 \tilde{\mathcal{N}}(\lambda) \log \frac{2}{\tau}}{n}}$$

with probability at least $1 - \tau$, *where* $\operatorname{ess\,sup}$ *denotes the essential supremum and*

$$\tilde{\mathcal{F}}_\infty(\lambda) := \operatorname*{ess\,sup}_{v \in \mathcal{H}} \|(Q + \lambda I)^{-1/2}v\|^2, \quad \tilde{\mathcal{N}}(\lambda) := \operatorname{Tr}((Q + \lambda I)^{-1}Q).$$

In particular, we recall that $\tilde{\mathcal{N}}(\lambda) \leq \tilde{\mathcal{F}}_\infty(\lambda) \leq \frac{\kappa^2}{\lambda}$.

Proof. Define for any $i \in \{1, \ldots, n\}$ the random operator $\zeta_i = (Q + \lambda I)^{-1/2}v_i \otimes z_i$. Note that $\mathbb{E}\zeta_i = (Q + \lambda I)^{-1/2}T$. Since ζ_i is a vector in the Hilbert space of Hilbert-Schmidt operators on $\mathcal{H}$, we study the moments of $\|\zeta_i - \mathbb{E}\zeta_i\|_{HS}$ in order to apply Prop. 2. We recall that

$$\operatorname{ess\,sup}\|\zeta_i - \mathbb{E}\zeta_i\|_{HS} \leq \operatorname{ess\,sup}\|\zeta_i\|_{HS} + \mathbb{E}\|\zeta_i\|_{HS} \leq 2\operatorname{ess\,sup}\|\zeta_i\|_{HS}$$

$$= 2\operatorname{ess\,sup}\|(Q + \lambda I)^{-1/2}v_i \otimes z_i\|_{HS} \leq \operatorname{ess\,sup}\|(Q + \lambda I)^{-1/2}v_i\|_{\mathcal{H}}\|z_i\|_{\mathcal{K}}$$

$$\leq 2\operatorname{ess\,sup}\|(Q + \lambda I)^{-1/2}v_i\|_{\mathcal{H}}\ \operatorname{ess\,sup}\|z_i\|_{\mathcal{K}} = 2\tilde{F}_\infty(\lambda)^{1/2}\kappa.$$

For any $p \geq 2$ we have

$$\mathbb{E}\,\|\zeta_i - \mathbb{E}\zeta_i\|_{HS}^p \leq (\operatorname*{ess\,sup}_z\|\zeta_i - \mathbb{E}\zeta_i\|^{p-2})(\mathbb{E}\,\|\zeta_i - \mathbb{E}\zeta_i\|_{HS}^2)$$

$$\leq (2\tilde{F}_\infty(\lambda)^{1/2}\kappa)^{p-2}\mathbb{E}\|\zeta_i - \mathbb{E}\zeta_i\|_{HS}^2.$$

Now we study $\mathbb{E}\|\zeta_i - \mathbb{E}\zeta_i\|_{HS}^2$,

$$\mathbb{E}\|\zeta_i - \mathbb{E}\zeta_i\|_{HS}^2 = \operatorname{Tr}(\mathbb{E}\zeta_i \otimes \zeta_i - (\mathbb{E}\zeta_i)^2) \leq \operatorname{Tr}(\mathbb{E}\zeta_i \otimes \zeta_i)$$

$$= \mathbb{E}\ \|z_i\|^2 \operatorname{Tr}((Q + \lambda I)^{-1/2}(v_i \otimes v_i)(Q + \lambda I)^{-1/2})$$

$$\leq \operatorname{ess\,sup}\|z_i\|_{\mathcal{K}}^2\ \mathbb{E}\operatorname{Tr}((Q + \lambda I)^{-1/2}(v_i \otimes v_i)(Q + \lambda I)^{-1/2})$$

$$\leq \operatorname{ess\,sup}\|z_i\|_{\mathcal{K}}^2\ \operatorname{Tr}((Q + \lambda I)^{-1/2}\mathbb{E}(v_i \otimes v_i)(Q + \lambda I)^{-1/2})$$

$$\leq \kappa^2 \operatorname{Tr}((Q + \lambda I)^{-1}Q) = \kappa^2 \tilde{\mathcal{N}}(\lambda),$$

for any $1 \leq i \leq n$. Therefore for any $p \geq 2$ we have

$$\mathbb{E}\left\|\zeta_i - \mathbb{E}\zeta_i\right\|_{HS}^p \leq \frac{1}{2}p!\sqrt{2\kappa^2\mathcal{N}(\lambda)}^2(2\tilde{F}_\infty(\lambda)^{1/2}\kappa)^{p-2}.$$

Finally we apply Prop. 2. $\qquad\square$

The following inequality, together with Prop. 8, is used in Prop. 10, Lemmas 10, 8. A similar technique can be found in [36].

Proposition 6. *Let $v_1, \ldots, v_n$ with $n \geq 1$, be independent and identically distributed random vectors on a separable Hilbert spaces $\mathcal{H}$ such that $Q = \mathbb{E}\, v \otimes v$ is trace class, and for any $\lambda > 0$ there exists a constant $\mathcal{F}_\infty(\lambda) < \infty$ such that $\langle v, (Q + \lambda I)^{-1}v \rangle \leq \mathcal{F}_\infty(\lambda)$ almost everywhere. Let $Q_n = \frac{1}{n}\sum_{i=1}^n v_i \otimes v_i$ and take $0 < \lambda \leq \|Q\|$. Then for any $\delta \geq 0$, the following holds with probability at least $1 - 2\delta$*

$$\|(Q + \lambda I)^{-1/2}(Q - Q_n)(Q + \lambda I)^{-1/2}\| \leq \frac{2\beta(1 + \mathcal{F}_\infty(\lambda))}{3n} + \sqrt{\frac{2\beta\mathcal{F}_\infty(\lambda)}{n}},$$

where $\beta = \log \frac{4\,\mathrm{Tr}\,Q}{\lambda\delta}$. Moreover, with the same probability

$$\lambda_{\max}\left((Q + \lambda I)^{-1/2}(Q - Q_n)(Q + \lambda I)^{-1/2}\right) \leq \frac{2\beta}{3n} + \sqrt{\frac{2\beta\mathcal{F}_\infty(\lambda)}{n}}.$$

