Fast kernel methods: Sobolev, physics-informed, and additive models

Nathan Doumèche [1] **Francis Bach** [2] **Gérard Biau** [1] **Claire Boyer** [3]

Abstract

Kernel methods are powerful tools in statistical learning, but their cubic complexity in the sample size n limits their use on large-scale datasets. In this work, we introduce a scalable framework for kernel regression with $\mathcal{O}(n \log n)$ complexity, and designed to fully leverage GPU acceleration. The approach is based on a Fourier representation of kernels combined with non-uniform fast Fourier transforms (NUFFT), enabling exact, fast, and memory-efficient computations. We instantiate our framework in three settings: Sobolev kernel regression, physics-informed regression, and additive models. The proposed estimators are shown to achieve minimax convergence rates, consistent with classical kernel theory. Empirical results demonstrate that our methods can process up to tens of billions of samples within minutes, providing both statistical accuracy and computational scalability. These contributions establish a flexible approach, paving the way for the routine application of kernel methods in large-scale learning tasks, whenever the kernel norm can be efficiently expressed in Fourier space and the ambient dimension d is small. Although the theoretical framework is valid in all dimensions d, the fast Sobolev regression package implemented in the paper relies on the CufiNUFFT package, which currently scales exponentially in d and is limited to $d \leq 3$. For similar reasons, the additive model package is only implemented for simple effects without interaction.

1. Introduction

Kernel methods. Kernel methods play a central role in statistical learning and nonparametric regression, providing flexible and theoretically sound tools for modeling complex data structures (Steinwart & Christmann, 2008). Despite their appeal, their practical application is often hindered by severe computational constraints: standard kernel ridge regression has a time complexity of $\mathcal{O}(n^3)$ and a memory cost of $\mathcal{O}(n^2)$, making it unsuitable for large-scale datasets (Bach, 2024, Chapter 7). Several approximation schemes, such as Nyström methods and random feature expansions, have been proposed to reduce this burden, reducing the computational cost to $\mathcal{O}(n^{3/2})$. However, these approximations introduce additional error terms, complicating theoretical guarantees and often degrading empirical performance (Meanti et al., 2020).

Contributions. We show that a broad class of kernel estimators can be trained exactly at a substantially reduced computational cost, achieving $\mathcal{O}(n \log n)$ complexity in terms of both time and memory. This improvement is made possible by taking advantage of the structural properties of kernels that can be expanded on a Fourier basis. Furthermore, our approach lends itself well to parallel computation and can be efficiently implemented on modern GPU architectures. These advances are a significant step forward in both the theory and practice of kernel methods, with clear potential for routine application to large-scale datasets. A key component of our approach is the recent implementation of the non-uniform fast Fourier transform on GPU (NUFFT, Shih et al., 2021) in dimension $d \leq 3$. The resulting computational efficiency enables the processing of tens of billions of data points in under a minute on a standard GPU, whereas conventional kernel approaches typically fail to scale beyond a few hundred thousand samples (Meanti et al., 2020; Doumèche et al., 2025a).

We illustrate our framework through three fundamental kernel learning tasks: Sobolev regression (Nemirovski, 2000), physics-informed regression (Doumèche et al., 2025b), and additive model regression (Hastie & Tibshirani, 1986). The ability to process massive datasets allows us to systematically assess the performance of the estimators, clearly highlighting the benefits of our approach. In particular, we introduce a kernel- and Fourier-based additive model with a GPU implementation that combines statistical accuracy, favorable scaling in the input dimension d, and high computational efficiency into a unified and theoretically sound framework.

[1]LPSM, Sorbonne University, Paris, France [2]INRIA, Paris, France [3]LMO, Université Paris-Saclay, Orsay, France. Correspondence to: Nathan Doumeche <nathan.doumeche@gmail.com>, Claire Boyer <claire.boyer@universite-paris-saclay.fr >.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

Table 1 completes Rudi et al. (2017, Table 1) and compares the performance of our algorithms to state-of-the-art kernel speed-up methods.

Table 1. Comparison of computational complexity of kernel methods. Logarithmic terms are not shown. In this paper, $n^{d/(2s+d)} = o(n^{1/2})$ as soon as $s > d/2$.

Algorithm	train time	kernel evals	memory	test time
SVM / KRR + direct method	n^3	n^2	n^2	n
KRR + iterative (Caponnetto & Yao, 2010)	$n^2\sqrt{n}$	n	n^2	n^2
Doubly stochastic (Dai et al., 2014)	$n^2\sqrt{n}$	$n^2\sqrt{n}$	n	n
Pegasos / KRR + SGD (Shalev-Shwartz et al., 2011)	n^2	n^2	n	n
KRR + iterative + preconditioning (Ma & Belkin, 2017)	n^2	n^2	n	n
Divide & Conquer (Zhang et al., 2013)	n^2	$\sqrt{n}$	n	n
Nyström / Random Features (Williams & Seeger, 2000)	n^2	$\sqrt{n}$	n	$\sqrt{n}$
Nyström + iterative (Camoriano et al., 2016)	n^2	$\sqrt{n}$	n	$\sqrt{n}$
Nyström + SGD (Dieuleveut & Bach, 2016)	n^2	$\sqrt{n}$	n	$\sqrt{n}$
FALKON (Rudi et al., 2017)	$n\sqrt{n}$	$n\sqrt{n}$	n	$\sqrt{n}$
NUFFT kernels (this article)	n	$N.A.$	n	$n^{d/(2s+d)}$

Related works. The usage of NUFFT to quickly evaluate translation-invariant kernels (i.e., kernels such that $K(x, y) = k(x - y)$ for some function k, and such that the Fourier series of k is known) was introduced by Potts & Steidl (2003) and Potts et al. (2004). The applications of NUFFT in learning theory to reduce to $O(n \log n)$ the overall training cost of kernel methods have then been pioneered by Barnett et al. (2024); Greengard (2025); Greengard et al. (2025), but were still limited to translation-invariant kernels whose Fourier series is easy to compute. However, neither of these two conditions is satisfied by physics-informed kernels, for which the geometry of the problem plays a central role (for a counter-example, see Doumèche et al., 2024, Proposition 5.1). Instead of focusing on the Fourier representation of the kernel, the method developed in the present article can be applied whenever the kernel norm can be quickly computed.

Most of the literature on kernel regression focuses on the asymptotic properties of infinite dimensional RKHS, such as Caponnetto & Vito (2007). This approach is mainly motivated by kernel universality, which is often considered as a requirement for a kernel function. Our point of view here is different and focuses on finding the best estimator for a given regression task. In particular, we show that the L^2-Fourier kernel of Proposition 3.3, which does not approximate any infinite-dimensional kernel, achieves better performance than the Sobolev kernel of Proposition 3.1 when learning a Sobolev function. Our finite-dimensional setting allows us to directly handle the low-smoothness regime $s < d/2$ in Sobolev regression, whereas more standard approaches require adapting high-smoothness kernels together with a source condition (see, e.g., Fischer & Steinwart, 2020; Zhang et al., 2023).

To the best of our knowledge, our contribution is the first one to provide an $O(d^2 n \log n)$ algorithm for additive regression. We believe it to be a key methodological contribution, given that the current state-of-the-art method runs in $O(\max(n^{7/5}, d^2 n))$ (Wood et al., 2014), and considering how massively used additive models are (e.g., Gilbert et al., 2022; Dai et al., 2023). This is particularly relevant for the optimization of additive models on large datasets (i.e., when nd^2 is large), and especially for hyperparameter fitting (which usually requires hundreds of kernel regressions). Real-data applications include grid load prediction with $n = 10^6$ and $d = 10$ (Wood et al., 2014), pollution tracking with $n = 10^8$ and $d = 13$ (Wood et al., 2017), tree survival modeling with $n = 10^5$ and $d = 60$ (Augustin et al., 2025), demand forecasting with $n = 10^5$ and $d = 11$ (Doumèche et al., 2026), electricity demand forecasting with $n = 10^7$ and $d = 15$ (Capezza et al., 2021), and car price prediction with $n = 10^4$ and $d = 10$ (Gabidolla & Carreira-Perpiñán, 2025). Moreover, though the kernel for additive models is translation-invariant, naively applying the framework of Barnett et al. (2024); Greengard (2025); Greengard et al. (2025) would scale exponentially in d.

Outline. In Section 2, we introduce the Fourier-based kernel approximation and the fast optimization strategy. Section 3 develops the Sobolev kernel regression framework and establishes its statistical guarantees. Section 4 extends the method to physics-informed regression, Section 5 to additive models, and Section 6 to translation-invariant kernels. The code for our experiments is available at github.com/NathanDoumeche/fast-kernel. Fast additive models can be used via the fastgam Python package.

2. Common setting

2.1. Nonparametric regression with kernel methods

General assumptions. Throughout this article, we adopt the standard nonparametric regression setting, in which the objective is to estimate an unknown function $f^\star$ from n i.i.d. observations $(X_1, Y_1), \dots, (X_n, Y_n)$ of a random pair (X, Y), under the following assumptions:

(i) The input variable X takes values in an open domain $\Omega \subseteq \mathbb{R}^d$ for some $d \geq 1$, and the output Y is real-valued;

(ii) The output is modeled as $Y = f^\star(X) + \varepsilon$, where the noise term ε satisfies $\mathbb{E}(\varepsilon \mid X) = 0$ and $\mathbb{E}(\varepsilon^2 \mid X) \leq \sigma^2$ for some $\sigma > 0$.

