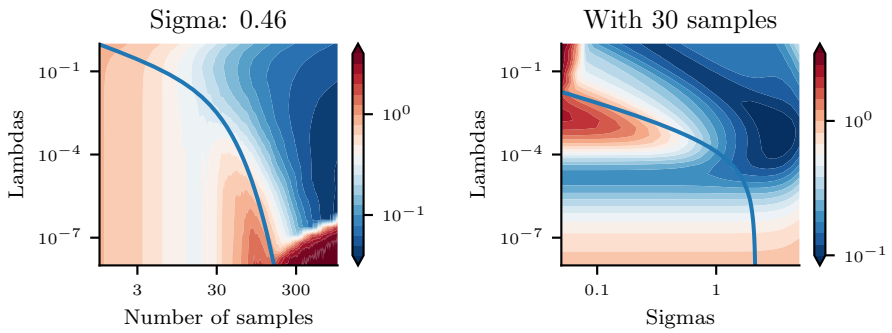

Figure 4: Two other cuts of Figure 1.

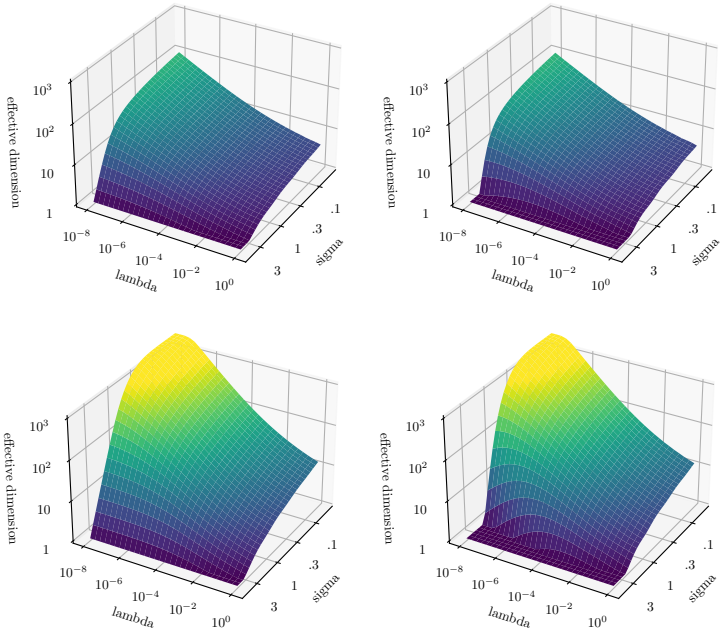

Figure 5: Effective dimensions $\mathcal{N}_1$ (left) and $\mathcal{N}_2$ (right) as a function of (λ, σ) in one dimension (top) and two dimension (bottom) when $\rho_{\mathcal{X}}$ is uniform on $\mathcal{X} = [-1, 1]^d$, and k is the Gaussian kernel.

A Generic proofs and discussions

A1 What do we mean by transitory regimes?

In essence, by transitory regimes we mean any finite-time behavior that does not match an expected long-time horizon "stationary" behavior. More precisely, let $\Gamma = \{(n, \mathbb{E}_{\mathcal{D}_n}[\mathcal{E}(f_n)]) \,|\, n \in \mathbb{N}\}$ be the graph of the expected excess risk. Theorem 2 provides a lower-upper bound of the form $\Gamma \subset \{(n, cn^{-\gamma}(1 + ah(n))) \,|\, n \in \mathbb{N}, a \in [-1, 1]\}$ with c, γ two constants and h a function that goes to zero when its argument goes to infinity. This shows that, as n grows large, $\mathbb{E}_{\mathcal{D}_n}[\mathcal{E}(f_n)]$ will behave as $cn^{-\gamma}$. However, this stationary behavior in $cn^{-\gamma}$ might take time to kick in, and when only accessing a small number of samples n, our bound does not lead to strong constraints on $\mathbb{E}_{\mathcal{D}_n}[\mathcal{E}(f_n)]$, which might arguably exhibit a very different profile. We illustrate this idea on Figure 6.

A2 Regularization is a change of kernel

Kernel ridge regression (6) with a kernel k_1 and a regularization parameter λ_1 is equivalent to kernel ridge regression with a kernel $k_2 = \lambda_1^{-1} k_1$ and a regularization parameter $\lambda_2 = 1$. Indeed, In the definition of the regularized risk (6), the regularization parameter λ and the kernel k only appear in

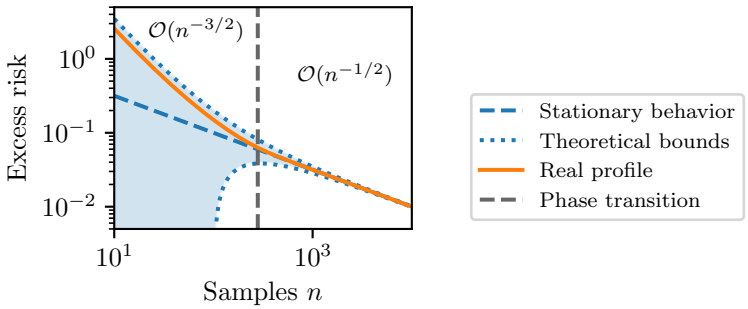

Figure 6: Illustration of transitory regimes. In essence, Theorem 2 states that $\mathcal{E}_n := \mathbb{E}_{\mathcal{D}_n}[\mathcal{E}(f_{n,\lambda})] = A(n,\lambda)(1+h(n,\lambda))$ for $h = O(\mathcal{N}_\infty(\lambda)/n)$. We illustrate our upper-lower bound when $A(n,\lambda_n) = n^{-1/2}$ and $nh(n,\lambda_n)$ is known to be in $[-10^2, 10^2]$. The upper-lower bound forces $\mathcal{E}_n$ to behave in $n^{-1/2}$ when n goes to infinity, yet when n is small, it can showcase quite different "transitory" behaviors.

one term $\lambda \|f\|_{\mathcal{F}}$. This term can be written as

$$\lambda \left\langle f, K^{-1} f \right\rangle_{L^2(\rho)} = \left\langle f, (\lambda^{-1} K)^{-1} f \right\rangle_{L^2(\rho)}.$$

Because K depends linearly on k, $\lambda^{-1}K$ is the integral operator linked with the kernel $\lambda^{-1}k$ when k is the kernel associated with the operator K.

The interpolation setting where $\lambda_1 = 0$ but f is searched in $\mathcal{F} = \operatorname{im} K^{1/2}$ corresponds to the barrier regularization $\chi_{\mathcal{F}}$, where $\chi_A(x) = 0$ in $x \in A$ and $+\infty$ otherwise. In other terms, the limiting case where $\lambda_1 = 0$ corresponds to kernel ridge regression with $\lambda_2 = 1$, $\|f\|_{\mathcal{F}_2} = +\infty$ if $f \notin \mathcal{F}_2$ and zero otherwise.

A3 Excess risk bounds - Proof of Theorem 2 and corollaries

For ease of notation, we will use the finite-dimensional notation $u^\top w$ also to denote the inner product $\langle u, v \rangle$ in (infinite-dimensional) Hilbert spaces. Moreover, we will simply write $\|\cdot\|$ for both $\|\cdot\|_{\mathcal{H}}$ and the operator norm on $\mathcal{H}$ (depending on context), L^2 for $L^2(\rho_{\mathcal{X}})$, and $\|\cdot\|_2$ for both $\|\cdot\|_{L^2}$ and the operator norm on L^2. While in our statements, for the sake of clarity, we have expressed everything in terms of operators on L^2, for the proofs it is more convenient to work on $\mathcal{H}$. Let us introduce the embedding

$$S : \mathcal{H} \to L^2, \qquad \theta \mapsto (x \mapsto \theta^\top \varphi(x)).$$

From S, one can take its adjoint S^* and check that $K = SS^*$. K is isometric to the (non-centered) covariance operator

$$\Sigma = S^* S = \mathbb{E}[\varphi(X) \otimes \varphi(X)].$$

Note that

$$\|S\theta\|_2^2 = \mathbb{E}[(\varphi(X)^\top \theta)^2] \leq \mathbb{E}[\|\varphi(X)\|^2 \|\theta\|^2] = \|\varphi\|_2^2 \|\theta\|^2,$$

which implies that K is a continuous operator as soon as $\varphi \in L^2$. The kernel ridge regression estimator (6) is characterized as

$$f_n = S(\Sigma_n + \lambda)^{-1} S_n^* \mathbb{Y},$$

where

$$S_n : \mathcal{H} \to \mathbb{R}^n,\ \theta \mapsto (\theta^\top \varphi(X_i))_{i\in[n]}, \qquad \Sigma_n = S_n^* S_n, \qquad \mathbb{Y} = (Y_i)_{i\in[n]} \in \mathbb{R}^n.$$

Endowing $\mathbb{R}^n$ with the scalar product $\langle a, b \rangle = \frac{1}{n}\sum_{i\in[n]} a_i b_i$, we have

$$S_n^* \mathbb{Y} = \frac{1}{n} \sum_{i\in[n]} Y_i \varphi(X_i), \qquad \Sigma_n = \frac{1}{n} \sum_{i=1}^n \varphi(X_i) \otimes \varphi(X_i).$$

It is useful to define ε_i as the difference between Y_i and $f^*(X_i) = \mathbb{E}[Y_i \,|\, X = X_i]$, which can be seen as the labeling noise and average to zero. We have, with $E = (\varepsilon_i)_{i\in[n]} \in \mathbb{R}^n$,

$$Y_i = f^*(X_i) + \varepsilon_i = \varphi(X_i)^\top \theta_* + \varepsilon_i, \qquad \mathbb{Y} = S_n \theta_* + E.$$

As a consequence,

$$f_n = S(\Sigma_n + \lambda)^{-1}\Sigma_n\theta_* + S(\Sigma_n + \lambda)^{-1}S_n^*E.$$

Let $\mathbb{X} = (X_i)_{i\in[n]}$. When our model is well specified so that $f^* = S\theta_*$, we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_n}\left[\mathcal{E}(f_n)\,|\,\mathbb{X}\right] &= \mathbb{E}_{\mathcal{D}_n}\left[\|f_n - f^*\|_2^2 \mid \mathbb{X}\right] \\
&= \left\|S(\Sigma_n + \lambda)^{-1}\Sigma_n\theta_* - S\theta_*\right\|_2^2 \\
&\qquad + \mathbb{E}_{\mathcal{D}_n}\left[\left\|S(\Sigma_n + \lambda)^{-1}S_n^*E\right\|_2^2 \mid \mathbb{X}\right] \\
&\qquad + 2\mathbb{E}_{\mathcal{D}_n}\left[\left(S(\Sigma_n + \lambda)^{-1}\Sigma_n\theta_* - S\theta_*\right)^\top S(\Sigma_n + \lambda)^{-1}S_n^*E \mid \mathbb{X}\right] \\
&= \lambda^2\left\|S(\Sigma_n + \lambda)^{-1}\theta_*\right\|_2^2 \\
&\qquad + \mathbb{E}_{\mathcal{D}_n}\left[\left\|S(\Sigma_n + \lambda)^{-1}S_n^*E\right\|_2^2 \mid \mathbb{X}\right] \\
&\qquad + 2\left(S_n(\Sigma_n + \lambda)^{-1}\Sigma(\Sigma_n + \lambda)^{-1}\theta_*\right)^\top \mathbb{E}_{\mathcal{D}_n}[E \mid \mathbb{X}] \\
&= \lambda^2\left\|S(\Sigma_n + \lambda)^{-1}\theta_*\right\|_2^2 + \mathbb{E}_{\mathcal{D}_n}\left[\left\|S(\Sigma_n + \lambda)^{-1}S_n^*E\right\|_2^2 \mid \mathbb{X}\right],
\end{aligned}$$

where in the third equality we used $I - (\Sigma_n + \lambda)^{-1}\Sigma_n = \lambda(\Sigma_n + \lambda)^{-1}$, and in the last one $\mathbb{E}_{\mathcal{D}_n}[E\,|\,\mathbb{X}] = 0$. Assuming for simplicity that the noise is homoscedastic so that $\mathbb{E}[EE^\top] = \varepsilon^2 I$ for $\varepsilon > 0$, we obtain

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_n}\left[\left\|S(\Sigma_n + \lambda)^{-1}S_n^*E\right\|_2^2 \mid \mathbb{X}\right] &= \frac{1}{n}\operatorname{Tr}\left(S_n(\Sigma_n + \lambda)^{-1}\Sigma(\Sigma_n + \lambda)^{-1}S_n^*\mathbb{E}_{\mathcal{D}_n}\left[EE^\top \mid \mathbb{X}\right]\right) \\
&= \frac{\varepsilon^2}{n}\operatorname{Tr}\left(\Sigma(\Sigma_n + \lambda)^{-2}\Sigma_n\right),
\end{aligned}$$

where the $1/n$ factor arises from the fact that $E^* = E^\top/n$ in the geometry we have considered on $\mathbb{R}^n$. Finally we have retrieved the following standard bias-variance decomposition result.

Lemma 4 (Bias-Variance decomposition)**.** *When our model is well-specified so that $f^* = S\theta_*$ with $\theta_* \in \mathcal{H}$, and when the noise in the label is homoscedastic with variance ε^2, the estimator* (6) *verifies*

$$\mathbb{E}_{\mathcal{D}_n}\left[\mathcal{E}(f_n)\,|\,\mathbb{X}\right] = \lambda^2\underbrace{\left\|S(\Sigma_n + \lambda)^{-1}\theta_*\right\|_2^2}_{\mathcal{B}_n} + \frac{\varepsilon^2}{n}\underbrace{\operatorname{Tr}\left(\Sigma(\Sigma_n + \lambda)^{-2}\Sigma_n\right)}_{\mathcal{V}_n}. \tag{15}$$

We would like to get the limit when n goes to infinity in equation (15). We expect the first term to concentrate towards $\lambda^2\left\|S(\Sigma + \lambda)^{-1}\theta_*\right\|_2^2$, and the second term to $\operatorname{Tr}(\Sigma^2(\Sigma + \lambda)^{-2})$.