Proof. Let $Q_\lambda = Q + \lambda I$. Here we apply Prop. 3 on the random variables $Z_i = M - Q_\lambda^{-1/2}v_i \otimes Q_\lambda^{-1/2}v_i$ with $M = Q_\lambda^{-1/2}QQ_\lambda^{-1/2}$ for $1 \leq i \leq n$. Note that the expectation of Z_i is 0. The random vectors are bounded by

$$\|Q_\lambda^{-1/2}QQ_\lambda^{-1/2} - Q_\lambda^{-1/2}v_i \otimes Q_\lambda^{-1/2}v_i\| \leq \langle v_i, Q_\lambda^{-1}v_i \rangle + \|Q_\lambda^{-1/2}QQ_\lambda^{-1/2}\| \leq \mathcal{F}_\infty(\lambda) + 1,$$

almost everywhere, for any $1 \leq i \leq n$. The second order moment is

$$\begin{aligned}\mathbb{E}(Z_i)^2 = \mathbb{E}\ \langle v_i, Q_\lambda^{-1}v_i \rangle\ Q_\lambda^{-1/2}v_i \otimes Q_\lambda^{-1/2}v_i\ &-\ Q_\lambda^{-2}Q^2 \\ &\leq \mathcal{F}_\infty(\lambda)\mathbb{E}Q_\lambda^{-1/2}v_i \otimes Q_\lambda^{-1/2}v_i = \mathcal{F}_\infty(\lambda)Q = S,\end{aligned}$$

for $1 \leq i \leq n$. Now we can apply Prop. 3. Now some considerations on β. It is $\beta = \log \frac{2\,\mathrm{Tr}\,S}{\|S\|\delta} = \frac{2\,\mathrm{Tr}\,Q_\lambda^{-1}Q}{\|Q_\lambda^{-1}Q\|\delta}$, now $\mathrm{Tr}\,Q_\lambda^{-1}Q \leq \frac{1}{\lambda}\mathrm{Tr}\,Q$. We need a lower bound for $\|Q_\lambda^{-1}Q\| = \frac{\sigma_1}{\sigma_1 + \lambda}$ where $\sigma_1 = \|Q\|$ is the biggest eigenvalue of Q, now $\lambda \leq \sigma_1$ thus $\frac{\sigma_1}{\sigma_1 + \lambda} \geq 1/2$ and so $\beta \leq \log \frac{2\,\mathrm{Tr}\,Q_\lambda^{-1}Q}{\|Q_\lambda^{-1}Q\|\delta} \leq \log \frac{4\,\mathrm{Tr}\,Q}{\lambda\delta}$.

For the second bound of this proposition we use the second bound of Prop. 3, the analysis remains the same except for uniform bound on Z_1, that now is

$$\sup_{f \in \mathcal{H}}\langle f, Z_1 f \rangle = \sup_{f \in \mathcal{H}}\langle f, Q_\lambda^{-1}Qf \rangle - \left\langle f, Q_\lambda^{-1/2}v_i \right\rangle^2 \leq \sup_{f \in \mathcal{H}}\langle f, Q_\lambda^{-1}Qf \rangle \leq 1.$$

$\qquad\square$

In the following remark, we start from the result of the previous proposition, expressing the conditions on n and λ with respect to a given value for the bound.

Remark 10. *With the same notation of Prop. 6, assume that $\|v\| \leq \kappa$ almost everywhere[3], then we have that*

1. *for any $t \in (0, 1]$, when $n \geq \frac{2}{t^2}\left(\frac{2t}{3} + \mathcal{F}_\infty(\lambda)\right)\log\frac{4\kappa^2}{\lambda\delta}$ and $\lambda \leq \|Q\|$, we have*

$$\lambda_{\max}\left((Q + \lambda I)^{-1/2}(Q - Q_n)(Q + \lambda I)^{-1/2}\right) \leq t,$$

   with probability at least $1 - \delta$.

---

[3]Or equivalently define κ^2 with respect to $\mathcal{F}_\infty(\lambda)$ as $\kappa^2 = \inf_{\lambda>0}\mathcal{F}_\infty(\lambda)(\|Q\| + \lambda)$.

2. *The equation above holds with the same probability with $t = 1/2$, when $\frac{9\kappa^2}{n}\log\frac{n}{2\delta} \leq \lambda \leq \|Q\|$ and $n \geq 405\kappa^2 \vee 67\kappa^2\log\frac{\kappa^2}{2\delta}$.*

3. *The equation above holds with the same probability with $t = 1/3$, when $\frac{19\kappa^2}{n}\log\frac{n}{4\delta} \leq \lambda \leq \|Q\|$ and $n \geq 405\kappa^2 \vee 67\kappa^2\log\frac{\kappa^2}{2\delta}$.*

The next proposition, together with Prop. 10 are a restatement of Prop. 1 of [16]. In particular the next proposition performs the analytic decomposition of the difference between the empirical and the true effective dimension, while Prop. 10, bounds the decomposition in probability.