To guarantee that $f^\star$ is learnable from data, we impose the following regularity assumptions:

(iii) The target function belongs to a Sobolev space of smoothness $s > d/2$, that is, $f^\star \in H^s(\Omega)$ for some known $s \in \mathbb{R}^+$;

(iv) The domain Ω is a bounded Lipschitz domain of $\mathbb{R}^d$ (see, e.g., Agranovich, 2015), included in the hypercube $[-L, L]^d$ for some $L > 0$.

Recall that $H^s(\Omega)$ denotes the Sobolev space of functions whose weak derivatives up to order s belong to $L^2(\Omega)$. This space includes, for example, the Hölder space $C^s(\bar{\Omega})$ of continuously differentiable functions with bounded derivatives. The Lipschitz assumption on Ω allows the boundary of the domain to be non-differentiable, encompassing a broad class of bounded sets, including smooth manifolds and common geometries, such as hypercubes.

Kernel methods. In kernel-based methods, this setting naturally leads to the minimization of the regularized empirical risk functional

$$\mathcal{R}(f) = \frac{1}{n}\sum_{j=1}^{n}(f(X_j) - Y_j)^2 + \lambda\|f\|_{\mathcal{H}}^2,$$

over a reproducing kernel Hilbert space (RKHS) $\mathcal{H} \subseteq H^s(\Omega)$, where $\lambda > 0$ is a regularization parameter and $s > d/2$. The norm $\|\cdot\|_{\mathcal{H}}$ serves as a regularizer and reflects the prior assumption that the true function $f^\star$ has a bounded RKHS norm, i.e., $\|f^\star\|_{\mathcal{H}} < \infty$. The resulting estimator $\hat{f} = \arg\min_{f\in\mathcal{H}} \mathcal{R}(f)$ serves as an estimator of the target function $f^\star$. A classical result in kernel theory states that the estimator $\hat{f}$ can be computed in closed form as $\hat{f}(x) = \big(K(X_1, x), \dots, K(X_n, x)\big)\,(\mathbb{K}_n + \lambda I)^{-1}\mathbb{Y}$, where the function K is the reproducing kernel associated with the RKHS $\mathcal{H}$, and $\mathbb{K}_n$ is the $n \times n$ kernel matrix with entries $(\mathbb{K}_n)_{j_1,j_2} = K(X_{j_1}, X_{j_2})$ (e.g., Bach, 2024, Chapter 7). The main computational bottleneck of this approach lies in the analytical computation of the kernel function K and in the inversion of the kernel matrix $\mathbb{K}_n$, which incurs a cost of $\mathcal{O}(n^3)$ in time and $\mathcal{O}(n^2)$ in memory.

Finite-dimensional approximation. To make the computation of the kernel estimator more tractable, we consider a finite-dimensional approximation of $\mathcal{H}$ through a finite basis $\phi_1, \dots, \phi_D$ with $D \in \mathbb{N}^\star$ elements. Thus, any function $f \in \mathcal{H}$ can be approximated by a function f_θ of the form

$$f_\theta(x) = \theta_1\,\bar{\phi}_1(x) + \cdots + \theta_D\,\bar{\phi}_D(x) = \langle\phi(x), \theta\rangle, \quad (1)$$

where $\phi = (\phi_1, \dots, \phi_D)^\top$ denotes the truncated feature map, $\theta \in \mathbb{C}^D$ is the parameter vector, and the inner product is defined as $\langle x, y\rangle = \sum_{k=1}^{D} \bar{x}_k y_k$. In the present paper, we let ϕ be the truncated Fourier basis composed of trigonometric functions.

Since the RKHS approximation if finite-dimensional, there is a positive-definite matrix M such that, for all function $f_\theta \in \mathcal{H}$ given by Equation (1), we have $\|f_\theta\|_{\mathcal{H}}^2 = \|M\theta\|_2^2$. The empirical risk of f_θ then translates into an empirical risk R on the parameter θ:

$$\mathcal{R}(f_\theta) := R(\theta) = n^{-1}\|\Phi\theta - \mathbb{Y}\|_2^2 + \lambda\|M\theta\|_2^2,$$

where $\Phi = \big(\phi(X_1) \mid \cdots \mid \phi(X_n)\big)^*$ is the design matrix (with complex conjugate transpose), and $\mathbb{Y} = (Y_1, \dots, Y_n)^\top \in \mathbb{R}^n$ is the response vector. Minimizing this objective yields the parameter

$$\hat{\theta} = \arg\min_{\theta\in\mathbb{C}^D} R(\theta) = (n^{-1}\Phi^*\Phi + \lambda M^*M)^{-1} n^{-1}\Phi^*\mathbb{Y}, \quad (2)$$

and the resulting estimator is $\hat{f}_n(x) = \langle\phi(x), \hat{\theta}\rangle$.

2.2. Fourier expansion

Using a Fourier basis makes RKHS approximation errors precisely quantifiable, as shown below.

Fourier expansion on the Sobolev space $H^s(\Omega)$. The smoothness of a function in $H^s(\Omega)$ (endowed with any of the equivalent Sobolev norms of order s) is reflected in the decay rate of its Fourier coefficients (see Proposition A.2 in the appendices). Since the Fourier coefficient vector $\theta(f) \in \mathbb{C}^{\mathbb{Z}^d}$ is infinite-dimensional, it cannot be stored or manipulated directly. Thus, in line with the finite-dimensional approximation principle presented above, we truncate the Fourier expansion to a finite number of frequencies, indexed by a cutoff parameter $m \in \mathbb{N}^\star$. In other words, we consider the finite-dimensional approximation space

$$H_m = \Big\{ f : x \mapsto \sum_{k_1,\dots,k_d=-m}^{m} \theta_{k_1,\dots,k_d} \exp\big(i\frac{\pi}{2L}\langle k, x\rangle\big), \quad \theta \in \mathbb{C}^{(2m+1)^d} \Big\},$$

where $k = (k_1, \dots, k_d) \in \mathbb{Z}^d$ is a multi-index. The dimension of this kernel Hilbert space is thus $D = (2m+1)^d$, and the truncated feature map reads

$$\phi(x) = \Big(\exp\big(-i\frac{\pi}{2L}\langle k, x\rangle\big)\Big)_{\|k\|_\infty \le m}.$$

For any function $f \in H^s(\Omega)$, its truncated approximation $f_m \in H_m$ takes the form

$$f_m(x) = \langle\phi(x), (\theta(f)_k)_{\|k\|_\infty\le m}\rangle.$$

Fourier truncation. The truncation of high-frequency Fourier modes is justified by the following approximation result. Let $\mathbb{P}_X$ denote the marginal distribution of the input variable X on the domain Ω. From this point onward, we assume the following regularity condition on $\mathbb{P}_X$:

(v) $\mathbb{P}_X$ is a probability measure on Ω that admits a density with respect to the Lebesgue measure, bounded above by a constant $\kappa > 0$, i.e., $\frac{d\mathbb{P}_X}{dx} \le \kappa$.

Proposition 2.1 (Approximation of the Fourier expansion)**.** *Under assumption (v), there exists a constant $C > 0$, depending only on d and Ω, such that, for all $n \in \mathbb{N}^\star$, all $f \in H^s(\Omega)$, and all truncation levels $m \in \mathbb{N}^\star$, the following inequality holds:*

$$\|f - f_m\|^2_{L^2(\mathbb{P}_X)} \leq C\kappa \, \|f\|^2_{H^s(\Omega)} \, m^{-2s}.$$

Under Assumptions (i) to (v), any estimator $\hat{f}$ of $f^\star$ has an error $\mathbb{E}(\|f^\star - \hat{f}\|^2_{L^2(\mathbb{P}_X)})$ at least of the order of $n^{-2s/(2s+d)}$ (e.g., Tsybakov, 2009, Theorem 2.1). This convergence rate of $2s/(2s+d)$ is known as the Sobolev minimax rate. Letting $m = n^{1/(2s+d)}$ in Proposition 2.1 is thus sufficient to achieve a precision of $n^{-2s/(2s+d)}$. Therefore, since adding extra Fourier modes would not result in an improved asymptotical precision on $\hat{f}_n$, in what follows, we consider $m = n^{1/(2s+d)}$.

2.3. Fast optimization

From naive implementation to fast transforms. Since the total number of Fourier modes is $(2m+1)^d = \mathcal{O}(n^{d/(2s+d)})$, a naive implementation of (2) incurs a complexity of $\mathcal{O}(n^{1+2d/(2s+d)})$, mainly due to the matrix product $\Phi^*\Phi$. While already faster than the standard $\mathcal{O}(n^3)$ kernel approach, we show how this can be further reduced to $\mathcal{O}(n \log n)$—near-optimal, as even reading the data takes $\mathcal{O}(n)$. This speedup is enabled by using complex exponential bases to discretize the kernel, which are well-suited to fast Fourier transform (FFT) algorithms. However, since the exponentials are evaluated at non-uniform points, standard FFTs are not applicable. Instead, we use the non-uniform fast Fourier transform (NUFFT), which extends FFT efficiency to irregular sampling. These algorithms can be efficiently parallelized on modern GPUs, leveraging recent advances in hardware-optimized fast summation. In what follows, we rely on the `cuFINUFFT` library for the GPU-based implementation of the NUFFT, restricted to dimensions 1, 2, and 3 (Shih et al., 2021). The theory developed here holds for any dimension d, but the complexity of cufiNUFFT grows as exponentially in d.