A3.1 Bounding the bias term

Let us begin by working out the term $\mathcal{B}_n = \left\|S(\Sigma_n + \lambda)^{-1}\theta_*\right\|_2^2$. We first introduce some notation to make derivations shorter. Let $E_n = \Sigma_n - \Sigma$, $\Sigma_\lambda = \Sigma + \lambda$ and $F_n = -\Sigma_\lambda^{-1/2}E_n\Sigma_\lambda^{-1/2}$. As long as $\|F_n\| < 1$, we have

$$\begin{aligned}
\mathcal{B}_n &= \theta_*^\top(\Sigma_n + \lambda)^{-1}\Sigma(\Sigma_n + \lambda)^{-1}\theta_* \\
&= \theta_*^\top(\Sigma_\lambda + E_n)^{-1}\Sigma(\Sigma_\lambda + E_n)^{-1}\theta_* \\
&= \theta_*^\top\Sigma_\lambda^{-1/2}(I - F_n)^{-1}\Sigma_\lambda^{-1}\Sigma(I - F_n)^{-1}\Sigma_\lambda^{-1/2}\theta_* \\
&= \sum_{i,j\in\mathbb{N}}\theta_*^\top\Sigma_\lambda^{-1/2}F_n^i\Sigma_\lambda^{-1}\Sigma F_n^j\Sigma_\lambda^{-1/2}\theta_*.
\end{aligned}$$

Let us assume for a moment that

$$\left\|(\Sigma\Sigma_\lambda^{-1})^{-1/2}F_n(\Sigma\Sigma_\lambda^{-1})^{1/2}\right\| \le \|F_n\|. \tag{16}$$

If equation (16) holds, then

$$
\begin{aligned}
\left|\mathcal{B}_n - \theta_*^\top \Sigma_\lambda^{-2}\Sigma\theta_*\right| &= \left|\sum_{i,j\in\mathbb{N};i+j\neq 0} \theta_*^\top \Sigma_\lambda^{-1/2} F_n^i \Sigma_\lambda^{-1}\Sigma F_n^j \Sigma_\lambda^{-1/2}\theta_*\right| \\
&\leq \sum_{i+j\neq 0}\left|\theta_*^\top \Sigma_\lambda^{-1/2} F_n^i \Sigma_\lambda^{-1}\Sigma F_n^j \Sigma_\lambda^{-1/2}\theta_*\right| \\
&= \sum_{i+j\neq 0}\left|\left\langle \Sigma_\lambda^{-1/2}(\Sigma_\lambda^{-1}\Sigma)^{1/2}\theta_*, \left((\Sigma_\lambda^{-1}\Sigma)^{-1/2} F_n^i \Sigma_\lambda^{-1}\Sigma F_n^j (\Sigma_\lambda^{-1}\Sigma)^{-1/2}\right)(\Sigma_\lambda^{-1}\Sigma)^{1/2}\Sigma_\lambda^{-1/2}\theta_*\right\rangle\right| \\
&\leq \sum_{i+j\neq 0}\left\|\Sigma_\lambda^{-1/2}(\Sigma_\lambda^{-1}\Sigma)^{1/2}\theta_*\right\|^2 \left\|(\Sigma_\lambda^{-1}\Sigma)^{-1/2} F_n^i \Sigma_\lambda^{-1}\Sigma F_n^j (\Sigma_\lambda^{-1}\Sigma)^{-1/2}\right\| \\
&= \theta_*^\top \Sigma_\lambda^{-2}\Sigma\theta_* \sum_{i+j\neq 0}\left\|((\Sigma\Sigma_\lambda^{-1})^{-1/2} F_n (\Sigma\Sigma_\lambda^{-1})^{1/2})^{i+j}\right\| \\
&\leq \theta_*^\top \Sigma_\lambda^{-2}\Sigma\theta_* \sum_{i,j\in\mathbb{N};i+j\neq 0}\left\|(\Sigma\Sigma_\lambda^{-1})^{-1/2} F_n (\Sigma\Sigma_\lambda^{-1})^{1/2}\right\|^{i+j} \\
&= \theta_*^\top \Sigma_\lambda^{-2}\Sigma\theta_* \sum_{i\in\mathbb{N}}(i+2)\left\|(\Sigma\Sigma_\lambda^{-1})^{-1/2} F_n (\Sigma\Sigma_\lambda^{-1})^{1/2}\right\|^{i+1} \\
&\leq \theta_*^\top \Sigma_\lambda^{-2}\Sigma\theta_* \sum_{i\in\mathbb{N}}(i+2)\left\|F_n\right\|^{i+1} \\
&= \theta_*^\top \Sigma_\lambda^{-2}\Sigma\theta_* \int_0^\infty (\lfloor x\rfloor + 2)\left\|F_n\right\|^{\lceil x\rceil}\,\mathrm{d}x \\
&\leq \theta_*^\top \Sigma_\lambda^{-2}\Sigma\theta_* \int_0^\infty (x+2)\left\|F_n\right\|^{x}\,\mathrm{d}x \\
&= \theta_*^\top \Sigma_\lambda^{-2}\Sigma\theta_* \frac{1-2\log(\|F_n\|)}{\log^2(\|F_n\|)}.
\end{aligned}
$$

This inequality is useful as long as $\|F_n\|$ is small enough, which is not always true. When $\|F_n\|$ is large, we can instead proceed with the simpler bound

$$
\left|\mathcal{B}_n - \theta_*^\top \Sigma_\lambda^{-2}\Sigma\theta_*\right| \leq \mathcal{B}_n + \theta_*^\top \Sigma_\lambda^{-2}\Sigma\theta_* \leq 2\theta_*^\top \Sigma\theta_*\lambda^{-2} = 2\left\|f^*\right\|_2^2\lambda^{-2}.
$$

Therefore, rewriting the limit as

$$
\begin{aligned}
\lambda^2\theta_*^\top(\Sigma+\lambda)^{-2}\Sigma\theta_* &= \lambda^2(\Sigma^{1/2}\theta_*)^\top(\Sigma+\lambda)^{-2}\Sigma^{1/2}\theta_* = \lambda^2(S\theta_*)^\top(K+\lambda)^{-2}S\theta_* \\
&= \lambda^2(f^*)^\top(K+\lambda)^{-2}f^* = \left\|\lambda(K+\lambda)^{-1}f^*\right\|_2^2 = \mathcal{B}(\lambda),
\end{aligned}
$$

we split the bias error as

$$
\begin{aligned}
\left|\lambda^2\mathcal{B}_n - \lambda^2\theta_*^\top \Sigma_\lambda^{-2}\Sigma\theta_*\right| &\leq 2\left\|f^*\right\|_2^2 \mathbf{1}_{\|F_n\|>1/2} + \mathcal{B}(\lambda)\frac{1-2\log(\|F_n\|)}{\log^2(\|F_n\|)}\mathbf{1}_{\|F_n\|\leq 1/2} \\
&\leq 2\left\|f^*\right\|_2^2 \mathbf{1}_{\|F_n\|>1/2} - \frac{3\mathcal{B}(\lambda)}{\log(\|F_n\|)}\mathbf{1}_{\|F_n\|\leq 1/2}.
\end{aligned}
$$

Taking the expectation and using the convexity of the absolute value, we obtain

$$
\left|\mathbb{E}_{\mathcal{D}_n}[\lambda^2\mathcal{B}_n] - \lambda^2\theta_*^\top \Sigma_\lambda^{-2}\Sigma\theta_*\right| \leq 2\left\|f^*\right\|_2^2 \mathbb{P}(\|F_n\| > 1/2) - \mathcal{B}(\lambda)\int_0^{1/2}\frac{3\mathbb{P}\left(\|F_n\| > x\right)}{\log(x)}\,\mathrm{d}x.
$$

We now proceed with an exponential concentration inequality on $\|F_n\|$. We will use the one of Cabannes et al. [5], Eq. (25). As long as $\lambda \leq \|\Sigma\|$, we have

$$
\mathbb{P}(\|F_n\| > t) \leq 28\mathcal{N}_1(\lambda)\exp\left(-\frac{nt^2}{\mathcal{N}_\infty(\lambda)(1+t)}\right).
$$

This inequality shows the restrictive notion of effective dimension, which is useful to ensure the good conditioning of the linear system implicitly encoded in (6), and, in essence, bound all moments of $\varphi(X)$. As long as $\mathcal{N}_\infty(\lambda) \leq 3n/2$, we can compute the integral as

$$
\begin{aligned}
-\int_0^{1/2} \frac{\mathbb{P}\left(\|F_n\| > x\right)}{\log(x)} \, \mathrm{d}x &\leq -28\mathcal{N}_1(\lambda) \int_0^{1/2} \frac{\exp(-3nx^2/2\mathcal{N}_\infty(\lambda))}{\log(x)} \, \mathrm{d}x \\
&= -28\mathcal{N}_1(\lambda) \frac{\mathcal{N}_\infty^{1/2}(\lambda)}{1.5^{1/2} n^{1/2}} \int_0^{1/2} \frac{\exp(-u^2)}{\log(u) + \log(3n/2\mathcal{N}_\infty^2(\lambda))} \, \mathrm{d}u \\
&\leq -28 \frac{\mathcal{N}_1(\lambda)\mathcal{N}_\infty^{1/2}(\lambda)}{1.5^{1/2} n^{1/2}} \int_0^{1/2} \frac{\exp(-u^2)}{\log(u)} \, \mathrm{d}u \leq \frac{8\mathcal{N}_1(\lambda)\mathcal{N}_\infty^{1/2}(\lambda)}{n^{1/2}}.
\end{aligned}
$$

We recall that the bounds above were derived under condition (16), which is rather strong. However, the attentive reader would remark that a much laxer assumption is sufficient, which we introduce thereafter.

Assumption 1. *There exists a constant c such that, for all $i, j \in \mathbb{N}$,*

$$
\mathbb{E}\left[\left\|(\Sigma\Sigma_\lambda^{-1})^{-.5} F_n^i \Sigma\Sigma_\lambda^{-1} F_n^j (\Sigma\Sigma_\lambda^{-1})^{-.5}\right\| \,\middle|\, \|F_n\| \leq 1/2\right] \leq c^2 \mathbb{E}\left[\|F_n\|^{i+j} \,\middle|\, \|F_n\| \leq 1/2\right]. \quad (17)
$$

Assumption 1 notably holds when $\mathcal{F}$ is finite dimensional with $c^2 = \left\|K^{-1}\right\|_2^{-1} (\|K\|_2 + \lambda)$. As such all our lower-bound results can be cast with finite-dimensional approximation of infinite-dimensional RKHS. Under Assumption 1, we get

$$
\begin{aligned}
&\mathbb{E}\left[\left|\mathcal{B}_n - \theta_*^\top \Sigma_\lambda^{-2} \Sigma \theta_*\right| \,\middle|\, \|F_n\| \leq 1/2\right] \\
&\leq \sum_{i+j\neq 0} \left\|\Sigma_\lambda^{-1/2} (\Sigma_\lambda^{-1}\Sigma)^{1/2} \theta_*\right\|^2 \mathbb{E}\left[\left\|(\Sigma_\lambda^{-1}\Sigma)^{-1/2} F_n^i \Sigma_\lambda^{-1} \Sigma F_n^j (\Sigma_\lambda^{-1}\Sigma)^{-1/2}\right\| \,\middle|\, \|F_n\| \leq 1/2\right] \\
&\leq c^2 \sum_{i,j\in\mathbb{N}; i+j\neq 0} \left\|\Sigma_\lambda^{-1/2} (\Sigma_\lambda^{-1}\Sigma)^{1/2} \theta_*\right\|^2 \mathbb{E}\left[\|F_n\|^{i+j} \,\middle|\, \|F_n\| \leq 1/2\right],
\end{aligned}
$$

which allows us to proceed with the precedent derivations without assuming that (16) holds.

While the previous results were achieved for $f^* \in \mathcal{F}$, they can be extended by density to any f^* in the closure of $\mathcal{F}$ in $L^2(\rho_\mathcal{X})$, i.e. $f \in (\ker K)^\perp$, leading to the following result.

Proposition 5. *When $f^* \in (\ker K)^\perp$ and $\lambda \leq \|\Sigma\|$, under the technical Assumption 1, the bias term can be bounded from above and below by*

$$
\left|\mathbb{E}_{\mathcal{D}_n}[\mathcal{B}_n] - \mathcal{S}(\lambda)\right| \leq c^2 \mathcal{N}_1(\lambda) \left(\frac{56 \left\|f^*\right\|_2^2}{\lambda^2} \exp\left(-\frac{n}{6\mathcal{N}_\infty(\lambda)}\right) + \frac{8\mathcal{S}(\lambda)\mathcal{N}_\infty^{1/2}(\lambda)}{n^{1/2}}\right). \quad (18)
$$

A3.2 Discussion on the bias bound

More direct upper bound. The precedent derivations can be made more direct with the following series of implications, with $A \preceq B$ meaning that $x^\top A x \leq x^\top B x$ for every x:

$$
\begin{aligned}
\|F_n\| \leq 1/2 \quad &\Rightarrow \quad -1/2 I \preceq F_n \preceq 1/2 I \\
&\Rightarrow \quad 1/2 I \preceq I - F_n \preceq 3/2 I \\
&\Rightarrow \quad 4/9 I \preceq (I - F_n)^{-2} \preceq 4 I \\
&\Rightarrow \quad 4/9 \mathcal{B}(\lambda) \preceq \theta_*^\top \Sigma^{1/2} \Sigma_\lambda^{-1} (I - F_n)^{-2} \Sigma_\lambda^{-1} \Sigma^{1/2} \theta_* \preceq 4\mathcal{B}(\lambda).
\end{aligned}
$$

Let us assume that

$$
\begin{aligned}
&\mathbb{E}\left[(I - F_n)^{-1} \Sigma \Sigma_\lambda^{-1} (I - F_n)^{-1} \,\middle|\, \|F_n\| \leq 1/2\right] \\
&\qquad \preceq c^2 \mathbb{E}\left[\Sigma^{1/2} \Sigma_\lambda^{-1/2} (I - F_n)^{-2} \Sigma^{1/2} \Sigma_\lambda^{-1/2} \,\middle|\, \|F_n\| \leq 1/2\right],
\end{aligned} \quad (19)
$$

which is always verified for $c^2 = \min(\lambda^{-1}, \|K^{-1}\|_2^{-1})\,\|K\|_2$ (although taking $c^2 \propto \lambda$ would slow down our upper bound by a factor of λ). This leads to the simple upper bound

$$\begin{aligned}
\mathcal{B}_n &= \theta_*^\top \Sigma_\lambda^{-1/2}(I - F_n)^{-1}\Sigma_\lambda^{-1}\Sigma(I - F_n)^{-1}\Sigma_\lambda^{-1/2}\theta_* \\
&\leq c^2 \theta_*^\top \Sigma_\lambda^{-1/2}(\Sigma_\lambda^{-1})^{1/2}(I - F_n)^{-2}(\Sigma_\lambda^{-1})^{1/2}\Sigma_\lambda^{-1/2}\theta_* + \|f^*\|_{L^2}^2\,\mathbb{P}(\|F_n\| > 1/2) \\
&\leq 4c^2\mathcal{S}(\lambda) + \|f^*\|_2^2\,\mathcal{N}_1(\lambda)\exp\left(-\frac{n}{6\mathcal{N}_\infty(\lambda)}\right).
\end{aligned}$$

Improvement directions. In essence, we expect the bias upper-lower bound to behave as

$$\mathcal{S}(\lambda)(I - F_n)^{-2} - I \simeq \mathcal{S}(\lambda)F_n.$$

Getting this linear dependency in F_n explicitly would allow to improve the bound since

$$\mathbb{E}_{\mathcal{D}_n}\left[\|F_n\| \,|\, \|F_n\| \leq 1/2\right] \lesssim \mathcal{N}(\lambda)\mathcal{N}_\infty(\lambda)n^{-1}.$$

Moreover, going back to the definition of F_n,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_n}\theta_*^\top \Sigma\Sigma_\lambda^{-1}F_n\Sigma\Sigma_\lambda^{-1}\theta_* &= \mathbb{E}_{\mathcal{D}_n}\left[\left(\frac{1}{n}\sum_{i=1}^n \theta_*^\top \Sigma\Sigma_\lambda^{-3/2}\varphi(X_i)\right)^2\right] - \mathbb{E}_X\left[\theta_*^\top \Sigma\Sigma_\lambda^{-3/2}\varphi(X)\right]^2 \\
&= \mathbb{E}_{\mathcal{D}_n}\left[\left(\frac{1}{n}\sum_{i=1}^n \theta_*^\top \Sigma\Sigma_\lambda^{-3/2}\varphi(X_i) - \mathbb{E}_X\left[\theta_*^\top \Sigma\Sigma_\lambda^{-3/2}\varphi(X)\right]\right)^2\right],
\end{aligned}$$

which suggests possible improvements of the bound in $\mathcal{S}(\lambda)n^{-1}$.