Proposition 7 (Geometry of Empirical Effective Dimension). *Let $\mathcal{L}$ be an Hilbert space. Let $L, L_M : \mathcal{L} \to \mathcal{L}$ be two bounded positive linear operators, that are trace class. Given $\lambda > 0$, let*

$$\mathcal{N}(\lambda) = \mathrm{Tr}(LL_\lambda^{-1}) \quad and \quad \tilde{\mathcal{N}}(\lambda) = \mathrm{Tr}(L_M L_{M,\lambda}^{-1}),$$

with $L_\lambda = L + \lambda I$. Then

$$|\tilde{\mathcal{N}}(\lambda) - \mathcal{N}(\lambda)| \leq \lambda e(\lambda) + \frac{d(\lambda)^2}{1 - c(\lambda)}$$

where $c(\lambda) = \lambda_{\max}(\tilde{B})$, $d(\lambda) = \|\tilde{B}\|_{\mathrm{HS}}$ with $\tilde{B} = L_\lambda^{-1/2}(L - L_M)L_\lambda^{-1/2}$ and $e(\lambda) = \mathrm{Tr}(L_\lambda^{-1/2}\tilde{B}L_\lambda^{-1/2})$.

Proof. First of all note that $\lambda_{\max}(\tilde{B})$, the biggest eigenvalue of $\lambda_{\max}$, is smaller than 1 since $\tilde{B}$ is the difference of two positive operators and the biggest eigenvalue of the minuend operator is $\|L_\lambda^{-1/2}LL_\lambda^{-1/2}\| = \frac{\lambda_{\max}(L)}{\lambda_{\max}(L)+\lambda} < 1$. Then we can use the fact that $A(A+\lambda I)^{-1} = I - \lambda(A+\lambda I)^{-1}$ for any bounded linear operator A and that $L_{M,\lambda}^{-1} = L_\lambda^{-1/2}(I - \tilde{B})^{-1}L_\lambda^{-1/2}$ (see the proof of Prop. 8), since $\lambda_{\max}(\tilde{B}) < 1$, to obtain

$$|\tilde{\mathcal{N}}(\lambda) - \mathcal{N}(\lambda)| = |\mathrm{Tr}(L_{M,\lambda}^{-1}L_M - LL_\lambda^{-1})| = \lambda|\mathrm{Tr}(L_{M,\lambda}^{-1} - L_\lambda^{-1})| = |\lambda\mathrm{Tr}(L_{M,\lambda}^{-1}(L_M - L)L_\lambda^{-1})|$$

$$= |\lambda\mathrm{Tr}(L_\lambda^{-1/2}(I - \tilde{B})^{-1}L_\lambda^{-1/2}(L_M - L)L_\lambda^{-1/2}L_\lambda^{-1/2})|$$

$$= |\lambda\mathrm{Tr}(L_\lambda^{-1/2}(I - \tilde{B})^{-1}\tilde{B}L_\lambda^{-1/2})|.$$

Considering that for any bounded symmetric linear operator X the following identity holds

$$(I - X)^{-1}X = X + X(I - X)^{-1}X,$$

when $\lambda_{\max}(X) < 1$, we have

$$\lambda|\mathrm{Tr}(L_\lambda^{-1/2}(I - \tilde{B})^{-1}\tilde{B}L_\lambda^{-1/2})| \leq \underbrace{\lambda|\mathrm{Tr}(L_\lambda^{-1/2}\tilde{B}L_\lambda^{-1/2})|}_{A} + \underbrace{\lambda|\mathrm{Tr}(L_\lambda^{-1/2}\tilde{B}(I - \tilde{B})^{-1}\tilde{B}L_\lambda^{-1/2})|}_{B}.$$

The term A is just equal to $\lambda e(\lambda)$. Now, by definition of Hilbert-Schmidt norm, the term B can be written as $B = \|\lambda^{1/2}L_\lambda^{-1/2}\tilde{B}(I - \tilde{B})^{-1/2}\|_{\mathrm{HS}}^2$, thus we have

$$B = \|\lambda^{1/2}L_\lambda^{-1/2}\tilde{B}(I - \tilde{B})^{-1/2}\|_{\mathrm{HS}}^2 \leq \|\lambda^{1/2}L_\lambda^{-1/2}\|^2\|\tilde{B}\|_{\mathrm{HS}}^2\|(I - \tilde{B})^{-1/2}\|^2 \leq (1-c(\lambda))^{-1}d(\lambda)^2,$$

since $\|(I - \tilde{B})^{-1/2}\|^2 = (1 - \lambda_{\max}(\tilde{B}))^{-1}$ because the spectral function $(1 - \sigma)^{-1}$ is increasing and positive on $[-\infty, 1)$. $\square$

The next result is essentially Prop. 7 of [16], while a similar technique can be found in [36]. It is used, mainly together with Prop. 6, to give multiplicative bounds to empirical operators. It is used in the analytic decomposition of the excess risk, in Prop. 7, Thm. 4.

Proposition 8. *Let $\mathcal{H}$ be a separable Hilbert space, let A, B two bounded self-adjoint positive linear operators on $\mathcal{H}$ and $\lambda > 0$. Then*

$$\|(A + \lambda I)^{-1/2}B^{1/2}\| \leq \|(A + \lambda I)^{-1/2}(B + \lambda I)^{1/2}\| \leq (1 - \beta)^{-1/2}$$

with

$$\beta = \lambda_{\max}\left[(B + \lambda I)^{-1/2}(B - A)(B + \lambda I)^{-1/2}\right].$$

Note that $\beta \leq \frac{\lambda_{\max}(B)}{\lambda_{\max}(B)+\lambda} < 1$ by definition.