Covariance vector and NUFFT. A naive computation of the covariance vector $v = \Phi^*\mathbb{Y}/n$ via explicit matrix–vector multiplication leads to a complexity of $\mathcal{O}(n^{1+d/(2s+d)})$. However, each component of v can be written explicitly as

$$v_k = \frac{1}{n}\sum_{j=1}^{n} Y_j \exp\big(i\pi\langle k, \frac{X_j}{2L}\rangle\big),$$

which is the definition of the type-I NUFFT of the weighted signal $\mathbb{Y}$ with respect to the non-uniform spatial locations $(X_1, \ldots, X_n)$, leading to time and memory complexity $\mathcal{O}(n \log n)$.

Covariance matrix and NUFFT. A naive computation of the empirical covariance matrix $\hat{\Sigma} = \Phi^*\Phi/n$ via explicit matrix–matrix multiplication incurs a computational complexity of $\mathcal{O}(n^{1+2d/(2s+d)})$. However, one can observe that $\hat{\Sigma}$ is both Hermitian and d-level block Toeplitz. Specifically, it satisfies $\hat{\Sigma}^* = \hat{\Sigma}$ and, for all $k_1, k_2 \in \{-m, \ldots, m\}^d$, the relation $\hat{\Sigma}_{k_1,k_2} = \hat{\Sigma}_{0,k_2-k_1}$ (by elementary properties of the exponential function). This means that the matrix $\hat{\Sigma}$ is fully determined by its first row, i.e., the $(2m+1)^d$ values $(\hat{\Sigma}_{0,k})_{\|k\|_\infty \leq m}$. Each entry can be expressed as

$$\hat{\Sigma}_{0,k} = \frac{1}{n}\sum_{j=1}^{n} \exp\big(i\pi\langle k, \frac{X_j}{2L}\rangle\big),$$

which corresponds to the type-I NUFFT of the constant signal $(1, \ldots, 1)$ evaluated at the sample points $(X_1, \ldots, X_n)$, leading to a time and memory complexity $\mathcal{O}(n \log n)$ with GPU acceleration. Furthermore, for any $x \in \mathbb{C}^{(2m+1)^d}$, the matrix–vector product $\hat{\Sigma}x$ can be performed efficiently via FFT in $\mathcal{O}((2m)^d \log m) = \mathcal{O}(2^d n^{d/(2s+d)} \log n)$ time; see Golub & Van Loan (Theorem 4.8.2, 2013) for the one-dimensional case and the multidimensional construction of Lee (1986).

Matrix inversion. We assume that for any $x \in \mathbb{C}^{(2m+1)^d}$, the matrix–vector product M^*Mx can be performed in $\mathcal{O}(m^d \log m) = \mathcal{O}(n^{d/(2s+d)} \log n)$. This condition is satisfied if M is diagonal or, as shown above, if M^*M is a covariance matrix. Then, since the matrix–vector product $\hat{\Sigma}x$ can also be performed in $\mathcal{O}(n^{d/(2s+d)} \log n)$ time via FFT, linear systems involving $\hat{\Sigma} + \lambda M^*M$ can be solved efficiently using the conjugate gradient method.

The conjugate gradient algorithm is guaranteed to converge in at most $(2m+1)^d$ iterations. Thus, if $s > d/2$ (which is a common assumption in Sobolev kernel learning), the total complexity of solving such systems is reduced to $\mathcal{O}(n^{2d/(2s+d)} \log n) = O(n \log n)$, which is near-linear in the number of data points. In Section 3, we will show that the algorithm for Sobolev regression still converges in $O(n \log n)$ with high probability when $0 < s \leq d/2$.

3. Sobolev regression

3.1. Sobolev (or Matérn) kernel regression

In this section, we study the Sobolev kernel regression, also known as the Matérn kernel setting, where $\mathcal{H} = H^s(\Omega)$ and the RKHS norm is given by $\|f\|^2_{\mathcal{H}} = \sum_{k\in\mathbb{Z}^d} |\theta(f)_k|^2 \, (1 + \|k\|_2^{2s})$. (Since all Sobolev norms are equivalent, the choice of the Sobolev norm does not impact the rate of convergence of the kernel estimator.) In the finite-dimensional approximation using H_m, one has $\|f_\theta\|^2_{\mathcal{H}} = \|S\theta\|_2^2$, where the matrix $S \in \mathbb{C}^{(2m+1)^d \times (2m+1)^d}$ is diagonal with entries $S_{k_1,k_2} = \sqrt{1 + \|k\|_2^{2s}}\mathbf{1}_{k_1=k_2}$. Using tools from kernel the-

ory (Mourtada & Rosasco, 2022; Bach, 2024), we derive the convergence rate of the estimator $\hat{\theta}$ as a function of the sample size n. This result is encapsulated in the following proposition.

Proposition 3.1 (Sobolev kernel regression)**.** *Let $\hat{\theta}$ be the Sobolev kernel estimator (2) with $M = S$, i.e.,*

$$\hat{\theta} = \underset{\theta \in \mathbb{C}^{(2m+1)^d}}{\text{argmin}} \; n^{-1}\|\Phi\theta - \mathbb{Y}\|_2^2 + \lambda\|S\theta\|_2^2$$
$$= (\hat{\Sigma} + \lambda S^2)^{-1} n^{-1} \Phi^* \mathbb{Y}.$$

Under Assumptions (i)-(v),

$$\mathbb{E}(\|\hat{f}_n - f^\star\|^2_{L^2(\mathbb{P}_X)}) \leq \sigma^2 n^{-1}\big(1 + \frac{\alpha^2}{\lambda n}\big)(2m+1)^d +$$
$$\inf_{\theta \in \mathbb{C}^{(2m+1)^d}} \Big\{ \big(2 + 6\big(1 + \frac{\alpha^2}{\lambda n}\big)^2\big)\mathbb{E}(\|f^\star - f_\theta\|^2_{L^2(\mathbb{P}_X)}) +$$
$$\lambda\big(1 + \frac{\alpha^2}{\lambda n}\big)^2 \|S\theta\|_2^2 \Big\},$$

where $\alpha = \sum_{k \in \mathbb{Z}^d} \frac{1}{1+\|k\|_2^{2s}} < \infty$. Thus, choosing the regularization parameter in the range $n^{-1} \leq \lambda \leq n^{-2s/(2s+d)}$ and setting the truncation level as $m = n^{1/(2s+d)}$ yields the convergence rate

$$\mathbb{E}(\|f_{\hat{\theta}} - f^\star\|^2_{L^2(\mathbb{P}_X)}) = \mathcal{O}\big(n^{-2s/(2s+d)}\big).$$

This convergence rate of $2s/(2s+d)$ matches the Sobolev minimax rate, meaning that it is the fastest rate achievable under Assumptions (i) to (v). The existence of minimax Sobolev estimators is well-known in the statistical learning literature (e.g., Nemirovski, 2000, Chapter 2). However, Proposition 3.1 establishes a clear connection between kernel methods and Fourier analysis. If $s > d/2$, then since S is diagonal, this estimator can be computed exactly with complexity $\mathcal{O}(n \log n)$, as shown in Section 2.3. The following Proposition establishes the $O(n \log n)$ complexity of the Sobolev kernel regression in the case where $s \leq d/2$.

Proposition 3.2 (Sobolev kernel algorithmic complexity)**.** *Let $s < d/2$, and $(\theta_k)_{k \in \mathbb{N}}$ be the conjugate gradient iterates. Then, stopping the conjugate gradient when reaching the criterion $\|(\hat{\Sigma} + \lambda S^2)\theta_k - n^{-1}\Phi^*\mathbb{Y}\|_2 \leq \frac{1}{n^3}$ results in the error $\|\theta_k - \hat{\theta}\|_2 \leq 1/n$. The resulting estimator satisfies the minimax rate $\mathbb{E}(\|f_{\theta_k} - f^\star\|^2_{L^2(\mathbb{P}_X)}) = \mathcal{O}\big(n^{-2s/(2s+d)}\big)$. Moreover, with probability greater than $1 - \exp(-n^{2s/(2s+d)})$, this criterion is reached in a complexity $O(n \log(n))$.*

This complexity bounds, combined with the GPU implementation, enables kernel methods to scale to large datasets.

One dimensional example. The goal of this experiment is to demonstrate the ability of our method to efficiently handle datasets containing tens of billions of samples, while experimentally recovering the theoretical convergence rate. We consider a simple one-dimensional regression setting where the input variable X is uniformly distributed over the interval $\Omega =]0, 1[$ and the output is $Y = \exp(X) + \varepsilon$, with a noise $\varepsilon \sim \mathcal{N}(0, 1)$ independent of X. We run the Sobolev kernel estimator with smoothness parameter $s = 1$, regularization $\lambda = n^{-2/3}$, and truncation level $m = n^{1/3}$. The algorithm successfully processes $n = 10^{10}$ data points in approximately one minute on a standard GPU (NVIDIA T4), using complex-128 precision. Figure 1 reports the test error $\mathbb{E}(\|f_{\hat{\theta}} - f^\star\|^2_{L^2(\mathbb{P}_X)})$, estimated over a separate test set of 10^4 samples, and averaged over 20 resamples. The observed performance matches the Sobolev minimax rate of $n^{-2/3}$, providing empirical evidence of the method's numerical stability and scalability.