A3.3 Bounding the variance term

Let us now work on the term $\mathcal{V}_n = \operatorname{Tr}\Sigma(\Sigma_n + \lambda)^{-2}\Sigma_n$. It works similarly to the bias term, concentrating towards $\mathcal{N}_2(\lambda)$. With $\Sigma_{n,\lambda} = \Sigma_n + \lambda I$, we have

$$\operatorname{Tr}\left(\Sigma\Sigma_n\Sigma_{n,\lambda}^{-2} - \Sigma^2\Sigma_\lambda^{-2}\right) = \operatorname{Tr}\left(\Sigma\Sigma_\lambda^{-1}\Sigma_n\Sigma_{n,\lambda}^{-1}(\Sigma_{n,\lambda}^{-1}\Sigma_\lambda - I) + \Sigma\Sigma_\lambda^{-1}(\Sigma_n\Sigma_{n,\lambda}^{-1} - \Sigma\Sigma_\lambda^{-1})\right).$$

Using that, for any A positive semi-definite and any B, $\operatorname{Tr}(AB) \leq \|B\|\operatorname{Tr}(A)$, it follows that

$$\begin{aligned}
|\mathcal{V}_n - \mathcal{N}_2(\lambda)| &\leq \mathcal{N}_1(\lambda)\left\|\Sigma_n\Sigma_{n,\lambda}^{-1}\right\|\left\|\Sigma_{n,\lambda}^{-1}\Sigma_\lambda - I\right\| + \mathcal{N}_1(\lambda)\left\|\Sigma_n\Sigma_{n,\lambda}^{-1} - \Sigma\Sigma_\lambda^{-1}\right\| \\
&\leq \mathcal{N}_1(\lambda)\left\|\Sigma_{n,\lambda}^{-1}\Sigma_\lambda - I\right\| + \mathcal{N}_1(\lambda)\left\|\Sigma_n\Sigma_{n,\lambda}^{-1} - \Sigma\Sigma_\lambda^{-1}\right\|.
\end{aligned}$$

Let us focus on the first term. Using $a^{-1} - b^{-1} = a^{-1}(b - a)b^{-1}$, we get

$$\begin{aligned}
\left\|\Sigma_\lambda\Sigma_{n,\lambda}^{-1} - I\right\| &= \left\|\Sigma_\lambda\Sigma_{n,\lambda}^{-1}(\Sigma - \Sigma_n)\Sigma_\lambda^{-1}\right\| \\
&\leq \left\|\Sigma_\lambda\Sigma_{n,\lambda}^{-1}\right\|\left\|\Sigma_\lambda^{1/2}F_n\Sigma_\lambda^{-1/2}\right\| \\
&\leq \left(\|I\| + \left\|\Sigma_\lambda\Sigma_{n,\lambda}^{-1} - I\right\|\right)\left\|\Sigma_\lambda^{1/2}F_n\Sigma_\lambda^{-1/2}\right\| \\
&\leq \sum_{i>0}\left\|\Sigma_\lambda^{1/2}F_n\Sigma_\lambda^{-1/2}\right\|^i.
\end{aligned}$$

For the second term,

$$\begin{aligned}
\left\|\Sigma_n\Sigma_{n,\lambda}^{-1} - \Sigma\Sigma_\lambda^{-1}\right\| &\leq \left\|\Sigma_n(\Sigma_{n,\lambda}^{-1} - \Sigma_\lambda^{-1}\right\| + \left\|(\Sigma_n - \Sigma)\Sigma_\lambda^{-1}\right\| \\
&\leq \left\|\Sigma_n\Sigma_{n,\lambda}^{-1}(\Sigma - \Sigma_n)\Sigma_\lambda^{-1}\right\| + \left\|(\Sigma_n - \Sigma)\Sigma_\lambda^{-1}\right\| \\
&\leq 2\left\|(\Sigma_n - \Sigma)\Sigma_\lambda^{-1}\right\| \\
&= 2\left\|\Sigma_\lambda^{1/2}F_n\Sigma_\lambda^{-1/2}\right\|.
\end{aligned}$$

Similarly as for (16), if we assume that

$$\left\|\Sigma_\lambda^{1/2} F_n \Sigma_\lambda^{-1/2}\right\| \leq c \left\|F_n\right\|, \tag{20}$$

then we have, as long as $\|F_n\| \leq 1/2c$,

$$\begin{aligned}|\mathcal{V}_n - \mathcal{N}_2(\lambda)| &\leq \sum_{i>0} \left\|\Sigma_\lambda^{1/2} F_n \Sigma_\lambda^{-1/2}\right\|^i + 2\left\|\Sigma_\lambda^{1/2} F_n \Sigma_\lambda^{-1/2}\right\| \\ &\leq \frac{c\left\|F_n\right\|}{1 - c\left\|F_n\right\|} + 2c\left\|F_n\right\| \leq 4c\left\|F_n\right\|.\end{aligned}$$

For the case when $\|F_n\|$ is large, we can again proceed with a simple bound:

$$|\mathcal{V}_n - \mathcal{N}_2(\lambda)| \leq \left|\operatorname{Tr}\left(\Sigma\Sigma_n(\Sigma_n + \lambda)^{-2}\right)\right| + \operatorname{Tr}\left(\Sigma^2\Sigma_\lambda^{-2}\right) \leq 2\operatorname{Tr}(\Sigma)\lambda^{-1}.$$

Splitting the full expectation as we did for the bias we thus obtain

$$\begin{aligned}&|\mathbb{E}_{\mathcal{D}_n}[\mathcal{V}_n] - \mathcal{N}_2(\lambda)| \leq \mathbb{E}_{\mathcal{D}_n}[|\mathcal{V}_n - \mathcal{N}_2(\lambda)|] \\ &\leq 2\operatorname{Tr}(\Sigma)\lambda^{-1}\mathbb{P}(\|F_n\| > 1/2c) + 4c\mathbb{E}\left[\|F_n\| \,|\, \|F_n\| \leq 1/2c\right]\mathbb{P}(\|F_n\| \leq 1/2c).\end{aligned}$$

We are left with the computation of two integrals. As long as $\lambda \leq \|\Sigma\|$, with a the coefficient appearing in the exponential

$$\begin{aligned}&\mathbb{P}(\|F_n\| \leq 1/2c)\mathbb{E}\left[\|F_n\| \,|\, \|F_n\| \leq 1/2c\right] \leq 28\mathcal{N}_1(\lambda)\int_0^{1/2c} x\exp(-3nx^2/2\mathcal{N}_\infty(\lambda))\,\mathrm{d}x \\ &= 28\mathcal{N}_1(\lambda)a^{-1}\int_0^{a^{1/2c}/2} x\exp(-x^2)\,\mathrm{d}x \leq \frac{10\mathcal{N}_1(\lambda)\mathcal{N}_\infty(\lambda)}{n}.\end{aligned}$$

Once again, the condition (20) can be relaxed with the following assumption.

Assumption 2. *There exists a constant c such that, for all $i \in \mathbb{N}$,*

$$\mathbb{E}\left[\left\|\Sigma_\lambda^{1/2} F_n \Sigma_\lambda^{-1/2}\right\|^i \,\middle|\, \|F_n\| \leq 1/2c\right] \leq c^i \mathbb{E}\left[\|F_n\|^i \,\middle|\, \|F_n\| \leq 1/2c\right]. \tag{21}$$

As for Assumption 1, Assumption 2 holds when $\mathcal{F}$ is finite-dimensional with $c^2 = \left\|K^{-1}\right\|(\|K\| + \lambda)$. It holds in general with $c^2 = \lambda^{-1}(\|\Sigma\| + \lambda)$, although this would deteriorate the bound by a factor of λ.

We are finally ready to collect the different pieces.

Proposition 6. *When $f^* \in (\ker K)^\perp$ and $\lambda \leq \|\Sigma\|$, under the technical Assumption 2, the variance term can be bounded from above and below by*

$$\frac{\varepsilon^2}{n}\left|\mathbb{E}_{\mathcal{D}_n}[\mathcal{V}_n] - \mathcal{N}_2(\lambda)\right| \leq \varepsilon^2\mathcal{N}_1^2(\lambda)\left(\frac{28\operatorname{Tr}(\Sigma)}{n\lambda}\exp\left(-\frac{c^2 n}{(4+2c)\mathcal{N}_\infty(\lambda)}\right) + \frac{40c\mathcal{N}_\infty(\lambda)}{n^2}\right). \tag{22}$$

A3.4 Discussion to the variance bound

Once again, the bound presented here is somewhat unsatisfying, as it will not necessarily decrease faster than $\mathcal{N}(\lambda)/n$ if one considers target functions that are far away from $\mathcal{F}$ and requires a large search space (i.e. $\mathcal{B}(\lambda)$ decreases slowly with λ, and given a fixed number of samples n the optimal λ_n is found for $\mathcal{N}(\lambda_n)$ quite large compared to n), so that $\mathcal{N}_\infty(\lambda)\mathcal{N}(\lambda)/n$ does not go to zero. Several directions can be taken to improve the bound. For example, when Y is bounded by M, the noise ε^2 can be replaced by M^2, and it is possible to get an upper bound of the form

$$\mathbb{E}_{\mathcal{D}_n}[\mathcal{E}(f_{\lambda,n}^{(\text{thres.})})] \leq \frac{8M^2}{n}\mathcal{N}_1(\lambda) + \inf_{f\in\mathcal{F}}\left(\|f - f^*\|_{L^2}^2 + \lambda\|f\|_{\mathcal{F}}^2\right),$$

for a truncated version $f_{\lambda,n}^{(\text{thres.})}$ of the estimator (6) as proved by Mourtada et al. [17] for ridge-less regression and extended to ridge regression in Mourtada et al. [18]. Moreover, retaking the analysis of Mourtada and Rosasco [16], one can get a lower bound of the form

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_n}\mathcal{V}_n &\geq \frac{n}{n+1}\mathbb{E}_{\mathcal{D}_{n+1}}\operatorname{Tr}\left(\Sigma_n\Sigma_{n,\lambda}^{-1}\Sigma_{n+1}\Sigma_{n+1,\lambda(n+1)/n}^{-1}\right) \\ &\geq \frac{n}{n+1}\mathbb{E}_{\mathcal{D}_n}\operatorname{Tr}\left(\Sigma_n\Sigma_{n,\lambda}^{-1}\right) - \lambda\mathbb{E}_{\mathcal{D}_{n+1}}\operatorname{Tr}\left(\Sigma_n\Sigma_{n,\lambda}^{-1}\Sigma_{n+1,\lambda(n+1)/n}^{-1}\right),\end{aligned}$$

which might lead to some lower bound with

$$\mathbb{E}_{\mathcal{D}_{n+1}} \operatorname{Tr}\left(\Sigma_n \Sigma_{n,\lambda}^{-1} \Sigma_{n+1,\lambda(n+1)/n}^{-1}\right) = \frac{n+1}{n} \mathbb{E}_{\mathcal{D}_n, X} \operatorname{Tr}\left(\Sigma_n \Sigma_{n,\lambda}^{-1} (\Sigma_{n,\lambda} + \varphi(X) \otimes \varphi(X))^{-1}\right)$$
$$\leq \frac{n+1}{n} \mathbb{E}_{\mathcal{D}_n} \operatorname{Tr}\left(\Sigma_n \Sigma_{n,\lambda}^{-1}\right) \mathbb{E}_X \left[\left\|(\Sigma_{n,\lambda} + \varphi(X) \otimes \varphi(X))^{-1}\right\|\right].$$

We also note that it should not be too hard to replace F_n by $\Sigma^{1/2}\Sigma_\lambda^{-1} E_n \Sigma^{1/2}\Sigma_\lambda^{-1}$, which would lead to $\varepsilon^2 \mathcal{N}_2(\lambda)/n$ instead of $\varepsilon^2 \mathcal{N}_1(\lambda)/n$ in the right-hand side of (10).

A4 Full theorem

Collecting the precedent results leads to the following theorem.

Theorem 3. *Under the technical Assumptions 1 and 2, as long as $\lambda \leq \|\Sigma\|$, when f^* belongs to the closure of $\mathcal{F}$ in $L^2(\rho_{\mathcal{X}})$, the estimator* (6) *verifies*

$$\left|\mathbb{E}_{\mathcal{D}_n}\left[\mathcal{E}(f_{\lambda,n})\right] - \frac{\varepsilon^2 \mathcal{N}_2(\lambda)}{n} - \lambda^2 \mathcal{S}(\lambda)\right| \leq \mathcal{N}_1(\lambda)\left(a_n \cdot \frac{\varepsilon^2 \mathcal{N}_1(\lambda)}{n} + a_n^{1/2} \lambda^2 \mathcal{S}(\lambda)\right)$$
$$+ \mathcal{N}_1(\lambda)\left(\frac{\operatorname{Tr}(\Sigma)\varepsilon^2}{n} + \|f^*\|_{L^2}^2\right) \exp(-c a_n), \tag{23}$$

where $a_n = \mathcal{N}_\infty(\lambda)/n$.

As long as its right-hand side decreases faster then $\mathbb{E}_{\mathcal{D}_n}[\mathcal{E}(f_n)]$, Theorem 3 states that $\mathbb{E}_{\mathcal{D}_n}[\mathcal{E}(f_n)]$ behaves like

$$\mathbb{E}_{\mathcal{D}_n}\left[\mathcal{E}(f_n)\right] \simeq \frac{\varepsilon^2 \mathcal{N}_2(\lambda)}{n} + \lambda^2 \mathcal{S}(\lambda).$$

When optimizing for λ, assuming that $\mathcal{N}_1 \simeq \mathcal{N}_2$, the right-hand side of Theorem 3 decreases faster than $\mathbb{E}_{\mathcal{D}_n}[\mathcal{E}(f_n)]$ if and only if $\mathcal{N}_1(\lambda) a_n^{-1/2}$ goes to zero with n. This implies $\mathcal{N}(\lambda)^3 \leq n$, which is a much stronger condition than the high-sample regime condition $\mathcal{N}(\lambda) \leq n$.