30

Proof. Let $B_\lambda = B + \lambda I$. First of all note that $\beta < 1$ for any $\lambda > 0$. Indeed, by exploiting the variational formulation of the biggest eigenvalue, we have

$$\beta = \lambda_{\max}(B_\lambda^{-1/2}(B - A)B_\lambda^{-1/2}) = \sup_{f \in \mathcal{H}, \|f\|_\mathcal{H} \leq 1} \left\langle f, B_\lambda^{-1/2}(B - A)B_\lambda^{-1/2}f \right\rangle$$

$$= \sup_{f \in \mathcal{H}, \|f\|_\mathcal{H} \leq 1} \left\langle f, B_\lambda^{-1/2}BB_\lambda^{-1/2}f \right\rangle - \left\langle f, B_\lambda^{-1/2}AB_\lambda^{-1/2}f \right\rangle$$

$$\leq \sup_{f \in \mathcal{H}, \|f\|_\mathcal{H} \leq 1} \left\langle f, B_\lambda^{-1/2}BB_\lambda^{-1/2}f \right\rangle = \lambda_{\max}(B_\lambda^{-1/2}BB_\lambda^{-1/2})$$

$$\leq \frac{\lambda_{\max}(B)}{\lambda_{\max}(B) + \lambda} < 1,$$

since A is a positive operator and thus $\left\langle f, B_\lambda^{-1/2}AB_\lambda^{-1/2}f \right\rangle \geq 0$ for any $f \in \mathcal{H}$. Now note that

$$(A + \lambda I)^{-1} = [(B + \lambda I) - (B - A)]^{-1}$$

$$= \left[B_\lambda^{1/2} \left(I - B_\lambda^{-1/2}(B - A)B_\lambda^{-1/2} \right) B_\lambda^{1/2} \right]^{-1}$$

$$= B_\lambda^{-1/2} \left[I - B_\lambda^{-1/2}(B - A)B_\lambda^{-1/2} \right]^{-1} B_\lambda^{-1/2}.$$

Now let $X = (I - B_\lambda^{-1/2}(B - A)B_\lambda^{-1/2})^{-1}$. We have that,

$$\|(A + \lambda I)^{-1/2}B_\lambda^{1/2}\| = \|B_\lambda^{1/2}(A + \lambda I)^{-1}B_\lambda^{1/2}\|^{1/2} = \|X\|^{1/2}$$

because $\|Z\| = \|Z^*Z\|^{1/2}$ for any bounded operator Z. Note that

$$\|(A + \lambda I)^{-1/2}B^{1/2}\| = \|(A + \lambda I)^{-1/2}B_\lambda^{1/2}B_\lambda^{-1/2}B^{1/2}\| \leq \|(A + \lambda I)^{-1/2}B_\lambda^{1/2}\| \|B_\lambda^{-1/2}B^{1/2}\|$$

$$\leq \|X\|^{1/2}\|B_\lambda^{-1/2}B^{1/2}\| \leq \|X\|^{1/2}.$$

Finally let $Y = B_\lambda^{-1/2}(B - A)B_\lambda^{-1/2}$, we have seen that $\beta = \lambda_{\max}(Y) < 1$, then

$$\|X\| = \|(I - Y)^{-1}\| = (1 - \lambda_{\max}(Y))^{-1},$$

since $X = w(Y)$ with $w(\sigma) = (1 - \sigma)^{-1}$ for $-\infty \leq \sigma < 1$, and w is positive and monotonically increasing on the domain. □

The following proposition is used to give a fine analytical decomposition of the excess risk in Prop. 7, Thm. 4. A similar interpolation inequality for finite dimensional matrices, can be found in [37]. Here we prove it for bounded linear operators on separable Hilbert spaces.

Proposition 9. *Let* $\mathcal{H}, \mathcal{K}$ *be two separable Hilbert spaces and* X, A *be bounded linear operators, with* $X : \mathcal{H} \to \mathcal{K}$ *and* $A : \mathcal{H} \to \mathcal{H}$ *be positive semidefinite.*

$$\|XA^\sigma\| \leq \|X\|^{1-\sigma}\|XA\|^\sigma, \qquad \forall \sigma \in [0, 1]. \tag{44}$$

Proof.

$$\|XA^\sigma\| = \|A^\sigma(X^*X)^{\frac{1}{\sigma}\sigma}A^\sigma\|^{\frac{1}{2}} \leq \|A(X^*X)^{\frac{1}{\sigma}}A\|^{\frac{\sigma}{2}} \tag{45}$$

where the last inequality is due to Cordes (see Proposition 4). Then we have that $(X^*X)^{\frac{1}{\sigma}} \preccurlyeq \|X\|^{\frac{2(1-\sigma)}{\sigma}}X^*X$ (where $\preccurlyeq$ is the Löwner partial ordering on positive operators) and so

$$A(X^*X)^{\frac{1}{\sigma}}A \preccurlyeq \|X\|^{\frac{2(1-\sigma)}{\sigma}}AX^*XA$$

that implies

$$\|A(X^*X)^{\frac{1}{\sigma}}A\|^{\frac{\sigma}{2}} \leq \|X\|^{1-\sigma}\|AX^*XA\|^{\frac{\sigma}{2}} = \|X\|^{1-\sigma}\|XA\|^\sigma \tag{46}$$

□

In the next proposition, we bound $\mathcal{N}_M(\lambda)$ in terms of the effective dimension $\mathcal{N}(\lambda)$ that is the one associated to the kernel K. The proof of Prop. 10 analogous to the one of Prop. 1 of [16], with simpler proof and slightly improved constants.