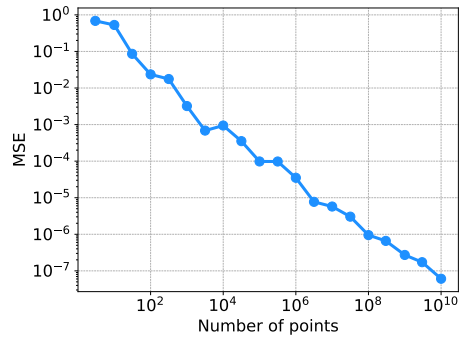

Figure 1. One-dimensional Sobolev kernel regression

In Appendix B.1, we replicate this experiment in the setting $s \leq d/2$ by replacing exp by the Heaviside function $1_{x>0}$, which belongs in all fractional Sobolev spaces $H^s([-1, 1])$ for $s \leq 1/2$, but not when $s > 1/2$. We experimentally recover the minimax rate of $n^{-1/2}$, while the algorithm still scales in $O(n \log n)$.

3.2. L^2-Fourier kernel regression

Dependency on the smoothness s. If the target function $f^\star$ is known to have smoothness s, then it also belongs to all Sobolev spaces $H^{s_2}(\Omega)$ for any $s_2 < s$. As a result, Proposition 3.1 guarantees that all Sobolev kernel estimators with smoothness parameter $s_2 < s$ are well-defined and provably converge to $f^\star$. From a theoretical point of view, higher values of s yield faster convergence rates. However, the non-asymptotic bound in Proposition 3.1 also reveals a bias term involving $\lambda\|S\theta\|_2^2 \geq \lambda\|f^\star\|^2_{H^s(\Omega)}$. Overall, this term scales as $n^{-2s/(2s+d)}\|f^\star\|^2_{H^s(\Omega)}$, highlighting the classical trade-off inherent in the choice of s: larger values improve the convergence rate but come at the cost of increased dependence on the Sobolev norm of the target function.

For example, consider the function $f(x) = x^k$ on the

interval $]-1,1[$. One can compute $\|f\|^2_{H^1(]-1,1[)} = \int_{-1}^{1} f(x)^2 + f'(x)^2\,dx = \frac{2}{2k+1} + \frac{2k^2}{2k-1}$, whereas $\|f\|^2_{H^k(]-1,1[)} \geq \int_{-1}^{1} \left(f^{(k)}(x)\right)^2 dx = 2(k!)^2$, showing that $\|f\|^2_{H^k(]-1,1[)} \gg \|f\|^2_{H^1(]-1,1[)}$ for large k. This illustrates how higher-order Sobolev norms can grow rapidly and lead to high regularization bias despite improved asymptotic rates.

To mitigate the increasing regularization bias associated with large values of s, we propose an alternative estimator that penalizes the unweighted ℓ_2 norm of the parameter vector, i.e., $\lambda\|\theta\|_2^2$, instead of the Sobolev-weighted norm $\lambda\|S\theta\|_2^2$.

Proposition 3.3 (L^2-Fourier kernel regression)**.** *Let $\hat{\theta}$ be the L^2-Fourier kernel estimator given by Equation (2) with $M = I$, where I is the identity matrix. Then,*

$$\begin{aligned}\hat{\theta} &= \underset{\theta\in\mathbb{C}^{(2m+1)^d}}{\text{argmin}}\; n^{-1}\|\Phi\theta - \mathbb{Y}\|_2^2 + \lambda\|\theta\|_2^2 \\ &= (n^{-1}\Phi^*\Phi + \lambda I)^{-1} n^{-1}\Phi^*\mathbb{Y} \\ &= (\hat{\Sigma} + \lambda I)^{-1} n^{-1}\Phi^*\mathbb{Y}.\end{aligned}$$

Under Assumptions (i)-(v),

$$\begin{aligned}\mathbb{E}(\|\hat{f}_n - f^\star\|^2_{L^2(\mathbb{P}_X)}) \leq \lambda\sigma^2 r(1+r) + \\ \inf_{\theta\in\mathbb{C}^{(2m+1)^d}} \left(2 + 6(1+r)^2\right)\mathbb{E}(\|f^\star - f_\theta\|^2_{L^2(\mathbb{P}_X)}) \\ + \lambda(1+r)^2\|\theta\|_2^2,\end{aligned}$$

where $r = \frac{(2m+1)^d}{\lambda n}$. Therefore, choosing the regularization parameter as $\lambda = n^{-2s/(2s+d)}$ and setting the truncation level as $m = n^{1/(2s+d)}$ yields the convergence rate

$$\mathbb{E}[\|f_{\hat{\theta}} - f^\star\|^2_{L^2(\mathbb{P}_X)}] = \mathcal{O}(n^{-2s/(2s+d)}).$$

The kernel associated with this estimator is known as the Dirichlet kernel (Smola & Schölkopf, 1997), i.e.,

$$K(x,y) = \frac{\sin((m+1/2)(x-y)\pi/(2L))}{\sin((x-y)/(4L))}.$$

As before, if $s > d/2$, this estimator can be computed exactly on a GPU with complexity $\mathcal{O}(n\log n)$. The following Proposition establishes the $O(n\log n)$ complexity of the Sobolev kernel regression in the case where $s \leq d/2$.

Proposition 3.4 (L^2-Fourier kernel algorithmic complexity)**.** *Let $s < d/2$, and $(\theta_k)_{k\in\mathbb{N}}$ be the conjugate gradient iterates. Then, stopping the conjugate gradient when reaching the criterion $\|(\hat{\Sigma} + \lambda I)\theta_k - n^{-1}\Phi^*\mathbb{Y}\|_2 \leq \frac{1}{n^3}$ results in the error $\|\theta_k - \hat{\theta}\|_2 \leq 1/n$. The resulting estimator satisfies the minimax rate $\mathbb{E}(\|f_{\theta_k} - f^\star\|^2_{L^2(\mathbb{P}_X)}) = \mathcal{O}\big(n^{-2s/(2s+d)}\big)$. Moreover, with probability greater than $1 - \exp(-n^{2s/(2s+d)})$, this criterion is reached in a complexity $O(n\log(n))$.*

Dependency in s in the one-dimensional case. The goal of this experiment is to compare the performance of the Sobolev kernel estimator with that of its L^2-Fourier variant. We consider a one-dimensional setting where the input variable X is uniformly distributed over the interval $\Omega =]0,1[$ and the response is given by $Y = 25\,|X - 0.5|^3 + \varepsilon$, with a noise $\varepsilon \sim \mathcal{N}(0,1)$ independent of X. We implement both kernel estimators across multiple sample sizes n, and evaluate their MSE on a fixed test set of 10^4 points. To ensure a meaningful comparison, we use the same regularization and truncation schedules for both methods, setting $\lambda_n = n^{-2s/(2s+1)}$ and $m = n^{1/(2s+1)}$ for each smoothness parameter s. We consider 40 values of s uniformly spaced between the minimal admissible value $s = d/2 = 1/2$ and $s = 10$. Figure 2 reports the MSE of the two kernel estimators, averaged over 10 independent resamples for each value of s. This large-scale experiment involves training 400 kernel estimators at sample size $n = 10^8$, and completes in approximately 5 minutes on a standard GPU, illustrating the computational efficiency of both kernel methods.

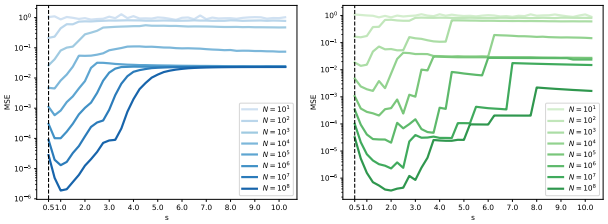

Figure 2. Sobolev regression (Left), and L^2-Fourier regression (Right) in dimension $d = 1$

As expected, for any fixed value of s, the error of both kernel estimators decreases as the number of observations n increases. In addition, we clearly observe the trade-off discussed earlier regarding the choice of the smoothness parameter s. Indeed, the target function $f^\star(x) = 25\,|x - 0.5|^3$ belongs to the Sobolev space $H^3([0,1])$ but not to $H^4([0,1])$, which implies that $s = 3$ is the largest integer smoothness level for which the minimax convergence rate remains theoretically valid. However, in practice, the value $s = 3$ does not correspond to the hyperparameter that yields the lowest MSE. Interestingly, in both kernel estimators, the optimal choice of s increases gradually with the sample size n. This empirical behavior is consistent with the theoretical results of Propositions 3.1 and 3.3, which indicate that the optimal s should converge to 3 as n becomes large.

In practice, the L^2-Fourier kernel consistently outperforms the Sobolev kernel. For example, at $n = 10^8$, its minimum MSE, achieved at $s = 2$, is approximately ten times lower than that of the Sobolev kernel, whose minimum occurs at $s = 1$. For the L^2-Fourier kernel, the shift toward higher values of s occurs more rapidly. This is due to its reduced sensitivity to the explosion of the Sobolev norm as

s increases, allowing it to benefit from the improved theoretical convergence rates associated with higher regularity at smaller sample sizes. Additional experiments can be found in Appendix B.2.