A4.1 Upper bound application to local polynomials

The following recalls a proof of convergence rates for local polynomial estimation. Consider the case where $\mathcal{X} = [0, 1)$ with uniform distribution, and $f = f^*$ is assumed to be (α, L)-Hölder,

$$\left|f^{(\lfloor\alpha\rfloor)}(x) - f^{(\lfloor\alpha\rfloor)}(y)\right| \leq L\,|x - y|^{\alpha - \lfloor\alpha\rfloor}.$$

Then, by fitting Taylor expansions on intervals $[(i-1)/m, i/m)$ for $i \in [m]$ and $m \in \mathbb{Z}_+$ with polynomials

$$\varphi(x) = \left(\left(x - \frac{2i-1}{2m}\right)^j \cdot \mathbf{1}_{x \in [\frac{i-1}{m}, \frac{i}{m})}\right)_{i \in [m], j \in [0, \lfloor\alpha\rfloor]},$$

one can ensure that [see 11, Lemma 11.1]

$$\|\Pi_{\mathcal{F}} f^* - f^*\|_2 \leq \|\Pi_{\mathcal{F}} f^* - f^*\|_\infty \leq \frac{L}{2^\alpha \lfloor\alpha\rfloor!\, m^\alpha},$$

where $\Pi_{\mathcal{F}}$ denotes the orthogonal projection from L^2 onto $\mathcal{F}$. The same type of result also holds in dimension d using m^d multivariate polynomials of degree less then $\lfloor\alpha\rfloor$. The number of such polynomials is $m^d h_{\lfloor\alpha\rfloor}(1_d)$, where $h_{\lfloor\alpha\rfloor}$ is the complete homogeneous symmetric polynomial of degree $\lfloor\alpha\rfloor$ in d variables and 1_d is the vector of all ones. Thus,

$$h_{\lfloor\alpha\rfloor}(1_d) = \binom{d + \lfloor\alpha\rfloor}{\alpha} = \frac{(d + \lfloor\alpha\rfloor)!}{d!\,\lfloor\alpha\rfloor!} \geq \frac{\lceil\alpha\rceil^d}{d!}.$$

Balancing the bias and the variance term, we get an excess risk that behaves as

$$\mathbb{E}\left[\|f_n - f^*\|^2\right] \lesssim \inf_{m \in \mathbb{Z}_+} \frac{\varepsilon^2 m^d \lceil\alpha\rceil^d}{n} + \frac{L^2}{2^{2\alpha} \lfloor\alpha\rfloor! m^{2\alpha}} = c n^{-2\alpha/(d+2\alpha)},$$

for some constant c that depends on α, d and grows with L, σ^2, the infimum being found for $m \propto (L^2 n)^{1/(d+2\alpha)}$. This shows an important point for practitioners: the size of the window should depend on how smooth f^* is expected to be among the functions in C^α.

Covering issues in high-dimension? Intuitively in high-dimension problems, where $d = \dim(\mathcal{X})$ is big, leveraging local properties is not very reasonable, since the covering of $\mathcal{X}$ with local neighborhoods grows exponentially with the dimension d, meaning that if one wants to have enough samples per neighborhood, n should scale exponentially with d. Because rates in $O(n^{-2\alpha/(2\alpha+d)})$ are minimax optimal and of the lower bound in (10), it appears explicitly that to ensure minimax optimal convergence rates (i.e. make sure the rates hold for all functions in C^α), the partition size of $\mathcal{X}$ should scale in $O(n^{d/(2\alpha+d)}) = O(n \times n^{-2\alpha/(2\alpha+d)})$ when trying to leverage the fact that $f^* \in C^\alpha$. Indeed and in contrast with the prior intuition, the partition size does not deteriorate with the dimension of the input space nor with the regularity of the target function, nor the percentage of the total volume contained in each region of the partition.[2] As a consequence, we expect the length of transitory regimes to suffer from the difficulty to leverage smoothness in high-dimension rather than to the fact that Taylor expansions need to be localized, and expect similar pessimistic pictures to take place when estimating target function through Fourier expansion.

A4.2 Upper bound application to translation-invariant kernels

Table 1 depicts two types of kernels: the Matérn kernels (the exponential kernel corresponding to a Matérn kernel of low smoothness), and the Gaussian kernel (which can be seen as the limit of a Matérn kernel to infinite smoothness). By balancing bias and variance, one can prove the usual convergence rates for functions in Sobolev spaces, i.e. $f^* \in H^\alpha$. We refer to Appendix B for an explanation of those bounds on the variance and bias terms.

For the Matérn kernels, the generalization error reads, with $\tau = 2\beta - d$,

$$\mathbb{E}\left[\|f_n - f^*\|^2\right] \lesssim \frac{(\sigma^\tau \lambda)^{-d/2\beta}}{n} + (\sigma^\tau \lambda)^{\alpha/\beta}.$$

This is optimized for

$$\sigma^\tau \lambda = n^{-2\beta/(2\alpha+d)},$$

leading to minimax convergence rates in $O(n^{-2\alpha/(2\alpha+d)})$.

For the Gaussian kernel, we get

$$\mathbb{E}\left[\|f_n - f^*\|^2\right] \lesssim \frac{\sigma^{-d}\log(\lambda^{-1}\sigma^d)^{d/2}}{n} + \sigma^{2\alpha}\log(\lambda^{-1}\sigma^d)^{-\alpha},$$

which is optimized for

$$\sigma^{-2}\log(\lambda^{-1}\sigma^d) = n^{2/(2\alpha+d)},$$

leading to the same minimax convergence rate. In particular, when σ is fixed, this leads to

$$\lambda = \lambda_n = \sigma^d \exp(-\sigma^2 n^{2/(2\alpha+d)}).$$

Based on Theorem 2, this is true as long as $\mathcal{N}(\lambda)\mathcal{N}_\infty^{1/2}(\lambda)/n^{1/2}$ goes to zero with n, which imposes some constraints on α when assuming $f^* \in H^\alpha$. However, considering the refinement of Mourtada et al. [18], the upper bound is actually true without this constraint, which allows to prove the convergence rates in $O(n^{2\alpha/(2\alpha+d)})$ for any α.

A4.3 Lower bound application to local polynomials - Proof of Proposition 1

Proposition 1 is a straightforward adaptation of Theorem 2, using the fact that $\mathcal{N}(0) = \dim \mathcal{F}$ is larger than the number of unknowns in a single Taylor expansion of order α in dimension d, which corresponds to the number of coefficients in a polynomial of degree α with d variables, and is equal to the number of sets of d elements among $d + \alpha$ elements.

[2] However, the radius of those regions will scale as $r = v^{1/d}$ for v the volume of those regions, meaning that when this volume will shrink to zero, the radius will shrink slower as the dimension grows, which will lead to a slower minimization of the approximation error.

A4.4 Lower bound application to translation-invariant kernels - Proof of Proposition 2

Using the bias and variance lower bounds decomposition and Proposition 14, we get, when $f^* = f_m$ is a single frequency function $\hat{f}(\omega) = \delta_m(\omega)$ with $m \in \mathbb{Z}^d$,

$$\frac{\varepsilon^2 \mathcal{N}_2(\lambda)}{n} + \mathcal{B}(\lambda) \geq \left(\frac{d\pi^{(d-1)/2}}{2^{d+2}\Gamma((d-1)/2)n} \lambda^{-d/2\beta} + \frac{\lambda^2}{((1+\|m\|^2)^{-\beta} + \lambda)^2} \right)$$
$$\geq \frac{1}{2} \left(\frac{d\pi^{(d-1)/2}}{2^{d+1}\Gamma((d-1)/2)n} \lambda^{-d/2\beta} + \min(\lambda^2 (1+\|m\|^2)^{2\beta}, 1) \right).$$

From the fact that

$$\inf_{x \in \mathbb{R}} ax^{-\alpha} + bx^2 = \left(\alpha^{2/(2+\alpha)} + \alpha^{-\alpha/(2+\alpha)} \right) b^{\alpha/(2+\alpha)} a^{2/(2+\alpha)} \geq b^{\alpha/(2+\alpha)} a^{2/(2+\alpha)},$$

we have that, whatsoever λ is (if it is fixed for all n independently of the realization $\mathcal{D}_n$).

$$\frac{d\pi^{(d-1)/2}}{2^{d+1}\Gamma((d-1)/2)n} \lambda^{-d/2\beta} + \lambda^2 (1+\|m\|^2)^{2\beta}$$
$$\geq \left(\frac{d\pi^{(d-1)/2}}{2^{d-1}\Gamma((d-1)/2)} \right)^{4\beta/(4\beta+d)} \left(1 + \|m\|^2 \right)^{2\beta d/(4\beta+d)}.$$

Once again, this lower bound is also true when d is replaced by any $l \in [d]$.

A5 Interpolation spaces, capacity and source conditions

In this section, we discuss the values of $\mathcal{N}(\lambda)$ and $\mathcal{B}(\lambda)$ for classical problems.

A5.1 Relation between variances

We begin with simple facts.

Proposition 7. *For any search space $\mathcal{F}$ and regularization $\lambda > 0$,*

$$\mathcal{N}_2(\lambda) \leq \mathcal{N}_1(\lambda) \leq \mathcal{N}_\infty(\lambda). \tag{24}$$

Once again, the precise study of the variance is easier in $\mathcal{H}$ with the operator Σ rather than in L^2. First of all, notice that $K = SS^*$ has the same spectrum as $\Sigma = S^*S$, so that, for $a \in [1, 2]$,

$$\mathcal{N}_a(\lambda) = \mathrm{Tr}((K+\lambda)^{-a} K^a) = \mathrm{Tr}((\Sigma+\lambda)^{-a}\Sigma^a) = \sum_{\mu \in \mathrm{spec}(\Sigma)} \frac{\mu^a}{(\lambda+\mu)^a}.$$

This shows the first part of the inequality (24):

$$\left(0 \leq \frac{x}{x+\lambda} \leq 1 \qquad \Rightarrow \qquad \frac{x}{x+\lambda} \leq \frac{x^a}{(x+\lambda)^a} \right) \qquad \Rightarrow \qquad \mathcal{N}_2(\lambda) \leq \mathcal{N}_1(\lambda).$$

For the second part of the inequality, we need to reformulate $\mathcal{N}_\infty(\lambda)$.

Lemma 8. *$\mathcal{N}_\infty(\lambda)$ can be expressed in $\mathcal{H}$ as*

$$\mathcal{N}_\infty(\lambda) = \operatorname*{ess\,sup}_{x \sim \rho_\mathcal{X}} \left\| (\Sigma+\lambda)^{-1/2} \varphi(x) \right\|^2.$$

Proof. Observe that, for $x \in \mathcal{X}$,

$$\left\| (\Sigma+\lambda)^{-1/2} \varphi(x) \right\|^2 = \varphi(x)^\top (\Sigma+\lambda)^{-1} \varphi(x) = \mathrm{Tr}\left((\Sigma+\lambda)^{-1} \varphi(x) \otimes \varphi(x) \right).$$

Let us introduce the operator

$$S_x : \mathcal{H} \to L^2(\rho_\mathcal{X}), \quad \theta \mapsto (x' \mapsto \varphi(x)^\top \theta).$$

From

$$\langle S_x \theta, g \rangle = \mathbb{E}[g(X) \varphi(x)^\top \theta] = \langle \theta, \mathbb{E}_X[g(X)\varphi(x)] \rangle$$

we get $S_x^* g = \mathbb{E}[g(X)\varphi(x)]$. Similarly, one can check that

$$K_x(g)(x') = (S_x S_x^* g)(x') = (S_x(\mathbb{E}_X[g(X)]\varphi(x)))(x') = \varphi(x)^\top \varphi(x)\mathbb{E}_X[g(X)] = \mathbb{E}[g]k(x,x),$$

and that

$$\Sigma_x \theta = S_x^* S_x \theta = \mathbb{E}_X[\varphi(x)^\top \theta]\varphi(x) = (\varphi(x) \otimes \varphi(x))\theta,$$

from which we deduce that there exists $\varepsilon \in \{-1, 1\}$ such that

$$\varepsilon \left\|(K+\lambda)^{-1} K_x\right\| = \operatorname{Tr}(K+\lambda)^{-1} K_x) = \operatorname{Tr}(\Sigma+\lambda)^{-1}\varphi(x)\otimes\varphi(x)) = \left\|(\Sigma+\lambda)^{-1/2}\varphi(x)\right\|.$$

Necessarily $\varepsilon = 1$ since the right term is positive. Taking the essential supremum ends the proof. □

In view of Lemma 8, the last inequality in (24) follows from

$$\mathcal{N}_2(\lambda) = \mathbb{E}_X\left[\operatorname{Tr}((\Sigma+\lambda)^{-1}\varphi(X)\otimes\varphi(X)\right] \leq \operatorname*{ess\,sup}_X \operatorname{Tr}((\Sigma+\lambda)^{-1}\varphi(X)\otimes\varphi(X)) = \mathcal{N}_\infty(\lambda).$$

A5.2 Bounding the variance with interpolation inequalities

The following is a reinterpretation of Proposition 29 of Cabannes et al. [5].

Proposition 9 (Capacity condition)**.** *When $K^p(L^2(\rho_{\mathcal{X}}))$ is continuously embedded in $L^\infty(\rho_{\mathcal{X}})$ with $p \leq 1/2$, there exists a constant c such that*

$$\mathcal{N}_\infty(\lambda) \leq c\lambda^{-2p}. \tag{25}$$

Proof. The continuous embedding means that there exists a constant c such that, for any $\lambda \geq 0$,

$$\|K^p f\|_\infty \leq c \|f\|_2.$$

Stated in $\mathcal{H}$, we get

$$\|S\theta\|_\infty \leq c \left\|K^{-p} S\theta\right\|_2 = c \left\|\Sigma^{1/2-p}\theta\right\|$$

for every $\theta \in \mathcal{H}$. In other terms,

$$\operatorname*{ess\,sup}_x \left|\varphi(x)^\top\theta\right| \leq c \left\|\Sigma^{1/2-p}\theta\right\|.$$

Let us denote by (λ_i, θ_i) the eigenvalue decomposition of Σ. Then

$$\operatorname*{ess\,sup}_x (\varphi(x)^\top\theta)^2 \leq c^2 \left\|\Sigma^{1/2-p}\theta\right\|^2 = c^2 \sum_{i\in\mathbb{N}} \lambda_i^{1-2p}(\theta_i^\top\theta)^2.$$

When considering $\theta = \theta_i$, this leads to

$$\left|\theta^\top\varphi(x)\right| \leq c\lambda_i^{1/2-p}.$$

Therefore,

$$\begin{aligned}
\mathcal{N}_\infty^{1/2}(\lambda) &= \sup_x \left\|(\Sigma+\lambda)^{-1/2}\varphi(x)\right\| = \sup_x \sup_{\theta;\|\theta\|\leq 1} \theta^\top \Sigma_\lambda^{-1/2}\varphi(x) \\
&= \sup_x \sup_{\theta;\|\theta\|\leq 1} \sum_{i\in\mathbb{N}} \frac{\theta^\top\theta_i\theta_i^\top\varphi(x)}{(\lambda+\lambda_i)^{1/2}} \leq c \sup_{a;\sum a_i^2\leq 1} \sum_{i\in\mathbb{N}} \frac{a_i\lambda_i^{1/2-p}}{(\lambda+\lambda_i)^{1/2}} \\
&= c \sup_{i\in\mathbb{N}} \frac{\lambda_i^{1/2-p}}{(\lambda+\lambda_i)^{1/2}} = c \sup_{t\in\operatorname{spec}(K)} \frac{t^{1/2-p}}{(\lambda+t)^{1/2}} \leq c \sup_{t\geq 0} \frac{t^{1/2-p}}{(\lambda+t)^{1/2}} \\
&= c(2p)^{-p}(1-2p)^{1/2-p}\lambda^{-p},
\end{aligned}$$

where the last equality follows from basic calculus. □

A5.3 Bounding the bias with source conditions

We now focus our attention on the bias term.