31

Proposition 10 (Bounds on the Effective Dimension). *Let $\tilde{\mathcal{N}}(\lambda) = \operatorname{Tr} L_M L_{M,\lambda}^{-1}$. Under Assumption 3, for any $\delta > 0$, $\lambda \leq \|L\|$ and $m \geq (4 + 18\mathcal{F}_\infty(\lambda))\log\frac{12\kappa^2}{\lambda\delta}$, then the following holds with probability at least $1 - \delta$,*

$$|\tilde{\mathcal{N}}(\lambda) - \mathcal{N}(\lambda)| \leq c(\lambda, m, \delta)\mathcal{N}(\lambda) \leq 1.55\mathcal{N}(\lambda),$$

with $c(\lambda, m, \delta) = \frac{q}{\mathcal{N}(\lambda)} + \sqrt{\frac{3q}{2\mathcal{N}(\lambda)}} + \frac{3}{2}\left(\frac{3q}{\sqrt{\mathcal{N}(\lambda)}} + \sqrt{3q}\right)^2$ and $q = \frac{4\mathcal{F}_\infty(\lambda)\log\frac{6}{\delta}}{3m}$.

Proof. First of all we recall that $\tilde{\mathcal{N}}(\lambda) = \operatorname{Tr} L_M L_{M,\lambda}^{-1}$ and that $\mathcal{N}(\lambda) = \operatorname{Tr} L L_\lambda^{-1}$. Let $\tau = \delta/3$. By Prop. 7 we know that

$$|\tilde{\mathcal{N}}(\lambda) - \mathcal{N}(\lambda)| \leq \left(\frac{d(\lambda)^2}{(1 - c(\lambda))\mathcal{N}(\lambda)} + \frac{\lambda e(\lambda)}{\mathcal{N}(\lambda)}\right)\mathcal{N}(\lambda)$$

where $c(\lambda) = \lambda_{\max}(\tilde{B})$, $d(\lambda) = \|\tilde{B}\|_{HS}$ with $\tilde{B} = L_\lambda^{-1/2}(L - L_M)L_\lambda^{-1/2}$ and $e(\lambda) = |\operatorname{Tr}(L_\lambda^{-1/2}\tilde{B}L_\lambda^{-1/2})|$. Thus, now we control $c(\lambda)$, $d(\lambda)$ and $e(\lambda)$ in probability. Choosing m such that $m \geq (4 + 18\mathcal{F}_\infty(\lambda))\log\frac{4\kappa^2}{\lambda\tau}$, Prop. 6 guarantees that $c(\lambda) = \lambda_{\max}(\tilde{B}) \leq 1/3$ with probability at least $1 - \tau$.

To find an upper bound for $\lambda e(\lambda)$ we define the i.i.d. random variables $\eta_i = \langle \psi_{\omega_i}, \lambda L_\lambda^{-2}\psi_{\omega_i}\rangle_{\rho_X} \in \mathbb{R}$ with $i \in \{1, \ldots, m\}$. By linearity of the trace and the expectation, we have $M = \mathbb{E}\eta_1 = \mathbb{E}\langle\psi_{\omega_1}, \lambda L_\lambda^{-2}\psi_{\omega_1}\rangle_{\rho_X} = \mathbb{E}\operatorname{Tr}(\lambda L_\lambda^{-2}\psi_{\omega_1} \otimes \psi_{\omega_1}) = \lambda\operatorname{Tr}(L_\lambda^{-2}L)$. Therefore,

$$\lambda e(\lambda, m) = \left|\lambda\operatorname{Tr}(L_\lambda^{-1/2}\tilde{B}L_\lambda^{-1/2})\right| = \left|\lambda\operatorname{Tr}\left(L_\lambda^{-1}LL_\lambda^{-1} - \frac{1}{m}\sum_{i=1}^m (L_\lambda^{-1}\psi_{\omega_i}) \otimes (L_\lambda^{-1}\psi_{\omega_i})\right)\right|$$

$$= \left|\operatorname{Tr}\left(\lambda L_\lambda^{-1}LL_\lambda^{-1}\right) - \frac{1}{m}\sum_{i=1}^m \langle\psi_{\omega_i}, \lambda L_\lambda^{-2}\psi_{\omega_i}\rangle_{\rho_X}\right| = \left|M - \frac{1}{m}\sum_{i=1}^m \eta_i\right|.$$

By noting that M is upper bounded by $M = \operatorname{Tr}(\lambda L_\lambda^{-2}L) = \operatorname{Tr}((I - L_\lambda^{-1}L)L_\lambda^{-1}L) \leq \mathcal{N}(\lambda)$, we can apply the Bernstein inequality (Prop. 1) where T and S are

$$\sup_{\omega \in \Omega}|M - \eta_1| \leq \lambda\|L_\lambda^{-1/2}\|^2\|L_\lambda^{-1/2}\psi_{\omega_1}\|^2 + M \leq \mathcal{F}_\infty(\lambda) + \mathcal{N}(\lambda) \leq 2\mathcal{F}_\infty(\lambda) = T,$$

$$\mathbb{E}(\eta_1 - M)^2 = \mathbb{E}\eta_1^2 - M^2 \leq \mathbb{E}\eta_1^2 \leq (\sup_{\omega \in \Omega}|\eta_1|)(\mathbb{E}\eta_1) \leq \mathcal{F}_\infty(\lambda)\mathcal{N}(\lambda) = S.$$

Thus, we have

$$\lambda|\operatorname{Tr} L_\lambda^{-1/2}\tilde{B}L_\lambda^{-1/2}| \leq \frac{4\mathcal{F}_\infty(\lambda)\log\frac{2}{\tau}}{3m} + \sqrt{\frac{2\mathcal{F}_\infty(\lambda)\mathcal{N}(\lambda)\log\frac{2}{\tau}}{m}},$$

with probability at least $1 - \tau$.