4. Physics-informed regression

Interestingly, our approach, which combines Fourier analysis with kernel methods, applies naturally to the framework of physics-informed machine learning (PIML). PIML incorporates prior physical knowledge–typically expressed as a partial differential equation (PDE)–into regression problems in order to regularize the empirical risk (Karniadakis et al., 2021). In this context, we assume that

(vi) There is a known linear differential operator $\mathcal{D}$ with constant coefficients such that $\forall x \in \Omega,\ \mathcal{D}(f^\star, x) = 0$.

Recall that $\mathcal{D}$ is linear with constant coefficients if $\mathcal{D}(f,x) = \sum_{|\alpha|\le s} a_\alpha \partial^\alpha f(x)$, with $a_\alpha \in \mathbb{R}$. It turns out that this regression setting can be formulated as a kernel method on $\mathcal{H} = H^s(\Omega)$ by minimizing the empirical risk $n^{-1}\sum_{j=1}^n (f_\theta(X_j) - Y_j)^2 + \lambda \|f_\theta\|^2_{H^s(\Omega)} + \mu \int_\Omega \mathcal{D}(f_\theta, x)^2 dx$, as studied by Doumèche et al. (2024); Doumèche et al. (2025a). In what follows, we show that this PDE-based penalty can be efficiently integrated into our Fourier framework and benefit from NUFFT acceleration.

Tractable domains. When the domain Ω is regular enough to allow for the analytical computation of its Fourier transform, the regression problem can be formulated equivalently on the finite-dimensional space H_m as a kernel estimator, of the form:

$$\begin{aligned}\hat\theta &= \operatorname*{argmin}_{\theta\in\mathbb{C}^{(2m+1)^d}} n^{-1}\sum_{j=1}^n (f_\theta(X_j) - Y_j)^2 + \lambda\|f_\theta\|^2_{H^s} \\ &\qquad + \mu\int_\Omega \mathcal{D}(f_\theta, x)^2 dx \\ &= \operatorname*{argmin}_{\theta\in\mathbb{C}^{(2m+1)^d}} n^{-1}\|\Phi\theta - \mathbb{Y}\|_2^2 + \lambda\|S\theta\|_2^2 + \mu\|C^{1/2}D\theta\|_2^2 \\ &= (n^{-1}\Phi^*\Phi + \lambda S^2 + \mu D^* C D)^{-1} n^{-1}\Phi^*\mathbb{Y},\end{aligned}$$

where S is the Sobolev matrix of Section 3.1, the matrix C is such that, for all $k_1, k_2 \in \mathbb{C}^{(2m+1)^d}$,

$$C_{k_1,k_2} = (4L)^{-d}\int_\Omega \exp(i\langle k_1 - k_2, x\rangle \pi/(2L))dx,$$

and $D \in \mathbb{C}^{(2m+1)^d\times(2m+1)^d}$ a diagonal matrix such that $D_{k,k} = \sum_{|\alpha|\le s} a_\alpha \left(\frac{-i\pi}{2L}\right)^{|\alpha|}\prod_{\ell=1}^d (k_\ell)^{\alpha_\ell}$. Since D is a diagonal matrix and C is a d-level block Toeplitz and Hermitian matrix, the matrix-vector product involving D^*CD can be performed in $\mathcal{O}(m^d \log m)$ operations. As a result, the conjugate gradient inversion of the full system matrix $(n^{-1}\Phi^*\Phi + \lambda S^2 + \mu D^* C D)$ can be carried out with overall complexity $\mathcal{O}(n\log n)$.

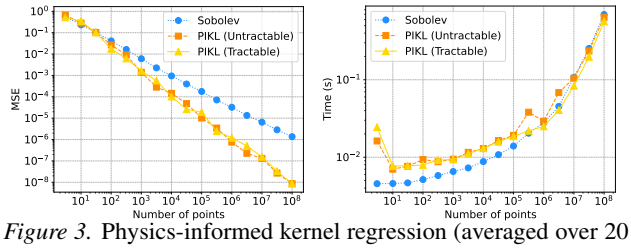

Figure 3. Physics-informed kernel regression (averaged over 20 samples)

Intractable domains. When the Fourier decomposition of the domain Ω is not available in closed form, a common strategy in physics-informed machine learning consists in sampling a set of $n_r \in \mathbb{N}^\star$ collocation points $(X_1^{(r)}, \dots, X_{n_r}^{(r)})$ from a known distribution $\mathbb{P}_{X^{(r)}}$ over Ω. Based on these points, we then define the following kernel estimator:

$$\begin{aligned}\hat\theta &= \operatorname*{argmin}_{\theta\in\mathbb{C}^{(2m+1)^d}} n^{-1}\sum_{j=1}^n (f_\theta(X_j) - Y_j)^2 + \lambda\|f_\theta\|^2_{H^s} \\ &\qquad + \mu n_r^{-1}\sum_{\ell=1}^{n_r} \mathcal{D}(f_\theta, X_j^{(r)})^2 \\ &= \operatorname*{argmin}_{\theta\in\mathbb{C}^{(2m+1)^d}} n^{-1}\|\Phi\theta - \mathbb{Y}\|_2^2 + \lambda\|S\theta\|_2^2 + \mu n_r^{-1}\|\Phi D\theta\|_2^2 \\ &= (n^{-1}\Phi^*\Phi + \lambda S^2 + \mu n_r^{-1} D^*\Phi^{(r)*}\Phi^{(r)} D)^{-1} n^{-1}\Phi^*\mathbb{Y},\end{aligned}$$

where $\Phi^{(r)} = (\phi(X_1^{(r)}) \mid \dots \mid \phi(X_n^{(r)}))^*$. Since D is a diagonal matrix and $(\Phi^{(r)})^*\Phi^{(r)}$ is a d-level block Toeplitz and Hermitian matrix, the associated matrix–vector products can be computed in $\mathcal{O}(m^d \log m)$ operations. Consequently, the conjugate gradient inversion of the system $(n^{-1}\Phi^*\Phi + \lambda S^2 + \mu n_r^{-1} D^*\Phi^{(r)*}\Phi^{(r)} D)$ can be performed with overall complexity $\mathcal{O}(n\log n)$.

One dimensional example. We consider a regression setting where the input variable X is uniformly distributed over the interval $\Omega =]0,1[\subseteq [-\pi/2, \pi/2]$, so that $L = \pi/2$. The response is given by $Y = \exp(X) + \varepsilon$, with a noise $\varepsilon \sim \mathcal{N}(0,1)$ independent of X. The prior knowledge takes the form of the linear differential constraint $(f^\star)' - f^\star = 0$. We set the parameters as $s = d = 1$, $m = n^{1/(2s+d)} = n^{1/3}$, $\lambda = n^{-2s/(2s+d)} = n^{-2/3}$, and $\mu = 1$.

All the kernel methods process $n = 10^8$ samples in under only one second on a standard GPU (NVIDIA L4). Figure 3 reports the test error $\mathbb{E}(\|f_{\hat\theta} - f^\star\|^2_{L^2([0,1])})$, evaluated on a test set of 10^4 samples, and then averaged over 20 resamples. We see that adding physics improves the performance of the estimator, while taking about the same running time.

5. Additive models

In this section, we show how our Fourier–NUFFT framework can incorporate linear constraints on the shape of the regression function such as additivity constraints, yielding a GPU-compatible algorithm with a $\mathcal{O}(n \log n)$ complexity.

Additive model. A well-known limitation of many machine learning methods is the curse of dimensionality, reflected for instance in the minimax rate $\mathcal{O}(n^{-2s/(2s+d)})$ of the Sobolev kernel estimator of Section 3 (achieving accuracy of δ requires at least $n = \delta^{-1-d/(2s)}$ samples). To mitigate this phenomenon, a common approach is to exclude interaction effects between input variables, resulting in the following additive model:

(vii) One has $f^\star(x_1, \dots, x_d) = g_1^\star(x_1) + \cdots + g_d^\star(x_d)$, where $g_1^\star, \dots, g_d^\star$ are univariate functions.

(We show in Appendix B.3 how to extend the framework to additive models with multivariate effects.) The smoothness assumption on $f^\star$ naturally transfers to the functions $g_\ell^\star$. In particular $f^\star$ lies in a Sobolev space of order s, if and only if each function $g_\ell^\star$ belongs in a Sobolev space of order s.

(iii)' There exists a known smoothness parameter $s > 1/2$ such that each component function $g_1^\star, \dots, g_d^\star$ belongs to the univariate Sobolev space of order s.