Proposition 10 (Capacity condition)**.** *When $f^* \in K^q(L^2(\rho_{\mathcal{X}})$ with $q \leq 1$, there exists a constant c such that, for any $\lambda \geq 0$,*

$$\mathcal{B}(\lambda) = \lambda^2 \mathcal{S}(\lambda) \leq c\lambda^{2q}. \tag{26}$$

Proof. The proof is straight-forward. If $f^* = K^q g$ with $g \in L^2$, then

$$\begin{aligned}\mathcal{S}(\lambda) &= \left\|(K+\lambda)^{-1} f^*\right\|_2 = \left\|(K+\lambda)^{-1} K^q g\right\|_2 \\ &= \left\|(K+\lambda)^{q-1}\right\|_2 \left\|(K+\lambda)^{-q} K^q\right\|_2 \|g\|_2 \leq \lambda^{q-1} \left\|K^{-p} f^*\right\|_2 .\end{aligned}$$

Squaring this term and multiplying it by λ^2 leads to the result. □

A5.4 Classical interpolation inequalities

We begin with a simple proposition.

Proposition 11. *When the function $x \to k(x,x)$ is bounded, the RKHS associated with k verifies*

$$\mathcal{N}_\infty(\lambda) = O(\lambda^{-1}).$$

Proof. This follows from the fact that $K^{1/2}(L^2) \hookrightarrow L^\infty$ as soon as φ is bounded since, for any $f = \varphi(\cdot)^\top \theta \in \mathcal{F} = S\mathcal{H} = K^{1/2}(L^2)$,

$$|f(x)| = \left|\varphi(x)^\top \theta\right| \leq \|\varphi\|_\infty \|\theta\| = \|\varphi\|_\infty \|f\|_{\mathcal{F}} = \|\varphi\|_\infty \left\|K^{-1/2} f\right\|_2 .$$

The previous characterization of $\mathcal{N}_\infty$ leads to the claim. □

We now turn ourselves to more complicated interpolations, and offer an informal proposition of facts that are well-known in approximation theory [28, 9].

Proposition 12 (Informal source condition)**.** *When $\mathcal{F} = H^\beta$ and $f^* \in H^\alpha$, it holds*

$$f^* \in K^{\alpha/2\beta}(L^2(\rho_{\mathcal{X}})).$$

Proof. In essence, as explained in Appendix B, K takes a function in $L^2(\rho_{\mathcal{X}})$ and multiply its Fourier transform by $\hat{q}(\omega)^{-1} = (1 + \|\omega\|^2)^\beta$, with q defining the Matérn kernel, making it 2β-smooth in the Sobolev sense. In harmonic settings where the Fourier functions diagonalize K and $\hat{q}(\omega)$ parameterizes the spectrum of K, the fractional operator K^p can be seen as multiplying the Fourier transform of f by $\hat{q}(\omega)^{-p}$, making it $2p\beta$-smooth. This fact can be extended beyond those harmonic settings, notably with interpolation inequalities as the one used for the last part of Proposition 3. On the opposite direction, any α-smooth function can be multiplied by $q(\omega)^{\alpha/2\beta}$ in Fourier while staying in $L^2(\rho_{\mathcal{X}})$, so that, if f^* is α-smooth, it belongs to $K^{\alpha/2\beta}$. □

Proposition 13 (Informal interpolation inequality)**.** *When $\mathcal{F} = H^\beta$ is the space of α-Sobolev functions,*

$$K^{d/2\beta}(L^2) \hookrightarrow L^\infty.$$

Proof. Note that $\mathcal{F} = S\mathcal{H} = K^{1/2}(L^2$. We have seen informally in the proof of the previous lemma how $K^p(L^2 \subset H^{2p\beta}$. Now, let us recall the Sobolev embedding theorems [1]. Under mild assumptions on $\rho_{\mathcal{X}}$, for $k, r, l, s > 0$

$$W^{k,r}(\rho_{\mathcal{X}}) \hookrightarrow W^{l,s}(\rho_{\mathcal{X}}), \qquad \text{as long as} \qquad \frac{1}{r} - \frac{k}{d} \leq \frac{1}{s} - \frac{l}{d}.$$

We want to use it with $k = 2p\beta$, $r = 2$, $l = 0$ and $s = +\infty$, which leads to $p = 4\beta/d$. □

These results partially explained Table 1, which we derive formally in Appendix B.

B Translation-invariant kernels and Fourier analysis

Let us recall basic facts about kernel methods and Fourier analysis, before providing proofs to Propositions 2 and 3.

B.1 Stylized analysis on the torus

When k is a translation-invariant kernel, i.e. $k(x, x') = q(x - x')$, the integral operator K is a convolution against q. Let us expand on the friendly case provided by the torus $\mathcal{X} = \mathbb{T}^d := \mathbb{R}^d/\mathbb{Z}^d = [0, 1]^d/\sim$, where $\sim$ is the relation identifying opposite faces of the hypercube, and $\rho_{\mathcal{X}}$ the uniform distribution. On the torus, a translation invariant kernel is defined through q being a one-periodic function on $\mathbb{R}^d$. The integral operator $K : L^2(\mathcal{X}, \mathrm{d}x) \to L^2(\mathcal{X}, \mathrm{d}x)$ is the convolution

$$Kf(x) = \int_{[0,1]^d} k(x, x')f(x')\,\mathrm{d}x' = \int_{[0,1]^d} q(x' - x)f(x')\,\mathrm{d}x' = q * f(x).$$

For $m \in \mathbb{Z}^d$, define the Fourier function $f_m : x \mapsto \exp(2i\pi\,\langle m, x\rangle)$. One can check that the f_m's form an orthonormal family that diagonalizes K with[3]

$$Kf_m = \widehat{q}_m f_m, \qquad \text{where} \qquad \widehat{q}_m = \int_{[0,1]^d} q(x)\exp(2i\pi\,\langle x, m\rangle)\,\mathrm{d}x.$$

Hence, using Pythagoras theorem, we can define the norm on $\mathcal{F}$ through its action on Fourier coefficients as

$$\|f\|_{\mathcal{F}}^2 = \left\langle f, K^{-1}f\right\rangle_{L^2(\rho_{\mathcal{X}})} = \sum_{m\in\mathbb{Z}^d} \widehat{q}_m^{-1}|\widehat{f}_m|^2 = \int_{\mathbb{R}^d} \widehat{q}(\omega)^{-1}|\widehat{f}(\omega)|^2 \#(\mathrm{d}\omega),$$

where $\widehat{f}_m = \langle f, f_m\rangle_{L^2(\rho_{\mathcal{X}})}$, and $\#$ is the counting measure on $\mathbb{Z}^d \subset \mathbb{R}^d$.

B.1.1 First part of the proof of Proposition 3

Since K is diagonalized in Fourier, we compute the size of $\mathcal{F}$ for $a \in \{1, 2\}$ with

$$\mathcal{N}_a(\lambda) = \mathrm{Tr}\left(K^a(K + \lambda)^{-a}\right) = \sum_{m\in\mathbb{Z}^d} \frac{\widehat{q}_m^a}{(\widehat{q}_m + \lambda)^a}.$$

For kernels whose scales are explicitly defined through $q_\sigma = q(x/\sigma)$, we have $\widehat{q}_\sigma(\omega) = \sigma^d\widehat{q}(\sigma\omega)$, which leads to (12).

Similarly, the bound on the bias term follows from Fourier analysis by

$$\mathcal{B}(\lambda) = \left\|K(K + \lambda)^{-1}f - f\right\|_{L^2(\rho_{\mathcal{X}})}^2 = \sum_{m\in\mathbb{Z}^d} \left|\widehat{f}_m\right| \left(\frac{\widehat{q}_m}{\widehat{q}_m + \lambda} - 1\right)^2 = \sum_{m\in\mathbb{Z}^d} \left|\widehat{f}_m\right| \left(\frac{\lambda}{\widehat{q}_m + \lambda}\right)^2.$$

which provides (13).

B.1.2 Second part of the proof of Proposition 3

When ρ is a distribution that is absolutely continuous with respect to the Lebesgue measure and whose density is bounded from above, we get

$$K \preceq \rho_\infty K_{\mathrm{d}x}, \qquad \text{with} \qquad \rho_\infty = \left\|\frac{\mathrm{d}\rho_{\mathcal{X}}}{\mathrm{d}x}\right\|_{L^\infty(\rho_{\mathcal{X}})},$$

where $K_{\mathrm{d}x}$ is the integral operator associated to the kernel k on $L^2(\mathrm{d}x)$ endowed with the Lebesgue measure. Using the fact the effective dimension is an increasing function of the eigenvalues (since

[3] Indeed, the Fourier transform of a function $f \in L^2(\mathbb{T}^d)$ can be defined as the mapping from $\mathbb{N}$ to $(\langle f_i, f\rangle)_{i\in\mathbb{N}}$ where (f_i) is a basis that diagonalizes all convolution operators (note that this definition is possible because $\rho_{\mathcal{X}}$ is uniform on the torus).

$x \mapsto x/(x+1)$ is increasing), and that eigenvalues are increasing with the Loewner order, this leads to, with the Fourier transform on $L^2(\mathrm{d}x)$,

$$\mathcal{N}(\lambda) \leq \rho_\infty \operatorname{Tr}\left((K_{\mathrm{d}x} + \lambda)^{-1} K_{\mathrm{d}x}\right) = \rho_\infty \int_{\mathbb{R}^d} \frac{\widehat{q}(\omega)}{\lambda + \widehat{q}(\omega)} \,\mathrm{d}\omega.$$

Note that those derivations are written informally (since K and $K_{\mathrm{d}x}$ do not act on the same space), but could be made formal with the isomorphic covariance operators on $\mathcal{H}$, plus some technicalities to make sure $\Sigma_{\mathrm{d}x}$ is well defined (assuming $\varphi(X)$ as a fourth-order moment against Lebesgue, or approaching $K_{\mathrm{d}x}$ within its action on compact space of $\mathcal{X}$ where it is bounded, before taking the limit of $\mathcal{N}_{\mathrm{d}x}(\lambda)$).

For the bias term, using the fact that $L^2(\rho_{\mathcal{X}})$ is continuously embedded in $L^2(\mathrm{d}x)$ and the isometry between the spatial and the Fourier domain

$$\|f - f^*\|_{L^2(\rho_{\mathcal{X}})} \leq \rho_\infty^{1/2} \|f - f^*\|_{L^2(\mathrm{d}x)} = \rho_\infty^{1/2} \|\widehat{f} - \widehat{f^*}\|_{L^2(\mathrm{d}x)}.$$

Finally, it should be noted that the norm associated with $\mathcal{F}$ does not depend on the density of X, hence the formula can be written independently of $\rho_{\mathcal{X}}$. Indeed, under definition assumption, this formula can even be written with a measure of infinite mass. For example, when $\mathcal{X} = \mathbb{R}^d$, one can consider the Fourier transform associated with $L^2(\mathrm{d}x)$ endowed with the Lebesgue measure, and get

$$\|f\|_{\mathcal{F}}^2 = \int_{\omega \in \mathcal{X}^d} \widehat{q}(\omega)^{-1} |\widehat{f}(\omega)|^2 \,\mathrm{d}\omega, \qquad \text{where} \qquad \widehat{q}(\omega) = \int_{\mathbb{R}^d} q(x) \exp(-2i\pi \langle x, \omega \rangle) \,\mathrm{d}x, \quad (27)$$

although some care is needed to deal with the continuous version of the spectral theorem (the set of eigenvalues being non-countable). From there the same derivations as for Proposition 3 lead to the desired result.

B.2 Sobolev spaces

Recall the action of derivation on the Fourier transform, for $m \in \mathbb{N}^d$, $|m| := \|m\|_1$, and $f \in L^2(\mathrm{d}x)$,

$$\widehat{\frac{\partial^{|m|} f}{\prod_{i \in [d]} \partial^{m_i} x_i}}(\omega) = (2i\pi)^{|m|} \prod_{i \in [d]} \omega_i^{m_i} \widehat{f}(\omega).$$

It characterizes the pseudo-norm

$$\|f\|_m^2 = \int_{\mathbb{R}^d} \left\| \frac{\partial^{|m|} f(x)}{\prod_{i \in [d]} \partial^{m_i} x_i} \right\|^2 \mathrm{d}x = (2\pi)^{2|m|} \int_{\mathbb{R}^d} \prod_{i \in [d]} \omega_i^{2m_i} \left| \widehat{f}(\omega) \right|^2 \mathrm{d}\omega.$$

This pseudo-norm is associated with the translation-invariant kernel such that $\widehat{q}(\omega) = \prod_{i \in [d]} \omega_i^{-2m_i}$ as per (27). Note that q is well defined when $\widehat{q}$ belongs to $L^1(\mathrm{d}x)$ (Bochner's theorem), that is $|m| > d$. Those observations are usual to deduce that the so-call Matérn kernel, which are defined from $\widehat{q}(\omega) \propto (1 + \|\omega\|_2^2)^{-\beta}$, corresponds to $\|\cdot\|_{\mathcal{F}}$ the Sobolev space $H^\beta(\mathrm{d}x)$ endowed with the norm

$$\|f\|_{H^\beta}^2 = \sum_{m; |m| \leq \beta} \|f\|_m^2 .$$

It follows from Bochner theorem that H^β is a reproducing kernel Hilbert space if only if $2\beta > d$. Remarkably, the exponential kernel corresponds to the Matérn kernel with $\beta = (d+1)/2$ [22]. For the Gaussian kernel, $\widehat{q}(\omega) = \pi^{-d/2} \exp(-\pi^2 \|\omega\|^2)$, and the associated function class $\mathcal{F}$ are analytic (Paley-Wiener theorem).

B.2.1 Functional sizes

Let us now express the capacity and bias bound within Sobolev spaces.