To find a bound for $d(\lambda)$ consider that $\tilde{B} = W - \frac{1}{m}\sum_{i=1}^m \zeta_i$ where ζ_i are i.i.d. random operators defined as $\zeta_i = L_\lambda^{-1/2}(\psi_{\omega_i} \otimes \psi_{\omega_i})L_\lambda^{-1/2} \in \mathcal{L}$ for all $i \in \{1, \ldots, m\}$, and $W = \mathbb{E}\zeta_1 = L_\lambda^{-1}L \in \mathcal{L}$. Then, by noting that $\|W\|_{HS} \leq \mathbb{E}\operatorname{Tr}(\zeta_1) = \mathcal{N}(\lambda)$, we can apply the Bernstein's inequality for random vectors on a Hilbert space (Prop. 2) where T and S are:

$$\|W - \zeta_1\|_{HS} \leq \|L_\lambda^{-1/2}\psi_{\omega_1}\|_{\rho_X}^2 + \|W\|_{HS} \leq \mathcal{F}_\infty(\lambda) + \|W\|_{HS} \leq 2\mathcal{F}_\infty(\lambda) = T,$$

$$\mathbb{E}\|\zeta_1 - W\|^2 = \mathbb{E}\operatorname{Tr}(\zeta_1^2 - W^2) \leq \mathbb{E}\operatorname{Tr}(\zeta_1^2) \leq (\sup_{\omega \in \Omega}\|\zeta_1\|)(\mathbb{E}\operatorname{Tr}(\zeta_1)) = \mathcal{F}_\infty(\lambda)\mathcal{N}(\lambda) = S,$$

obtaining

$$d(\lambda) = \|\tilde{B}\|_{HS} \leq \frac{4\mathcal{F}_\infty(\lambda)\log\frac{2}{\tau}}{m} + \sqrt{\frac{4\mathcal{F}_\infty(\lambda)\mathcal{N}(\lambda)\log\frac{2}{\tau}}{m}},$$

with probability at least $1 - \tau$. Then, by taking a union bound for the three events we have

$$|\tilde{\mathcal{N}}(\lambda) - \mathcal{N}(\lambda)| \leq \left(\frac{q}{\mathcal{N}(\lambda)} + \sqrt{\frac{3q}{2\mathcal{N}(\lambda)}} + \frac{3}{2} \left(\frac{3q}{\sqrt{\mathcal{N}(\lambda)}} + \sqrt{3q} \right)^2 \right) \mathcal{N}(\lambda),$$

with probability at least $1 - \delta$, where $q = \frac{4\mathcal{F}_\infty(\lambda) \log \frac{6}{\delta}}{3m}$. Finally, if $m \geq (4 + 18\mathcal{F}_\infty(\lambda)) \log \frac{12\kappa^2}{\lambda\delta}$, then we have $q \leq 2/27$. Noting that $\mathcal{N}(\lambda) \geq \|LL_\lambda^{-1}\| = \frac{\|L\|}{\|L\|+\lambda} \geq 1/2$, we have that

$$|\tilde{\mathcal{N}}(\lambda) - \mathcal{N}(\lambda)| \leq 1.55\mathcal{N}(\lambda).$$

$\square$

E   Examples of Random feature maps

A lot of works have been devoted to develop random feature maps in the the setting introduced above, or slight variations (see for example [13, 38, 20, 25, 39, 40, 41, 42, 43, 44] and references therein). In the rest of the section, we give several examples.

Random Features for Translation Invariant Kernels and extensions [13, 38, 41, 44]   This approximation method is defined in [13] for the translation invariant kernels when $X = \mathbb{R}^d$. A kernel $k : X \times X \to \mathbb{R}$ is *translation invariant*, when there exists a function $v : X \to \mathbb{R}$ such that $k(x, z) = v(x - z)$ for all $x, z \in X$. Now, let $\hat{v} : \Omega \to \mathbb{R}$ be the Fourier transform of v, with $\Omega = X \times [0, 2\pi]$. As shown in [13], by using the Fourier transform, we can express k as

$$k(x, z) = v(x - z) = \int_\Omega \psi(\omega, x)\psi(\omega, z)\hat{v}(\omega)d\pi, \quad \forall x, z \in X,$$

with π proportional to $\hat{v}$, $\psi(\omega, x) = \cos(w^\top x + b)$ and $\omega := (w, b) \in \Omega$, $x \in X$. Note that X, Ω, π, ψ satisfy Assumption 3.
In [38], they further randomize the construction above, by using results from locally sensitive hashing, to obtain a feature map which is a binary code. It can be shown that their methods satisfy Assumption 3 for an appropriate choice of Ω and the probability distribution π.
[41, 44] consider the setting of [13], but [41] selects $\omega_1, \ldots, \omega_m$ by means of the fast Welsh-Hadamard transform in order to improve the computational complexity for the algorithm computing K_M, while [44] selects them by using low-discrepancy sequences for quasi-random sampling to improve the statistical accuracy of K_M with respect to K.

Random Features Maps for the Gaussian Kernel, which are functions in $\mathcal{H}$.   This set of random features is related to the deterministic polynomial features in [45]. Let $X = [0, 1]^d$ and $\Omega = \mathbb{N}^d$. Let $\sigma > 0$. We have

$$\psi(\omega, x) = C^{1/2} e^{-\frac{\sigma^2}{2}\|x\|^2} \prod_{j=1}^d x_j^{\omega_j}, \qquad \pi(\omega) = \frac{\sigma^{-2\sum_j \omega_j}}{C\,\omega_1! \ldots \omega_d!},$$

where C is the normalization factor $C = e^{\sigma^2 d}$. Indeed by Taylor expansion of the exponential function and of the power of a multinomial, we have

$$e^{-\frac{\sigma^2}{2}\|x-x'\|^2} = e^{-\frac{\sigma^2}{2}(\|x\|^2+\|x'\|^2)}e^{\sigma^2 x^\top x} = e^{-\frac{\sigma^2}{2}(\|x\|^2+\|x'\|^2)} \sum_{t\geq 0} \frac{\sigma^{2t}}{t!}(x^\top x')^t$$

$$= e^{-\frac{\sigma^2}{2}(\|x\|^2+\|x'\|^2)} \sum_{t\geq 0} \frac{\sigma^{2t}}{t!}(\sum_{j=1}^d x_j x'_j)^t$$