The L^2-Fourier kernel can thus be extended to the additive model setting. Here, the regression function is parametrized by a concatenated vector $\theta = (\theta_1, \dots, \theta_d) \in \mathbb{C}^{d(2m+1)}$:

$$f_\theta(x_1, \dots, x_d) = \langle \phi(x_1), \theta_1 \rangle + \cdots + \langle \phi(x_d), \theta_d \rangle.$$

Proposition 5.1 (Additive kernel regression)**.** *Let $\hat\theta$ be the truncated Fourier estimator defined by*

$$\hat\theta = \underset{\theta \in \mathbb{C}^{d(2m+1)}}{\text{argmin}} \; n^{-1} \|\Phi_1\theta_1 + \cdots + \Phi_d\theta_d - \mathbb{Y}\|_2^2 + \lambda \|\theta\|_2^2$$
$$= (\hat\Sigma + \lambda I)^{-1} (n^{-1}\Phi_1^*\mathbb{Y}, \dots, n^{-1}\Phi_d^*\mathbb{Y})^\top$$

where $\Phi_\ell = \big(\phi(X_{1,\ell}) \mid \cdots \mid \phi(X_{n,\ell})\big)^*$, and $\hat\Sigma \in \mathbb{C}^{d(2m+1)\times d(2m+1)}$ is the block-matrix such that the (ℓ_1, ℓ_2)-block is $n^{-1}\Phi_{\ell_1}^*\Phi_{\ell_2} \in \mathbb{C}^{d\times d}$. Then, under Assumptions (i), (ii), (iii)', (iv), (v), and (vii), choosing $\lambda = n^{-2s/(2s+1)}$ and $m = n^{1/(2s+1)}/d$ yields the convergence rate

$$\mathbb{E}(\|f_{\hat\theta} - f^\star\|_{L^2(\mathbb{P}_X)}^2) = \mathcal{O}(n^{-2s/(2s+1)}).$$

The kernel associated with this estimator is the sum Dirichlet kernel, i.e., for all $x = (x_1, \dots, x_d)$, and $y = (y_1, \dots, y_d)$,

$$K_d(x, y) = \sum_{j=1}^{d} \frac{\sin((m + 1/2)(x_j - y_j)\pi/(2L))}{\sin((x_j - y_j)/(4L))}. \quad (3)$$

Notice that the additive assumption allows us to recover the more favorable univariate minimax rate of $n^{-2s/(2s+1)}$, as in the case $d = 1$. The scaling $m = n^{1/(2s+1)}/d$ is consistent with existing results on generalized additive models (GAMs) in high dimensions (Wood et al., 2014, Section 3.3). We provide in Appendix A.2 an explicit error bound for the additive kernel regression.

Complexity. Key components of the matrix $\hat\Sigma$—both diagonal and off-diagonal blocks—can be efficiently computed using NUFFT in $\mathcal{O}(d^2 n \log n)$ time, enabling the full system to be solved in nearly linear time when $s \geq 1$ (see Appendix A.2 for details) with GPU acceleration. This is significantly more efficient than GAMs, which require $\mathcal{O}(d^2 n m^2) = \mathcal{O}(\max(d^2 n, n^{7/5}))$ operations (since $m \geq 1$, m scales as $n^{1/5}$ and $s = 2$ for GAMs) and cannot benefit from NUFFT acceleration due to their reliance on spline bases rather than Fourier representations (Wood et al., 2014). In addition, the memory footprint of our kernel method is $\mathcal{O}(n \log n + m^2 d^2) = \mathcal{O}(n \log n)$, while GAMs require $\mathcal{O}(ndm)$ memory to store the design matrix.

Simple additive model without grid search. We consider an additive regression setting in dimension $d = 5$ where

(i) The input variable X is uniformly distributed over the hypercube $\Omega =]0, 1[^5$;

(ii) The response is $Y = \sum_{\ell=1}^{5} \left(\exp\left(\frac{X_\ell}{\ell+1}\right) - 1 \right) + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 1)$ is independent of X.

We implement the L^2-Fourier additive kernel regression estimator with smoothness parameter $s = 2$, and evaluate its performance over 16 different values of n ranging from 1 to 10^8. In Figure 4, we compare our L^2-Fourier additive kernel estimator implemented on CPU (CPU NUFFT) and on a NVIDIA T4 GPU (GPU NUFFT) with the `PyGAM` package, which runs on CPU only (Servén & Brummitt, 2018), and the `FALKON` Python package for both CPU and GPU (Rudi et al., 2017). For a fair comparison, the `PyGAM` estimator is given $2m + 1$ splines for each of the five univariate components, where $m = 1 + \lfloor n^{-1/(2s+1)}/d \rfloor$ with $s = 2$. The `FALKON` algorithm is implemented using Equation (3). This ensures that both kernels and the spline-based estimators use the same number of parameters. In all methods, the regularization parameter is set to $\lambda_n = n^{-2s/(2s+1)}$. Figure 4 (Left) shows that all estimators achieve comparable predictive performance. Figure 4 (Right) highlights the computational advantages of our method. The NUFFT kernel estimators exhibit significantly faster runtimes compared to `PyGAM` and `FALKON`, especially as n increases. In particular, `PyGAM` exceeds the available CPU memory (15 GB) at $n = 10^7$, making it infeasible to evaluate for

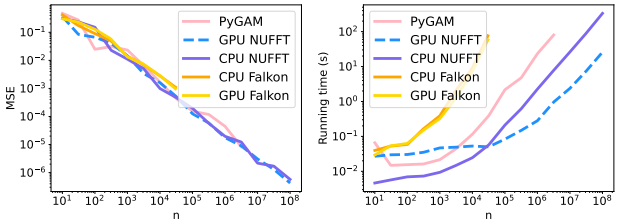

Figure 4. Additive model with $d = 5$

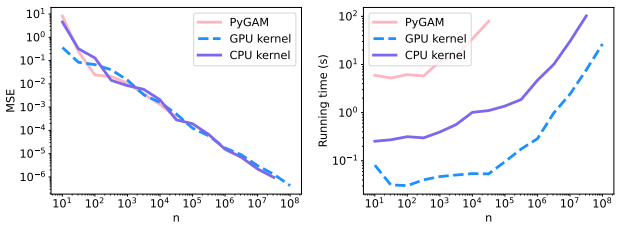

Figure 5. Grid search with 300 hyperparameters for the additive model with $d = 5$.

larger datasets. In contrast, our GPU implementation remains scalable and efficient. For instance, at $n = 10^8$, the GPU kernel estimator is approximately 14 times faster than its CPU counterpart. The GPU implementation of `FALKON` only speeds up the computation by approximately 15%.

Grid search for λ. Interestingly, the most computationally demanding step in the kernel estimation pipeline is the NUFFT computation, which has a complexity of $\mathcal{O}(d^2 n \log n)$. In contrast, the subsequent matrix inversion step is significantly faster, with a complexity of $\mathcal{O}(d^3 m^3) = \mathcal{O}(d^3 n^{3/5})$ when $s = 2$. This observation is particularly relevant because it implies that multiple values of the regularization parameter λ can be tested efficiently, making grid search highly tractable. Figure 5 compares the `gam.gridsearch` function from `PyGAM` to our kernel method implemented on both CPU and GPU, using a grid of 300 candidate values for λ. While the runtime of the `PyGAM` grid search becomes prohibitive at $n = 10^5$, our GPU-based kernel implementation completes the entire grid search in under 30 seconds even for $n = 10^8$.

6. Translation-invariant kernels

General formulation. Our framework can be directly applied to translation-invariant kernels on bounded domains $\Omega \subset [-L, L]^d$, i.e., kernels such that $K(x, y) = \tilde{k}(x - y)$ for some function $\tilde{k} \in H^s(\Omega - \Omega)$, where $\Omega - \Omega = \{x - y \mid x, y \in \Omega\}$. Indeed, from Proposition A.2, we know that k can be extended to $[-2L, 2L]^d$ by a Fourier series $\tilde{k}(x) = \sum_{k \in \mathbb{Z}^d} \psi_k \exp(i \frac{\pi}{2L} \langle k, x \rangle)$ where $\sum_{k \in \mathbb{Z}^d} |\psi_k|_2^2 \|k\|_2^{2s} < \infty$. Thus, by noting D the $(2m+1)^d \times (2m+1)^d$ diagonal matrix with $D_{k,k} = \psi_k$, we have that $K(x, y) = \sum_{k \in \mathbb{Z}^d} \psi_k \exp(i \frac{\pi}{2L} \langle k, x - y \rangle)$, which can be approximated by the truncation $\sum_{k \in \{-m,m\}^d} \psi_k \exp(i \frac{\pi}{2L} \langle k, x - y \rangle) = \langle \phi(x), D\phi(y) \rangle$. The truncated kernel corresponds to the empirical risk $L(\theta) = n^{-1} \|\Phi\theta - \mathbb{Y}\|_2^2 + \lambda \|D^{-1/2}\theta\|_2^2$. The estimator $\hat{\theta} = (\hat{\Sigma} + \lambda D^{-1}) \Phi^\star \mathbb{Y} / n$ is an instantiation of Equation (2), and thus benefits from the NUFFT framework.

Gaussian kernel. An easy way to apply this to the Gaussian kernel $K(x, y) = \exp(-\|x - y\|_2^2)$ on Ω^2 is to let L slowly increase (say $L = \log n$). In this limit, we can approximate the Fourier coefficients by $\psi_k = (4L)^{-d} \int_{[-2L,2L]^d} \tilde{k}(x) e^{-i\pi\langle k,x \rangle / 2L} dx \simeq (4L)^{-d} \int_{\mathbb{R}^d} e^{-\|x\|_2^2 + i \frac{\pi}{2L} \langle k,x \rangle} dx = (4L)^{-d} \exp(-\frac{\pi^2 \|k\|_2^2}{16 L^2})$.

Conclusion

In this work, we have introduced a scalable framework for kernel methods with $\mathcal{O}(n \log n)$ complexity, designed to fully exploit GPU acceleration. The approach combines a Fourier representation of the kernel with the non-uniform fast Fourier transform (NUFFT), yielding fast and memory-efficient computations. We instantiate the framework for Sobolev, physics-informed, and additive kernel models, thereby demonstrating its flexibility and broad applicability. From a theoretical perspective, we establish that the proposed Sobolev and additive kernels achieve optimal minimax convergence rates, in line with classical results from kernel theory.