Proposition 14 (Sobolev capacity)**.** *When $\widehat{q}(\omega) = (1 + \|\omega\|^2)^{-\beta}$ for $\beta > d$, when $\lambda\sigma^{-d}$ is bounded, and ρ has a bounded density*

$$\mathcal{N}_1(\lambda, \sigma) \leq \frac{2\beta\rho_\infty \pi^{(d+1)/2}}{\Gamma((d-1)/2)} \lambda^{-d/2\beta} \sigma^{-d(2\beta - d)/2\beta}.$$

Moreover when $\mathcal{X} = \mathbb{T}^d$ and $\rho_{\mathcal{X}}$ is uniform, we get

$$\mathcal{N}_2(\lambda, \sigma = 1) \geq \max_{l \in [d]} \frac{l\pi^{(l+1)/2}}{2^{l+1}\Gamma((l-1)/2)} \lambda^{-l/2\beta}.$$

Proof. In this setting, Proposition 3 leads to

$$\begin{aligned}
\int_{\mathbb{R}^d} \frac{1}{1 + \lambda \widehat{q_\sigma}(\omega)^{-1}} \,\mathrm{d}\omega &= \int_{\mathbb{R}^d} \frac{1}{1 + \lambda \sigma^{-d} \widehat{q}(\sigma\omega)^{-1}} \,\mathrm{d}\omega = \int_{\mathbb{R}^d} \frac{1}{1 + \lambda \sigma^{-d}(1 + \sigma^2 \left\|\omega\right\|^2)^\beta} \,\mathrm{d}\omega \\
&= \operatorname{surf}(\mathcal{S}^{d+1}) \int_{\mathbb{R}_+} \frac{r^{d-1} \,\mathrm{d}r}{1 + \lambda \sigma^{-d}(1 + \sigma^2 r^2)^\beta} \\
&= 2\pi \operatorname{vol}(\mathcal{S}^d) \int_{\lambda^{1/\beta}\sigma^{-d/\beta}}^\infty \frac{(u - \lambda^{1/\beta}\sigma^{-d/\beta})^{d/2-1} \,\mathrm{d}u}{\lambda^{d/2\beta}\sigma^{d-d^2/2\beta}(1 + u^\beta)} \\
&= 2\pi \operatorname{vol}(\mathcal{S}^d) \lambda^{-d/2\beta}\sigma^{d(d-2\beta)/2\beta} \int_{\mathbb{R}_+} \frac{x^{d/2-1} \,\mathrm{d}x}{1 + (x + \lambda^{1/\beta}\sigma^{-d/\beta})^\beta} \\
&\leq 2\pi\beta \operatorname{vol}(\mathcal{S}^d) \lambda^{-d/2\beta}\sigma^{d(d-2\beta)/2\beta}.
\end{aligned}$$

Where we have used the fact that

$$\begin{aligned}
\int_0^\infty \frac{x^{d/2-1} \,\mathrm{d}x}{1 + (x + \lambda^{1/\beta}\sigma^{-d/\beta})^\beta} &\leq \int_0^\infty \frac{x^{d/2-1} \,\mathrm{d}x}{\max(1, \max(x^\beta, \lambda\sigma^{-d}))} \\
&\leq \int_0^1 x^{d/2-1} \,\mathrm{d}x + \int_1^\infty \frac{x^{d/2-1} \,\mathrm{d}x}{x^\beta} = d/2 - (d/2 - \beta) = \beta,
\end{aligned}$$

which is true as long as $\beta > d/2$ to ensure proper convergence of the last integral.

For the part on the torus, in order to get a sharp learning limit, we need to be slightly more precise. In particular, we want to relate the discrete Fourier transform integral of Proposition 3 with the continuous one through series-integral comparison, and to get a lower bound on the last integral. We will fix $\sigma = 1$ for simplicity. A simple cut of $\mathbb{R}^d$ into unit cubes, and the fact that our integrand is decreasing leads to

$$\sum_{m \in \mathcal{Z}^d} \mathbf{1}_{0 \notin m} \frac{\widehat{q}_m^2}{(\widehat{q}_m + \lambda)^2} \leq \int \frac{\widehat{q}(\omega)^2}{(\widehat{q}(\omega) + \lambda)^2} \,\mathrm{d}\omega \leq \sum_{m \in \mathcal{Z}^d} 2^{\#\{i \in [d] \,|\, m_i = 0\}} \frac{\widehat{q}_m^2}{(\widehat{q}_m + \lambda)^2}.$$

We simplify it as

$$\mathcal{N}_2(\sigma, \lambda) \geq 2^{-d} \int \frac{\widehat{q}(\omega)^2}{(\widehat{q}(\omega) + \lambda)^2} \,\mathrm{d}\omega.$$

We compute the integral with the same techniques as before

$$\begin{aligned}
\int \frac{\widehat{q}(\omega)^2}{(\widehat{q}(\omega) + \lambda)^2} \,\mathrm{d}\omega &= \int \frac{1}{(1 + \widehat{q}(\omega)^{-1}\lambda)^2} \,\mathrm{d}\omega = \int \frac{1}{(1 + (1 + \left\|\omega\right\|^2)^\beta \lambda)^2} \,\mathrm{d}\omega \\
&= 2\pi \operatorname{vol}(\mathcal{S}^d) \int \frac{x^{d-1}}{(1 + (1 + x^2)^\beta \lambda)^2} \,\mathrm{d}x \\
&= 2\pi \operatorname{vol}(\mathcal{S}^d) \lambda^{-d/2\beta} \int \frac{x^{d-1}}{(1 + (\lambda^{1/\beta} + x^2)^\beta)^2} \,\mathrm{d}x \\
&\geq 2\pi \operatorname{vol}(\mathcal{S}^d) \lambda^{-d/2\beta} \int \frac{x^{d-1}}{4 \max(1, 4^\beta \max(\lambda^2, x^{4\beta}))} \,\mathrm{d}x \\
&= 2^{-1}\pi \operatorname{vol}(\mathcal{S}^d) \lambda^{-d/2\beta} \left(\int_0^1 x^{d-1} \,\mathrm{d}x + 4^{-\beta} \int_1^\infty \frac{x^{d-1}}{x^{4\beta}} \,\mathrm{d}x \right) \\
&= 2^{-1}\pi \operatorname{vol}(\mathcal{S}^d) \lambda^{-d/2\beta} (d + 4^{-\beta}(4\beta - d)) \\
&\geq 2^{-1}\pi \operatorname{vol}(\mathcal{S}^d) \lambda^{-d/2\beta} d.
\end{aligned}$$

It should be noted that this last bound is somewhat too lax, as it tends to zero when the dimension increases. Since, $\sum_{m \in \mathbb{Z}^d} a(\|m\|)$ for $a > 0$ strictly increasing with d, we deduce that this lower bound holds for any $k \leq d$. □

Proposition 15 (Gaussian capacity)**.** *When $\widehat{q}(\omega) = \exp(-\|\omega\|^2)$, and ρ has a bounded density*

$$\mathcal{N}_1(\lambda, \sigma) \leq \frac{\rho_\infty \pi^{(d-1)/2} d}{2\sigma^d} L(\lambda^{-1}\sigma^d),$$

where L is defined by Eq. (28)*. In particular, $L(x) \leq x$ when $x < 1$, and $L(x) \lesssim \log(x)^{d/2}$ when x gets large. Moreover when $\mathcal{X} = \mathbb{T}^d$ and $\rho_\mathcal{X}$ is uniform, we get*

$$\mathcal{N}_2(\lambda, \sigma) \geq \frac{\pi^{(d-1)/2}}{2^{d+1}\sigma^d} L(\lambda^{-1}\sigma^d),$$

Proof. With the Gaussian kernel, Proposition 3 leads to

$$\begin{aligned}
\int_{\mathbb{R}^d} \frac{1}{1 + \lambda\sigma^{-d}\exp(\sigma^2\|\omega\|^2)}\,\mathrm{d}\omega &= \mathrm{vol}(\mathbb{S}^d)\int_{\mathbb{R}_+} \frac{2x^{d-1}}{1+\lambda\sigma^{-d}\exp(\sigma^2 x^2)}\,\mathrm{d}x \\
&= \mathrm{vol}(\mathbb{S}^d)\sigma^{-d}\int_{\mathbb{R}_+} \frac{u^{d/2-1}}{1+\lambda\sigma^{-d}\exp(u)}\,\mathrm{d}u \\
&= \mathrm{vol}(\mathbb{S}^d)\Gamma(d/2)\frac{-\operatorname{Li}_{d/2}\left(-\sigma^d/\lambda\right)}{\sigma^d},
\end{aligned}$$

where Li is the polylogarithm function, hence the definition of L as

$$L(x) = -\operatorname{Li}_{d/2}(-x) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}x^k}{k^{d/2}} \tag{28}$$

We recognize an alternating sequence, whose terms amplitudes are decreasing as a function of $k \in \mathbb{N}$ when $x \leq 1$, which explains that $L(x)$ is smaller than the first term in this case. The expansion of the polylogarithm function in infinity leads to the upper bound when x goes to infinity.

When it comes to an lower bound, we can proceed as the precedent lemma with

$$\mathcal{N}_2(\sigma, \lambda) \geq 2^{-d}\int \frac{\mathrm{d}\omega}{(1+\lambda\widehat{q}(\omega)^{-1})^2} \geq 2^{-d}\int \frac{\mathrm{d}\omega}{1+\lambda\widehat{q}(\omega)^{-1}},$$

which corresponds to the integral computed for the upper bound. Once again, this also holds when d is replaced by any $l \in [d]$. □

B.2.2 Adherences

Let us now turn our attention to the bias, i.e. the adherence of functions in those spaces. For proof readability, we will assume that $\rho_\mathcal{X}$ has compact support. In this setting, $W^{\alpha,2}$ is continuously embedded in $C^{\alpha+d/2} = W^{\alpha+d/2,\infty}$, which can be defined as

$$C^{\alpha+d/2} = \left\{ f : \mathbb{R}^d \to R \;\middle|\; \|f\|_\alpha := \operatorname*{ess\,sup}_{\omega\in\mathbb{R}^d} \left|\widehat{f}(\omega)\right| (1+\|\omega\|)^{\alpha+d/2} < +\infty \right\}.$$

Proposition 16 (Adherence of H^α in H^β)**.** *When $\widehat{q}(\omega) = (1+\|\omega\|^2)^{-\beta}$ hence $\mathcal{F} = H^\beta$, if $\alpha > 2\beta$, for any $f^* \in H^\alpha(\rho_\mathcal{X})$, and λ small enough,*

$$\mathcal{B}(\sigma, \lambda) \leq \lambda^2 \left\|K^{-1}f\right\|^2_{L^2(\rho_X)}.$$

If $\alpha < \beta$, and $\rho_\mathcal{X}$ has a bounded density, then for any function $f^* \in C^{\alpha+d/2}$

$$\mathcal{B}(\sigma, \lambda) \leq \rho_\infty \frac{4\beta\pi^{(d+1)/2}\rho_\infty \|f\|^2_\alpha \beta}{(\beta^2-\alpha^2)\Gamma((d-1)/2)}\lambda^{\alpha/\beta}\sigma^{(2\beta-d)\alpha/\beta}. \tag{29}$$

Proof. The first part results from previous consideration on the source condition since we have the inclusion $f^* \in H^\alpha \subset H^{2\beta} = K(L^2(\rho_\mathcal{X}))$. The second part follows from an $L^1 - L^\infty$ Hölder

inequality

$$\begin{aligned}
\mathcal{B}(\sigma,\lambda) &\leq \rho_\infty \int_{\mathbb{R}^d} \frac{\left|\widehat{f}(\omega)\right|^2}{(\lambda^{-1}\sigma^d \widehat{q}(\sigma\omega)+1)^2}\,\mathrm{d}\omega \leq \rho_\infty \|f\|_\alpha^2 \int_{\mathbb{R}^d} \frac{(1+\|\omega\|^2)^{-(d/2+\alpha)}}{(\lambda^{-1}\sigma^d(1+\sigma^2\|\omega\|^2)^{-\beta}+1)^2}\,\mathrm{d}\omega \\
&= 2\pi \operatorname{vol}(\mathbb{S}^d)\rho_\infty \|f\|_\alpha^2 \int_{\mathbb{R}_+} \frac{2(1+x^2)^{-(d/2+\alpha)}x^{d-1}}{(\lambda^{-1}\sigma^d(1+\sigma^2x^2)^{-\beta}+1)^2}\,\mathrm{d}x \\
&= 2\pi \operatorname{vol}(\mathbb{S}^d)\rho_\infty \|f\|_\alpha^2 a^\alpha \sigma^{2\alpha} \int_a^\infty \frac{(u-a+a\sigma^2)^{-(d/2+\alpha)}(u-a)^{d/2-1}}{(u^{-\beta}+1)^2}\,\mathrm{d}u \\
&\leq \frac{4\beta\pi \operatorname{vol}(\mathbb{S}^d)\rho_\infty \|f\|_\alpha^2 \beta}{\beta^2-\alpha^2}\lambda^{\alpha/\beta}\sigma^{(2\beta-d)\alpha/\beta},
\end{aligned}$$

where a was set to be $\lambda^{1/\beta}\sigma^{-d/\beta}$, and the last integral can be computed with

$$\begin{aligned}
&\int_a^\infty \frac{(u-a+a\sigma^2)^{-(d/2+\alpha)}(u-a)^{d/2-1}}{(u^{-\beta}+1)^2}\,\mathrm{d}u \leq \int_0^\infty \frac{u^{d/2+\beta-1}}{(u+a(\sigma^2-1))^{d/2+\alpha}(1+u^\beta)^2}\,\mathrm{d}u \\
&\leq \int_0^\infty \frac{u^{d/2+\beta-1}}{(u+a(\sigma^2-1))^{d/2+\alpha}(1+u^\beta)^2}\,\mathrm{d}u \leq \int_0^1 \frac{u^{d/2+\beta-1}}{u^{d/2+\alpha}}\,\mathrm{d}u + \int_1^{+\infty} \frac{u^{d/2+\beta-1}}{u^{d/2+\alpha}u^2\beta}\,\mathrm{d}u \\
&= \frac{1}{\beta-\alpha}+\frac{1}{\beta+\alpha} = \frac{\beta}{\alpha(\beta-\alpha)}.
\end{aligned}$$

The volume of the sphere ends the proof. □

Proposition 17 (Adherence of H^α in Gaussian RKHS)**.** *When $\mathcal{F}$ is associated with the Gaussian kernel and $\rho_\mathcal{X}$ has a bounded density, for any $f^* \in C^{\alpha+d/2}$,*

$$\mathcal{B}(\sigma,\lambda) \leq \frac{2\pi^{(d+1)/2}\rho_\infty \|f\|_\alpha^2}{\Gamma((d-1)/2)}\left(\frac{1}{\sigma^{2d+4\alpha}}+\frac{2\log(2)^{-(d/2+2\alpha)}}{d+2\alpha}\right)\sigma^{2\alpha}\log(\lambda^{-1}\sigma^d)^{-\alpha}.$$