$$= e^{-\frac{\sigma^2}{2}(\|x\|^2+\|x'\|^2)} \sum_{t\geq 0} \frac{\sigma^{2t}}{t!} \sum_{\omega_1+\cdots+\omega_d=t} \frac{t!}{\omega_1!\ldots\omega_d!} \prod_{j=1}^d (x_j x'_j)^{\omega_j}$$

$$= \sum_{t\geq 0} \frac{\sigma^{2t}}{Ct!} \sum_{\omega_1+\cdots+\omega_d=t} \frac{t!}{\omega_1!\ldots\omega_d!} \left(C^{1/2}e^{-\frac{\sigma^2}{2}\|x\|^2}\prod_{j=1}^d x_j^{\omega_j}\right) \left(C^{1/2}e^{-\frac{\sigma^2}{2}\|x'\|^2}\prod_{j=1}^d x_j'^{\omega_j}\right)$$

$$= \sum_{\omega\in\mathbb{N}^d} \psi(x,\omega)\psi(x',\omega)\pi(\omega).$$

Note that it is possible to sample from π in the following way. By the steps above it is clear that a sample from π can be obtained by first sampling t from a Poisson distribution with parameter $\sigma^2 d$ and then sampling $\omega_1,\ldots,\omega_d$ from a multinomial distribution with probability $p_1 = \cdots = p_d = 1/d$ and number of trials t.

Finally note that $\psi(\omega, \cdot)$ is in $\mathcal{H}$, the RKHS induced by the Gaussian kernel (to prove it apply Prop. 3.6 of [46] with $b_\nu = \delta_{\nu=(\omega_1,\ldots,\omega_d)}$ and note that e_ν is exactly a multiple of $\psi(\omega, \cdot)$). Additionally note that $\sup_{\omega,x} |\psi(\omega, x)| < \infty$ and moreover $\|\psi(\omega, \cdot)\|_\mathcal{H} < \infty$ for any $\omega \in \mathbb{N}^d$. However $\lim_{\|\omega\|\to\infty}\|\psi(\omega, \cdot)\|_\mathcal{H} = \infty$.

Random Features Maps for Dot Product Kernels [39, 40, 43] This approximation method is defined for the dot product kernels when X is the ball of $\mathbb{R}^d$ of radius R, with $R > 0$. The considered kernels are of the form $k(x, z) = v(x^\top z)$ for a $v : \mathbb{R} \to \mathbb{R}$ such that

$$v(t) = \sum_{p=0}^\infty c_p t^p, \quad \text{with} \quad c_p \geq 0 \ \forall p \geq 0.$$

[39], start from the consideration that when $w \in \{-1, 1\}^d$ is a vector of d independent random variables with probability at least $1/2$, then $\mathbb{E}\, ww^\top = I$. Thus

$$\mathbb{E}_w\, (x^\top w)(z^\top w) = \mathbb{E}\, x^\top (ww^\top)z = x^\top \mathbb{E}(ww^\top)z = x^\top z,$$

and $(x^\top z)^p$ can be approximated by $g(W_p, x)^\top g(W_p, z)$ with $g(W_p, x) = \prod_{i=1}^p x^\top w_i$ and $W_p = (w_1, \ldots, w_p) \in \{-1, 1\}^{p\times d}$ a matrix of independent random variables with probability at least $1/2$. Indeed it holds,

$$\mathbb{E}_{W_p}\, g(W_p, x)g(W_p, z) = \mathbb{E}_{W_p} \prod_{i=1}^p (x^\top w_i) \prod_{i=1}^p (z^\top w_i) = \mathbb{E}_{W_p} \prod_{i=1}^p (x^\top w_i)(z^\top w_i)$$

$$= \prod_{i=1}^p \mathbb{E}_{w_i}\, (x^\top w_i)(z^\top w_i) = \prod_{i=1}^p x^\top z = (x^\top z)^p.$$

Therefore the idea is to define the following sample space $\Omega = \mathbb{N}_0 \times (\bigcup_{p=1}^\infty T^p)$ with $T^p = \{-1, 1\}^{p\times d}$, the probability π on Ω as $\pi(p, W_q) = \pi_\mathbb{N}(p)\pi(W_q|p)$ for any $p, q \in \mathbb{N}_0$, $W_q \in T^q$ with $\pi_\mathbb{N}(p) = \frac{\tau^{-p-1}}{\tau-1}$ for a $\tau > 1$ and $\pi(W_q|p) = 2^{-pd}\delta_{pq}$ and the function

$$\psi(\omega, x) = \sqrt{c_p \tau^{p+1}(\tau - 1)}g(W_p, x), \quad \forall \omega = (p, W_p) \in \Omega$$

where w_i is the i-th row of W_p with $1 \leq i \leq p$. Now note that Eq. (6) is satisfied, indeed for any $x, z \in X$

$$\int_\Omega \psi(\omega, x)\psi(\omega, z)d\pi = \int_{\mathbb{N}_0 \times (\bigcup_{p=1}^\infty T^p)} c_p g(W_p, x) g(W_p, z) dp d\pi(W_q | p)$$

$$= \sum_{p=0}^\infty c_p \, \mathbb{E}_{W_p} g(W_p, x) g(W_p, z)$$

$$= \sum_{p=0}^\infty c_p \, (x^\top z)^p = v(x^\top z) = K(x, z).$$

Now note that Assumption 3 is satisfied when $v(\tau R^2 d) < \infty$. Indeed, considering that $(x^\top w)^2 \leq \|x\|^2 \|w\|^2 \leq R^2 d$ for any $x \in X$ and $w \in \{-1, 1\}^d$, we have