Future works. In Proposition 3.1, the upper bound on the error involves the Sobolev regularization, vs. the L^2 regularization in Proposition 3.3. This suggests that the L^2-Fourier estimator has lower error, since the L^2 norm is always lower than the Sobolev norm. This improvement is observed in our numerical experiments (see Figure 2). Future works could focus on deriving the corresponding lower bounds, potentially leveraging close-form results on the bias in kernel regression (Bach, 2024, Equation 7.22).

cuFINUFFT is currently limited to $d \leq 3$, although other libraries (e.g., PyNUFFT) handle higher dimensions. Achieving precision ε typically requires the NUFFT complexity to scale as $\log(\varepsilon^{-1})^d$, which grows exponentially in d and limits high-dimensional use. However, the least accurate contributions correspond to high frequencies, which (i) are truncated at $n^{d/(2s+d)}$ in this work and (ii) are damped by the kernel regularization (e.g., Sobolev norms). Future work could exploit this structure to mitigate the apparent exponential dependence on d, or explore other NUFFT algorithms (Yang, 2019).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are no potential societal consequences of our work.

References

Agranovich, M. S. *Sobolev Spaces, Their Generalizations and Elliptic Problems in Smooth and Lipschitz Domains*. Springer, Cham, 2015.

Augustin, N., Albrecht, A., Anaya-Izquierdo, K., Davis, A., Meining, S., Puhlmann, H., and Wood, S. Modeling tree survival for investigating climate change effects. *Journal of the American Statistical Association*, 120:2079–2089, 2025.

Bach, F. *Learning Theory from First Principles*. MIT Press, Cambridge, Massachusetts, 2024.

Barnett, A., Greengard, P., and Rachh, M. Uniform approximation of common Gaussian process kernels using equispaced Fourier grids. *Applied and Computational Harmonic Analysis*, 71:101640, 2024.

Benyi, A. and Oh, T. The sobolev inequality on the torus revisited. *Publicationes Mathematicae Debrecen*, 83: 359–374, 2012.

Camoriano, R., Angles, T., Rudi, A., and Rosasco, L. NYTRO: When subsampling meets early stopping. In Gretton, A. and Robert, C. C. (eds.), *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 1403–1411. PMLR, 2016.

Capezza, C., Palumbo, B., Goude, Y., Wood, S. N., and Fasiolo, M. Additive stacking for disaggregate electricity demand forecasting. *The Annals of Applied Statistics*, 15: 727 – 746, 2021.

Caponnetto, A. and Vito, E. D. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.

Caponnetto, A. and Yao, Y. Cross-validation based adaption for regularization operators in learning theory. *Analysis and Applications*, 8, 2010.

Dai, B., Xie, B., He, N., Liang, Y., Raj, A., Balcan, M.-F., and Song, L. Scalable kernel methods via doubly stochastic gradients. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Dai, Y., Yang, S., Zhao, D., Hu, C., Xu, W., Anderson, D. M., Li, Y., Song, X.-P., Boyce, D. G., Gibson, L., Zheng, C., and Feng, L. Coastal phytoplankton blooms expand and intensify in the 21st century. *Nature*, 615 (7951):280–284, 2023.

Di Nezza, E., Palatucci, G., and Valdinoci, E. Hitchhiker's guide to the fractional sobolev spaces. *Bulletin des Sciences Mathématiques*, 136:521–573, 2012.

Dieuleveut, A. and Bach, F. Non-parametric stochastic approximation with large step sizes. *The Annals of Statistics*, 14:1363–1399, 2016.

Doumèche, N., Bach, F., Biau, G., and Boyer, C. Physics-informed machine learning as a kernel method. In Agrawal, S. and Roth, A. (eds.), *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 1399–1450. PMLR, 2024.

Doumèche, N., Bach, F., Biau, G., and Boyer, C. Physics-informed kernel learning. *Journal of Machine Learning Research*, 26(124):1–39, 2025a.

Doumèche, N., Bach, F., Éloi Bedek, Biau, G., Boyer, C., and Goude, Y. Forecasting time series with constraints. *arXiv:2502.10485*, 2025b.

Doumèche, N., Goude, Y., Rubrichi, S., and Allioux, Y. Human spatial dynamics for electricity demand forecasting. *IEEE Transactions on Power Systems*, 41:1882–1891, 2026.

Fischer, S. and Steinwart, I. Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 21:1–38, 2020.

Gabidolla, M. and Carreira-Perpiñán, M. A. Generalized additive models via direct optimization of regularized decision stump forests. In Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., and Zhu, J. (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 18047–18061. PMLR, 2025.

Gilbert, P. B., Montefiori, D. C., McDermott, A. B., Fong, Y., Benkeser, D., Deng, W., Zhou, H., Houchens, C. R., Martins, K., Jayashankar, L., Castellino, F., Flach, B., Lin, B. C., O'Connell, S., McDanal, C., Eaton, A., Sarzotti-Kelsoe, M., Lu, Y., Yu, C., Borate, B., van der Laan, L. W. P., Hejazi, N. S., Huynh, C., Miller, J., Sahly, H. M. E., Baden, L. R., Baron, M., Cruz, L. D. L., Gay, C., Kalams, S., Kelley, C. F., Andrasik, M. P., Kublin, J. G., Corey, L., Neuzil, K. M., Carpp, L. N., Pajon, R., Follmann, D., Donis, R. O., Koup, R. A., Team§, I. A., Moderna,

I. T., Team§, C. V. P. N. C. E. C., and Team§, U. S. G. U. B. Immune correlates analysis of the mRNA-1273 COVID-19 vaccine efficacy clinical trial. *Science*, 375 (6576):43–50, 2022.

Golub, G. H. and Van Loan, C. F. *Matrix Computations*. The Johns Hopkins University Press, Philadelphia, 2013.

Greengard, P. Efficient Fourier representations of families of Gaussian processes. *SIAM Journal on Scientific Computing*, 47:A971–A990, 2025.

Greengard, P., Rachh, M., and Barnett, A. H. Equispaced Fourier representations for efficient Gaussian process regression from a billion data points. *SIAM/ASA Journal on Uncertainty Quantification*, 13:63–89, 2025.

Hastie, T. and Tibshirani, R. Generalized additive models. *Statistical Science*, 1:297–310, 1986.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Wang, S., Perdikaris, P., and Yang, L. Physics-informed machine learning. *Nature Reviews Physics*, 3:422–440, 2021.

Lee, D. Fast multiplication of a recursive block Toeplitz matrix by a vector and its application. *Journal of Complexity*, 2:295–305, 1986.

Ma, S. and Belkin, M. Diving into the shallows: a computational perspective on large-scale shallow learning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Meanti, G., Carratino, L., Rosasco, L., and Rudi, A. Kernel methods through the roof: Handling billions of points efficiently. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 14410–14422. Curran Associates, Inc., 2020.

Mourtada, J. and Rosasco, L. An elementary analysis of ridge regression with random design. *Comptes Rendus Mathématique*, 360:1055–1063, 2022.

Nemirovski, A. Estimating signals satisfying differential inequalities. In Bernard, P. (ed.), *Lectures on Probability Theory and Statistics: Ecole d'Eté de Probabilités de Saint-Flour XXVIII - 1998*, pp. 155–182. Springer Berlin, 2000.

Nocedal, J. and Wright, S. J. (eds.). *Conjugate Gradient Methods*, pp. 100–133. Springer New York, New York, NY, 1999.

Potts, D. and Steidl, G. Fast summation at nonequispaced knots by NFFT. *SIAM Journal on Scientific Computing*, 24:2013–2037, 2003.

Potts, D., Steidl, G., and Nieslony, A. Fast convolution with radial kernels at nonequispaced knots. *Numerische Mathematik*, 98:329–351, 2004.

Rudi, A., Carratino, L., and Rosasco, L. FALKON: An optimal large scale kernel method. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Servén, D. and Brummitt, C. pyGAM: Generalized additive models in Python, 2018. URL https://doi.org/10.5281/zenodo.1208723.

Shalev-Shwartz, S., Singer, Y., and Srebro, N. Pegasos: Primal estimated sub-gradient solver for SVM. *Mathematical Programming*, 127:3–30, 2011.

Shih, Y.-h., Wright, G., Anden, J., Blaschke, J., and Barnett, A. H. cuFINUFFT: A load-balanced GPU library for general-purpose nonuniform FFTs . In *2021 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pp. 688–697, Los Alamitos, 2021. IEEE Computer Society.

Smola, A. and Schölkopf, B. From regularization operators to support vector kernels. In Jordan, M., Kearns, M., and Solla, S. (eds.), *Advances in Neural Information Processing Systems*, volume 10. MIT Press, 1997.

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

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer, New York, 2009.

Williams, C. and Seeger, M. Using the Nyström method to speed up kernel machines. In Leen, T., Dietterich, T., and Tresp, V. (eds.), *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000.

Wood, S. N., Goude, Y., and Shaw, S. Generalized additive models for large data sets. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 64:139–155, 2014.

Wood, S. N., Li, Z., Shaddick, G., and Augustin, N. H. Generalized additive models for gigadata: Modeling the u.k. black smoke network daily data. *Journal of the American Statistical Association*, 112:1199–1210, 2017.

Yang, H. A unified framework for oscillatory integral transforms: When to use nufft or butterfly factorization? *Journal of Computational Physics*, 388:103–122, 2019.