Proof. We follow the same path as the adherence of H^α in H^β,

$$\begin{aligned}
\mathcal{B}(\sigma,\lambda) &\leq \rho_\infty \|f\|_\alpha^2 \int_{\mathbb{R}^d} \frac{(1+\|\omega\|^2)^{-(d/2+\alpha)}}{(\lambda^{-1}\sigma^d\widehat{q}(\sigma\omega)+1)^2}\,\mathrm{d}\omega \\
&= \rho_\infty \|f\|_\alpha^2 \int_{\mathbb{R}^d} \frac{(1+\|\omega\|^2)^{-(d/2+\alpha)}}{(\lambda^{-1}\sigma^d\exp(-\sigma^2\|\omega\|^2)+1)^2}\,\mathrm{d}\omega \\
&= 2\pi \operatorname{vol}(\mathbb{S}^d)\rho_\infty \|f\|_\alpha^2 \int_{\mathbb{R}_+} \frac{(1+x^2)^{-(d/2+\alpha)}x^{d-1}}{(\lambda^{-1}\sigma^d\exp(-\sigma^2x^2)+1)^2}\,\mathrm{d}x \\
&= 2\pi \operatorname{vol}(\mathbb{S}^d)\rho_\infty \|f\|_\alpha^2 \sigma^{2\alpha} \int_1^{+\infty} \frac{\log(x)^{d/2-1}}{(\sigma^2+\log(x))^{d+2\alpha}(a+x)}\,\mathrm{d}x \\
&\leq 2\pi \operatorname{vol}(\mathbb{S}^d)\rho_\infty \|f\|_\alpha^2 \sigma^{2\alpha}\log(\lambda^{-1}\sigma^d)^{-\alpha}\left(\frac{1}{\sigma^{2d+4\alpha}}+\frac{\log(2)^{-\alpha}}{\alpha}\right).
\end{aligned}$$

where the integral was computed with

$$\begin{aligned}
\int_1^{+\infty} \frac{\log(x)^{d/2-1}}{(\sigma^2+\log(x))^{d+2\alpha}(a+x)}\,\mathrm{d}x &\leq \int_1^e \frac{\log(x)}{\sigma^{2d+4\alpha}}\,\mathrm{d}x + \int_e^{+\infty} \frac{1}{(\log(x))^{d/2+2\alpha+1}x}\,\mathrm{d}x \\
&= \frac{1}{\sigma^{2d+4\alpha}}+\frac{\log(2)^{-(d/2+2\alpha)}}{d/2+2\alpha}.
\end{aligned}$$

The same type of derivations can be made for the bias in the lower bound. □

Note that the proofs also work when the L^1-L^∞ Hölder inequality is replaced by a $L^\infty-L^1$ one, showcasing the norm of f in H^α instead of in $C^{\alpha+d/2}$. We refer to Bach [4] for details.

Remark 18 (Blessing of dimensionality)**.** *It should be noted that all our integral calculations show a constant $2\pi\rho_\infty \operatorname{vol}(\mathbb{S}^d)$ which will be present in front of the excess risk. As d increases, this constant goes to zero faster than any exponential profile. To see that, note how the volume of the d-sphere is always smaller than twice the one of its inscribed hypercube, whose volume is $d^{-d/2}$. We do not have clear intuition to understand this behavior at time of writing.*

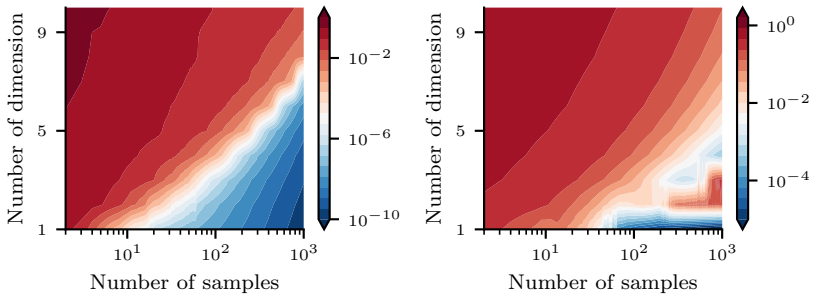

Figure 7: (Right) Noise-free convergence rates for $f^*(x) = x_1^5$ with $k(x,y) \propto (1 + x^\top y)^5$. We observed similar deterioration of convergence rates as a function of the dimension d as on Figure 2. The fact that the error is not exactly zero when $d = 1$ and $n \geq 5$ is due to a small regularization added in our algorithm to avoid running into computational issues when inverting a matrix online. (Left) Convergence rates for $f^*(x) = \cos(4\pi x_1)$ on the torus $\mathcal{X} = \mathbb{T}^d$ with the (periodic) exponential kernel $k(x,y) = q(-100\,\|x-y\|^2/d)$. We observe similar behavior as on Figure 2, the picture being worse because the kernel weights all frequencies in the Fourier domain, and not only the first $\binom{d+5}{5}$ ones.

C Experimental details

C.1 Online solving of problems of increasing size

In order to solve a big number of least-squares problems with increasing numbers of samples, one can use recursive matrix inversion. When $A \in \mathbb{R}^{n\times n}$, $x \in \mathbb{R}^n$ and $b \in \mathbb{R}$, one can check that

$$\begin{pmatrix} A & x \\ x^\top & b \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + cyy^\top & -cy \\ -cy^\top & c \end{pmatrix}$$

where

$$c = \frac{1}{b - x^\top y}, \qquad \text{and} \qquad y = A^{-1}x.$$

This allows efficient computation of matrix inversion online as the number of samples is increasing.

C.2 Example of convergence rates for "sparse" functions

Polynomial estimation. Consider the target function $f^*(x) = 63x_1^5 - 70x_1^3 + 15x_1$, learned with the polynomial kernel $k(x,y) = (1 + x^\top y)^d$ with $\mathcal{X} = [0,1]^d$ and ρ uniform, in the interpolation regime $\lambda = 0$. In this setting, $\mathcal{F}$ is exactly the space of polynomials of degree less than 5, which follows from the fact that k can be rewritten through a vector φ that enumerates monomials

$$(1 + x^\top y)^d = \sum_{i\in[0,d]} \binom{n}{i} \sum_{(i_j)_j;\sum_j i_j = i} \binom{i}{(i_j)_j} \prod x_j^{i_j} y_j^{i_j} = \varphi(x)^\top \varphi(y).$$

As a consequence, f^* belongs to $\mathcal{F}$, and we expect the bias term to be zero. Yet, we expect the variance, hence the generalization error, to behave in $\varepsilon^2 \dim \mathcal{F}/n$ where ε corresponds to some notion of variability between labels. The dimension of dimensionality of the class of functions made by polynomials of degree q verifies $\mathcal{F} = \binom{q+d}{d}$. In particular, in dimension $d = 100$, a polynomial of degree at most $q = 5$ can have up to one hundred million coefficients, meaning that, one need about one hundred million observations to enter the high-sample regime and expect an excess risk $\mathcal{R}(f_n) - \mathcal{R}(f^*)$ of order ε^2 as per Theorem 2. While one could fit a smaller polynomial of degree four instead of five, since f^* is actually orthogonal to all polynomials of lower degree, there is no hope to get better rates with such a smaller functional space. On Figure 2, the target function is $f^*(x) = x_1^5$, and the polynomial kernel is normalized as

$$k(x,y) = \left(\frac{1 + x^\top y}{1 + d}\right)^5,$$

to avoid running into computational issues. The noise level ε is set to 10^{-2}, and the lower bound in $\varepsilon^2 \dim \mathcal{F}/n$ is plotted on Figure 2. In practice, this lower bound describes well the learning dynamic

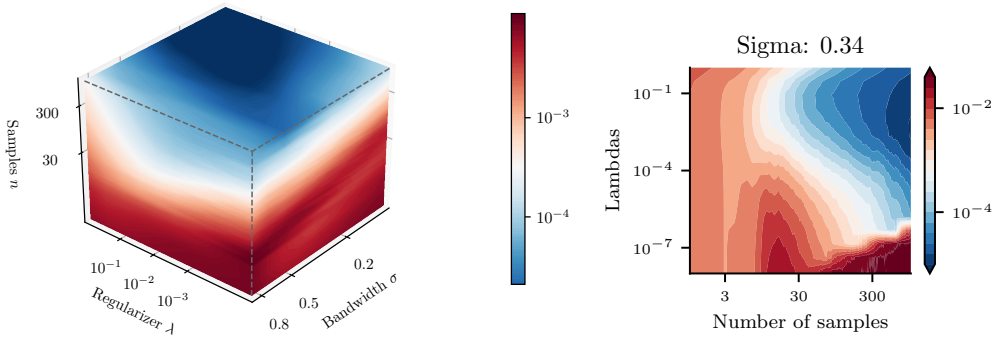

Figure 8: Excess risk when the target at the top is taken as $f^*(x) = \exp(-\max(x^2, C)) - \exp(C)$ with $C = 1/4$, and $x \in \mathbb{R}$ with unit Gaussian distribution. Note that the learning of the smooth part is more efficient when the regularizer is big, which forces the reconstruction to be smooth.

when the number of samples is high compared to the effective dimension of the space of functions considered.

Same example in Fourier. To transpose the previous example in the Fourier domain, one can consider $f^*(x) = \cos(\omega_0 x_1)$ together with some translation-invariant kernel. We illustrate it on Figure 7. The deterioration of the rates with respect to dimension can be understood precisely. In harmonic setting, such as on the torus with uniform measure, one can consider f^* as an eigenfunction of the integral operator K, associated with the eigenvalue λ_{ω_0}. The lower bound on the bias is given by

$$\mathcal{B} = \frac{\lambda^2}{(\lambda + \lambda_{\omega_0})^2}.$$

When k is translation invariant with $q(0)$, we get that

$$\operatorname{Tr} K = \sum_{\omega \in \mathbb{Z}^d} \lambda_\omega = \operatorname{Tr}\left(\mathbb{E}[\varphi(X)\varphi(X)^\top]\right) = \mathbb{E}[\varphi(X)^\top \varphi(X)] = \mathbb{E}[k(X, X)] = q(0).$$

When $q(0)$ does not depend on d, this quantity is constant. On the other hand, we expect the λ_ω to decrease with $\|\omega\|$, and because the number of frequencies below $\|\omega_0\|$ grows exponentially with the dimension, in order to keep this sum constant, λ_{ω_0} has to decrease exponentially fast with the dimension, hence the bias increase exponentially fast with the dimension.

Example of "wrongfully" arbitrarily fast convergence rates. To further emphasis on the importance of constants and transitory regimes, let us discuss an even simpler example. Assume that one wants to learn a polynomial of a unknown degree $s \in \mathbb{N}$ in a noiseless setting; or equivalently, learn an analytical function such that $f^{(s+1)} = 0$ for an unknown $s \in \mathbb{N}$. This polynomial can be learned exactly when provided with as many points as there are unknown coefficients in the polynomials. Meaning that the generalization error will almost surely go zero when provided enough points, as a consequence

$$\forall\, h : \mathbb{N} \to \mathbb{R}, \qquad \mathcal{E}(f_n) \leq O(h(n)).$$

where f_n is defined in (5) with $\mathcal{F}$ the space of polynomials of any degree. In other terms, we are able to prove *arbitrarily fast convergence rates*. Yet those convergence rates hide constants that are the real quantities governing convergence behaviors of any learning procedure. Figure 7 shows how the number of coefficients in a Taylor expansion of order s is once again the right quantity to look at.

C.3 Different convergence rates profiles

Figure 3 was computed with the Gaussian kernel on either the torus or $\mathbb{R}$. One hundred runs were launched and averaged to get a meaningful estimate of the expected excess risks. Convergence rates were computed for different hyperparameters, and the best set of hyperparameters (changing with respect to the number of samples but constant over run) was taken to show the best achievable convergence rates.

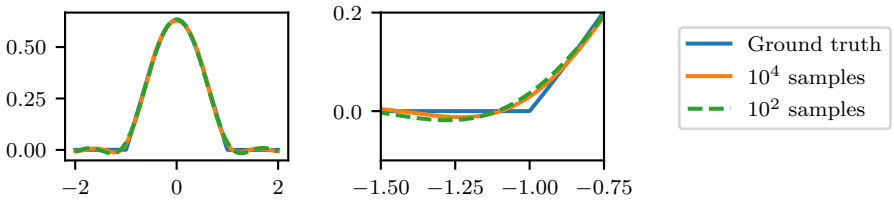

Figure 9: Example of a smooth target function with a C^1-singularity whose estimation is expected to showcase convergence rates that decrease fast first then slowly. The x-axis represent the input space $\mathcal{X} = [-2, 2]$, the y-axis represent the output space $f^*(x)$ and $f_{n,\lambda}$ for $n = 10^2$ and $n = 10^4$. The first fast decrease of excess risk corresponds to the easy estimation of the coarse details of the function, while the then slow decrease corresponds to the precise estimation of the C^1-singularity. The target function $x \to \exp(-\max(x^2, 1)) - \exp(-1)$ is represented in blue, while the estimation with 10^4 samples is represented in orange, and the one with 10^2 samples is represented in dashed green. The left picture zooms in on the estimation of the singularity, we see that the increase from 10^2 to 10^4 does not lead to a much better estimate.

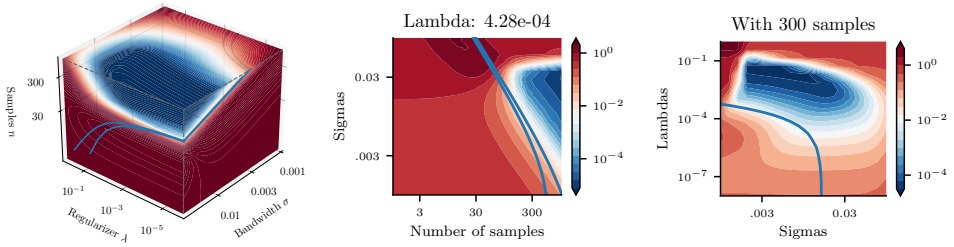

Figure 10: Excess of risk when the target at the top is taken as $f^*(x) = \cos(2\pi\omega x)$ with $\omega = 20$, and $x \in \mathbb{S} = \mathbb{R}/[-1, 1]$ uniform on the circle. Observe how the risk first stalls, before learning the function quite fast. The two graphs $\{(n, \mathcal{N}_a(\lambda, \sigma)\}$ for $a \in \{1, 2\}$ are plotted with the blue lines.

Fast then slow profile. Let us focus on the example provided by $f^* : x \to \exp(-\max(\|x\|^2, C)) - \exp(-C)$. Note that for $q_\sigma = \exp(-x^2/\sigma)$, the convolution $q_\sigma * f$ for a large σ will not modify f much, while making it analytical. This follows from Fourier analysis. If f is integrable, its Fourier transform is bounded. Since a convolution corresponds to a product in Fourier, and since the Fourier transform q_σ decays exponentially fast, so does $f * q_\sigma$, which implies its analytical property. As a consequence, all functions are close to analytical functions whose approximation should exhibit convergence rates in n^{-1}. In particular for f^* defined as before for a big C, one does not have enough observations, one will not be able to distinguish between f^* and $q_\sigma * f^*$, and as the number of sample first increase, one will learn quite fast a smooth version of f^*. After a certain number of samples, the learning will stall until enough points are provided to distinguish between f^* and its smoothing, and learn the C^1-singularity of the former. Figure 8 illustrates this observation. Note that similar reasoning could be made for any RKHS that is dense in L^2.