$$\sup_{\omega \in \Omega, x \in X} |\psi(\omega, x)|^2 = \sup_{p \in \mathbb{N}_0, W_p \in T^p, x \in X} c_p \tau^{p+1}(\tau - 1) g_p(W_p, x) g_p(W_p, z)$$

$$= \sup_{p \in \mathbb{N}_0} c_p \tau^{p+1}(\tau - 1) \sup_{W_p \in T^p, x \in X} \prod_{i=1}^p (x^\top w_i)^2$$

$$\leq \sup_{p \in \mathbb{N}_0} c_p \tau^{p+1}(\tau - 1)(R^2 d)^p \leq \frac{\tau}{\tau - 1} \sum_{p=0}^\infty c_p \tau^p (R^2 d)^p$$

$$= \frac{\tau}{\tau - 1} v(\tau R^2 d).$$

[40, 43] approximate the construction above by using randomized tensor sketching and Johnson-Lindenstrauss random projections. It can be shown that even their methods satisfy Assumption 3 for an appropriate choice of Ω and the probability distribution π.

Random Laplace Feature Maps for Semigroup invariant Kernels [42]   The considered input space is $X = [0, \infty)^d$ and the considered kernels are of the form

$$K(x, z) = v(x + z), \quad \forall x, z \in X,$$

and $v : X \to \mathbb{R}$ is a function that is positive semidefinite. By Berg's theorem, it is equivalent to the fact that $\breve{v}$, the Laplace transform of v is such that $\breve{v}(\omega) \geq 0$, for all $\omega \in X$ and that $\int_X \breve{v}(\omega) d\omega = V < \infty$. It means that we can express K by Eq. (6), where $\Omega = X$, the feature map is $\psi(\omega, x) = \sqrt{V} e^{-\omega^\top x}$, for all $\omega \in \Omega, x \in X$ and the probability density is $\pi(\omega) = \frac{\breve{v}(\omega)}{V}$. Note that Assumption 3 is satisfied.

Homogeneous additive Kernels [47]   In this work they focus on $X = [0, 1]^d$ and on additive homogeneous kernels, that are of the form

$$K(x, z) = \sum_{i=1}^d k(x_i, z_i), \quad \forall x, z \in (\mathbb{R}^+)^d,$$

where k is an γ-homogeneous kernel, that is a kernel k on $\mathbb{R}^+$ such that

$$k(cs, ct) = c^\gamma k(s, t), \, \forall c, s, t \in \mathbb{R}^+.$$

As pointed out in [47], this family of kernels is particularly useful on histograms. Exploiting the homogeneous property, the kernel k is rewritten as follows

$$k(s, t) = (st)^{\frac{\gamma}{2}} v(\log s - \log t), \, \forall s, t \in \mathbb{R}^+ \quad \text{with} \quad v(r) = k(e^{r/2}, e^{-r/2}), \, \forall r \in \mathbb{R}.$$

Let $\hat{v}$ be the Fourier transform of v. In [47], Lemma 1, they prove that v is a positive definite function, that is equivalent, by the Bochner theorem, to the fact that $\hat{v}(\omega) \geq 0$, for $\omega \in \mathbb{R}$, and $\int \hat{v}(\omega) d\omega \leq V < \infty$. Therefore k satisfies Eq. 6 with $\Omega = \mathbb{R} \times [0, 2\pi]$, a feature map $\psi_0((w, b), s) = \sqrt{V} s^{\frac{\gamma}{2}} \cos(b + w\omega \log s)$, with $(w, b) \in \Omega$ and a probability density π_0 defined as $\pi_0((w, b)) = \frac{\hat{v}(w)}{V} U(b)$ for all $\omega \in \Omega$, where U is the uniform distribution on $[0, 2\pi]$. Now note that K is expressed by Eq. 6, with the feature map $\psi(\omega, x) = (\psi_0(\omega, x_1), \ldots, \psi_0(\omega, x_d))$ and probability density $\pi(\omega) = \prod_{i=1}^d \pi_0(\omega_i)$. In contrast with the previous examples, [47] suggest to select $\omega_1, \ldots, \omega_m$ by using a deterministic approach, in order to achieve a better accuracy, with respect to the random sampling with respect to π.

Infinite one-layer Neural Nets and Group Invariant Kernels [25]   In this work a generalization of the ReLU activation function for one layer neural networks is considered, that is

$$\psi(\omega, x) = (\omega^\top x)^p \mathbf{1}(\omega^\top x), \quad \forall \omega \in \Omega, x \in X$$

for a given $p \geq 0$, where $\mathbf{1}(\cdot)$ is the step function and $\Omega = X = \mathbb{R}^d$. In the paper is studied the kernel K, given by Eq. 6 when the distribution π is a zero mean, unit variance Gaussian. The kernel has the following form

$$K(x, z) := \frac{1}{\pi} \|x\|^p \|z\|^p J_p(\theta(x, z)), \quad \forall x, z \in X.$$

where J_p is defined in Eq. 4 of [25] and $\theta(x, z)$ is the angle between the two vectors x and z. Examples of J_p are the following

$$
\begin{aligned}
J_0(\theta) &= \pi - \theta \\
J_1(\theta) &= \sin\theta + (\pi - \theta)\cos\theta \\
J_2(\theta) &= 3\sin\theta\cos\theta + (\pi - \theta)(1 + 2\cos^2\theta).
\end{aligned}
$$

Note that when X is a bounded subset of $\mathbb{R}^d$ then Assumption 3 is satisfied. Moreover in [25] it is shown that an infinite one-layer neural network with the ReLU activation function is equivalent to a kernel machine with kernel K. The units of a finite one-layer NN are obviously a subset of the infinite one-layer NN. While in the context of Deep Learning the subset is chosen by optimization, in the paper it is proposed to find it by randomization, in particular by sampling the distribution π.