Slow then fast profile. The slow then fast profile was computed with $\mathcal{X} = \mathbb{R}/[-1, 1]$ being the sphere, $\rho_\mathcal{X}$ being uniform and $f^* : x \to \cos(2\pi\omega x)$ with $\omega = 20$. One hundred runs were launched and averaged to get an estimate of the excess risk of the estimator in (6) for different values of σ and λ. Again, the best results for different sample sizes were reported to get an estimate of convergence rates on Figure 3. A log-log-log-log plot of the results is provided by Figure 10.

C.4 Exploring the low-sample regime

Looking at the population weights first. When given access to the knowledge of the full distribution ρ, the estimator in (6) can be rewritten as

$$f_{\infty,\lambda} = \mathbb{E}[Y \alpha_X], \qquad \alpha_X : x \to (K + \lambda I)^{-1} k(X, x). \tag{30}$$

This shows an interesting property of kernel ridge regression: it can be seen as learning in an unsupervised fashion the weights $\alpha : \mathcal{X} \to L^2(\rho)$, which then indicate on how to fold the input space to use information provided by the labels. At a high-level, one can think of a scheme, given some input points, to perform finite differences and leverage the result to build an estimate of the target

function from Taylor expansions, whatsoever would be the labels observations. Figure 11 shows how when λ is not to big, the reconstruction $f_{\infty,\lambda}(x_0)$ (x_0 being the same point at the bluest center on the different pictures on this Figure) depends on observations made far away from x_0 according to some periodic pattern, implicitly assuming that the target function should be regular when looked at in the Fourier domain. From this picture, one can build examples of non-smooth functions where this inductive bias will have adversarial effects.

Looking at the empirical weights. While the previous paragraph discusses the weights $\alpha_X(x)$ when given access to the full distribution, similar derivations can be made when accessing a finite number of samples. Indeed, kernel ridge regression reads

$$f_{\lambda,n}(x) = \sum_{i\in[n]} Y_i \hat{\alpha}_i(x), \qquad \hat{\alpha}(x) = (\hat{K} + n\lambda)^{-1}\hat{K}_x \in \mathbb{R}^n,$$

where

$$\hat{K} = (k(X_i, X_j))_{i,j\in[n]} \in \mathbb{R}^{n\times n}, \qquad \hat{K}_x = (k(X_i, x))_{i\in[n]} \in \mathbb{R}^n.$$

Note that $\hat{\alpha}_i(x) = \hat{\alpha}_{X_i|\mathcal{D}_{\mathcal{X},n}}(x)$ where $\mathcal{D}_{\mathcal{X},n} = (X_1, \cdots, X_n)$ is the input dataset. As a consequence, we check that

$$\mathbb{E}_{\mathcal{D}_n}[f_n(x)] = \sum_{i\in[n]} \mathbb{E}_{\mathcal{D}_n}[Y_i \hat{\alpha}_{X_i|\mathcal{D}_{\mathcal{X},n}}(x)] = n\cdots\mathbb{E}_{\mathcal{D}_n}[Y_1 \hat{\alpha}_{X_1|\mathcal{D}_{\mathcal{X},n}}(x)].$$

In other terms, f_n is a bias estimator whose average is defined as

$$\mathbb{E}_{\mathcal{D}_n}[f_n] = \mathbb{E}_{(X,Y)}[Y\hat{\alpha}_X], \qquad \hat{\alpha}_X = n\cdot\mathbb{E}_{\mathcal{D}_{\mathcal{X},n}}\left[\hat{\alpha}_{X_1|\mathcal{D}_{\mathcal{X},n}} \,\middle|\, X_1 = X\right]. \tag{31}$$

Those are the weights plotted on Figure 12. In order to compute those weights efficiently, one can use the block matrix inversion, we have, using classical matrix notations

$$\begin{aligned}\hat{\alpha}_{X_n|\mathcal{D}_n}(x) &= [(\hat{K} + n\lambda)^{-1}\hat{K}_X]_n = [(\hat{K} + n\lambda)^{-1}]_{n,:} \times \hat{K}_x \\ &= [(\hat{K} + n\lambda)^{-1}]_{n,:n-1} \times [\hat{K}_x]_{:n-1} + [(\hat{K} + n\lambda)^{-1}]_{n,n} \times [\hat{K}_x]_n \\ &= -(b - x^\top A^{-1}x)^{-1}([\hat{K}_x]_n - x^\top A^{-1} \times [\hat{K}_x]_{:n-1}).\end{aligned}$$

where

$$\hat{K} + n\lambda I = \begin{pmatrix} A & x \\ x^\top & b \end{pmatrix} = \begin{pmatrix} [\hat{K}]_{:n-1,:n-1} + n\lambda & [\hat{K}]_{n,:n-1} \\ [\hat{K}]_{n,:n-1}^\top & [\hat{K}]_{n,n} + n\lambda \end{pmatrix}$$

Let us denote

$$\tilde{K} = (k(X_i, X_j))_{i,j\in[n-1]} \in \mathbb{R}^{n-1\times n-1} \tilde{K}_x = (k(X_i, x))_{i\in[n-1]} \in \mathbb{R}^{n-1},$$

we get

$$\hat{\alpha}_{X|\mathcal{D}_n}(x) = \left(k(X,X) - Z_X^\top Z_X + n\lambda\right)^{-1}\left(k(X,x) - Z_X^\top Z_x\right), \qquad Z_x = (\tilde{K} + n\lambda)^{-1/2}\tilde{K}_x.$$

Those weights are plotted on Figures 12 and 13.

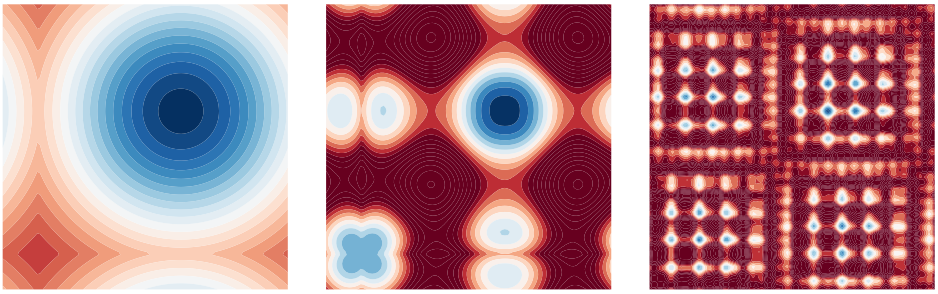

Figure 11: Level lines of the weights $x \to \alpha_x(x_0)$ (30) for a given $x_0 \in \mathcal{X}$, when $\mathcal{X}$ is the torus $\mathbb{R}^2/[-1,1]^2$ and the kernel is taken as the Gaussian kernel with the Riemannian metric on the torus (think of an unrolled donut). Parameters are taken as $\sigma = 1$ together with $\lambda = 10^6$ (left), $\lambda = 10^2$ (middle) or $\lambda = 1$ (right).

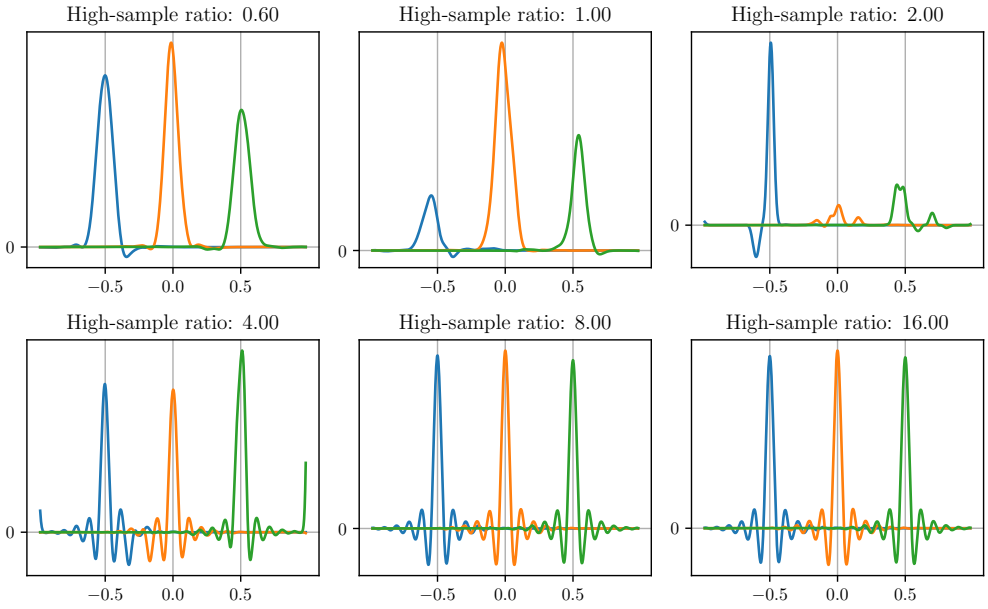

Figure 12: Weights $x \to \hat{\alpha}_X(x)$ as per (30), for $X = -1/2$ (blue), $X = 0$ (orange) and $X = 1/2$ (green), when $\mathcal{X} = [-1, 1]$ and $\rho_{\mathcal{X}}$ is uniform. The weights are computed with the Gaussian kernel with bandwidth $\sigma = .1$ and $\lambda = 10^{-5}$ which yields an effective dimension of $\mathcal{N} = 45$, and for $n \in \{27, 45, 90, 180, 360, 720\}$ which explains the high-sample ratio $n/\mathcal{N}$ seen on the titles of the different plots. When this ratio is close to one, the weights present weird behaviors which could explain the pic observed in convergence rates when transitioning from low-sample to high-sample regimes.

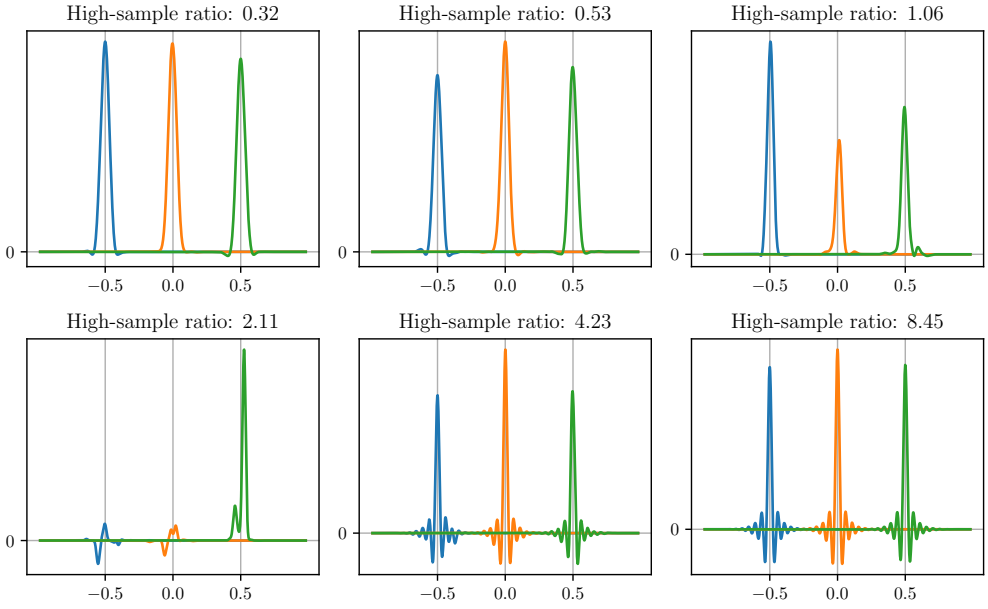

Figure 13: Same picture as Figure 12 yet with $\sigma = .05$, which leads to an effective dimension $\mathcal{N} = 85$.

References

[1] Robert Adams and John Fournier. *Sobolev Spaces*. Academic Press, 1975.

[2] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 1950.

[3] Francis Bach. High-dimensional analysis of double descent for linear regression with random projections. *arXiv preprint arXiv:2303.01372*, 2023.

[4] Francis Bach. *Learning Theory from First Principles*. MIT press (announced), 2023.

[5] Vivien Cabannes, Alessandro Rudi, and Francis Bach. Fast rates in structured prediction. In *Conference on Learning Theory*, 2021.

[6] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 2006.

[7] William Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 1979.

[8] Luc Devroye, Làslzó Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*. Springer, 2013.

[9] Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 2020.

[10] Evelyn Fix and Joseph Hodges. Discriminatory analysis. Nonparametric discrimination: Consistency properties. Technical report, School of Aviation Medicine, Randolph Field, Texas, 1951.

[11] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

[12] Andrey Kolmogorov and Vladimir Tikhomirov. ε-entropy and ε-capacity of sets in functional spaces. *Uspekhi Matematicheskikh Nauk*, 1979.

[13] Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 2020.

[14] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 2022.

[15] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. On the estimation of the derivatives of a function with the derivatives of an estimate. *Applied and Computational Harmonic Analysis*, 2022.

[16] Jaouad Mourtada and Lorenzo Rosasco. An elementary analysis of ridge regression with random design. *Comptes Rendus. Mathématique*, 2022.

[17] Jaouad Mourtada, Tomas Vaškevičius, and Nikita Zhivotovskiy. Distribution-free robust linear regression. *Mathematical Statistics and Learning*, 2022.

[18] Jaouad Mourtada, Tomas Vaškevičius, and Nikita Zhivotovskiy. Local risk bounds in statistical aggregation. *Preprint*, 2023.

[19] Nicolò Pagliana, Alessandro Rudi, Ernesto De Vito, and Lorenzo Rosasco. Interpolation and learning with scale dependent kernels. In *ArXiv*, 2020.

[20] Jaak Peetre. New thoughts on Besov spaces. *Duke University Mathematics Series*, 1976.

[21] Alexander Rakhlin and Xiyu Zhai. Consistency of interpolation with laplace kernels is a high-dimensional phenomenon. In *Conference on Learning Theory*, 2019.

[22] Carl Rasmussen and Christopher Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005.

[23] Bernhard Scholkopf and Alexander Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[24] Steve Smale and Ding-Xuan Zhou. Estimating the approximation error in learning theory. *Analysis and Applications*, 2003.

[25] Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 2007.

[26] Stefano Spigler, Mario Geiger, Stéphane d'Ascoli, Levent Sagun, Giulio Biroli, and Matthieu Wyart. A jamming transition from under- to over-parametrization affects loss landscape and generalization. *Journal of Physics A: Mathematical and Theoretical*, 2019.

[27] Charles Stone. Consistent nonparametric regression. *The Annals of Statistics*, 1977.

[28] Hans Triebel. *Interpolation Theory, Function Spaces, Differential Operators*. North-Holland Publishing Co., 1978.

[29] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[30] Ding-Xuan Zhou. The covering number in learning theory. *Journal of Complexity*, 2002.