Zhang, H., Li, Y., Lu, W., and Lin, Q. On the optimality of misspecified kernel ridge regression. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett,

J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 41331–41353. PMLR, 2023.

Zhang, Y., Duchi, J., and Wainwright, M. Divide and conquer kernel ridge regression. In Shalev-Shwartz, S. and Steinwart, I. (eds.), *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pp. 592–617. PMLR, 2013.

A. Additional results

A.1. Properties of Fourier expansion for Sobolev functions

The smoothness of a function in $H^s(\Omega)$ (endowed with any of the equivalent Sobolev norms of order s) is reflected in the decay rate of its Fourier coefficients, as formally stated in the next result.

Proposition A.1 (Fourier series representation, Doumèche et al., 2025a)**.** *Let $s \in \mathbb{N}^\star$. There exists a constant $C_\Omega > 0$ such that, for any function $f \in H^s(\Omega)$, there is a sequence of Fourier coefficients $\theta(f) \in \mathbb{C}^{\mathbb{Z}^d}$ such that:*

(i) $f(x) = \langle \phi_\infty(x), \theta(f) \rangle$, *where* $\phi_\infty(x) = \left(\exp\left(-i\frac{\pi}{2L}\langle k, x\rangle\right)\right)_{k\in\mathbb{Z}^d}$;

(ii) The following norm equivalence holds:

$$\|f\|^2_{H^s(\Omega)} \leq \sum_{k\in\mathbb{Z}^d} |\theta(f)_k|^2 \left(1 + \|k\|_2^{2s}\right) \leq C_\Omega \|f\|^2_{H^s(\Omega)}.$$

This property can be extended to the case where $s \in \mathbb{R}^+$. In such cases, $H^s(\Omega)$ is referred to as a fractional Sobolev space. There are several equivalent ways to define such spaces without having to construct a fractional derivative. Maybe the most straightforward one (see, e.g, Di Nezza et al., 2012, Equations (2.1) and (2.10)) defines $H^s(\Omega)$ when $0 < s < 1$ as the space of functions with finite Gagliardo semi-norm $\|\cdot\|_{G,s,\Omega}$, defined by

$$\|u\|^2_{G,s,\Omega} = \int_\Omega \int_\Omega \frac{(u(x) - u(y))^2}{|x-y|^{2s+d}} dxdy.$$

Then, for $s > 0$, the fractional Sobolev norm is defined by

$$\|u\|^2_{H^s(\Omega)} = \|u\|^2_{H^{\lfloor s \rfloor}(\Omega)} + \sum_{\|\alpha\|_1 = \lfloor s \rfloor} \|\partial^\alpha u\|^2_{G,s-\lfloor s \rfloor,\Omega}.$$

Using basic results on fractional Sobolev spaces, we can extend Proposition A.1 to non-integer values for s.

Proposition A.2 (Fourier series representation)**.** *Let $s \in \mathbb{R}^+$. There exist constants $C_1, C_2 > 0$ such that, for any function $f \in H^s(\Omega)$, there is a sequence of Fourier coefficients $\theta(f) \in \mathbb{C}^{\mathbb{Z}^d}$ such that:*

(i) $f(x) = \langle \phi_\infty(x), \theta(f) \rangle$, *where* $\phi_\infty(x) = \left(\exp\left(-i\frac{\pi}{2L}\langle k, x\rangle\right)\right)_{k\in\mathbb{Z}^d}$;

(ii) The following norm equivalence holds:

$$C_1\|f\|^2_{H^s(\Omega)} \leq \sum_{k\in\mathbb{Z}^d} |\theta(f)_k|^2 \left(1 + \|k\|_2^{2s}\right) \leq C_2 \|f\|^2_{H^s(\Omega)}.$$

Proof. Since Ω is a Lipschitz domain, there is a bounded extension E from $H^s(\Omega)$ to $H^s(\mathbb{R}^d)$, Formally, this means that there is a constant $C_1 > 0$ depending on Ω and s such that, for all function $f \in H^s(\Omega)$, there exists a function $E(f) \in H^s(\mathbb{R}^d)$, such that $\|E(f)\|_{H^s(\mathbb{R}^d)} \leq C_1\|f\|_{H^s(\Omega)}$ and $\forall x \in \Omega, E(f)(x) = f(x)$ (see, e.g, Di Nezza et al., 2012, Theorem 5.4). Now, let $\psi \in C^\infty(\mathbb{R}^d)$ be a smooth function such that

(i) for all $x \in \Omega$, $\psi(x) = 1$,

(ii) for all $x \in \mathbb{R}^d \backslash [-2L+\delta, 2L-\delta]^d$, $\psi(x) = 0$, where $\delta > 0$ is such that $\Omega \subseteq [-2L+\delta, 2L-\delta]^d$.

Then, the multiplication by ψ is continuous (see, e.g, Di Nezza et al., 2012, Lemma 5.3), meaning that there is a constant $C_2 > 0$ depending on s such that, for all function $g \in H^s(\mathbb{R}^d)$, $\|\psi \times g\|_{H^s(\mathbb{R}^d)} \leq C_2\|g\|_{H^s(\mathbb{R}^d)}$. Since all the derivatives of $\psi \times g$ vanish on $[-2L, 2L]^d \backslash [-2L+\delta, 2L-\delta]^d$, this means that the restriction of $\psi \times g$ to $[-2L, 2L]^d$ can be seen as a function in $H^s(\mathbb{T})$ where $\mathbb{T}$ is the torus $\mathbb{R}^d/[-2L, 2L]^d$. Thus, we just showed that for all function $f \in H^s(\Omega)$, there exists a function $h = \psi \times E(f) \in H^s(\mathbb{T})$, such that $\|h\|_{H^s(\mathbb{T})} \leq C_1C_2\|f\|_{H^s(\Omega)}$ and $\forall x \in \Omega, h(x) = f(x)$. Finally, the last

step of the proof is that there are constants $C_3, C_4 > 0$ such that any function $h \in H^s(\mathbb{T})$ can be expanded in a Fourier series in such a way that, almost everywhere with respect to the Lebesgue measure, $h(x) = \sum_{k\in\mathbb{Z}^d} \theta_k \exp\left(-i\frac{\pi}{2L}\langle k, x\rangle\right)$ and $C_3\|h\|^2_{H^s(\mathbb{T})} \leq \sum_{k\in\mathbb{Z}^d\setminus\{0\}} |\theta_k|^2 \left(1 + \|k\|_2^{2s}\right) \leq C_4\|h\|^2_{H^s(\mathbb{T})}$ (see, e.g., Benyi & Oh, 2012, Proposition 1.3). Since f and h coincide on Ω, we have trivially that $\|f\|^2_{H^s(\Omega)} = \|h\|^2_{H^s(\Omega)} \leq \|h\|^2_{H^s(\mathbb{T})}$. Putting everything together, we deduce that

$$C_3\|f\|^2_{H^s(\Omega)} \leq C_3\|h\|^2_{H^s(\mathbb{T})} \leq \sum_{k\in\mathbb{Z}^d\setminus\{0\}} |\theta_k|^2 \left(1 + \|k\|_2^{2s}\right) \leq C_4\|h\|^2_{H^s(\mathbb{T})} \leq C_1C_2C_4\|f\|^2_{H^s(\Omega)}.$$

□

A.2. Additional elements for additive models

In this section, we provide an explicit upper bound on the error of an additive kernel regressor depending on the regularization strength and the truncation level.

Proposition A.3 (Additive kernel regression)**.** *Let $\hat{\theta}$ be the truncated Fourier estimator defined by*

$$\begin{aligned}\hat{\theta} &= \underset{\theta\in\mathbb{C}^{d(2m+1)}}{\text{argmin}}\ n^{-1}\|\Phi_1\theta_1 + \cdots + \Phi_d\theta_d - \mathbb{Y}\|_2^2 + \lambda\|\theta\|_2^2 \\ &= (\hat{\Sigma} + \lambda I)^{-1}(n^{-1}\Phi_1^*\mathbb{Y}, \ldots, n^{-1}\Phi_d^\star\mathbb{Y})^\top,\end{aligned}$$

where $\Phi_\ell = \left(\phi(X_{1,\ell}) \mid \cdots \mid \phi(X_{n,\ell})\right)^\star$, I is the $d(2m+1)\times d(2m+1)$ identity matrix, and $\hat{\Sigma}$ is the block-matrix such that the (ℓ_1, ℓ_2)-block is $n^{-1}\Phi_{\ell_1}^*\Phi_{\ell_2}$, i.e.,

$$\hat{\Sigma} = \begin{pmatrix} n^{-1}\Phi_1^*\Phi_1 & \ldots & n^{-1}\Phi_1^*\Phi_d \\ \vdots & \ddots & \vdots \\ n^{-1}\Phi_d^*\Phi_1 & \ldots & n^{-1}\Phi_d^*\Phi_d \end{pmatrix}.$$

Then, under Assumptions (i), (ii), (iii)', (iv), (v), and (vii), one has

$$\begin{aligned}\mathbb{E}(\|\hat{f}_n - f^\star\|^2_{L^2(\mathbb{P}_X)}) \leq \inf_{\theta\in\mathbb{C}^{d(2m+1)}} &\left(\left(2 + 6(1+r)^2\right)\mathbb{E}(\|f^\star - f_\theta\|^2_{L^2(\mathbb{P}_X)}) + \lambda(1+r)^2\|\theta\|_2^2\right) \\ &+ \lambda\sigma^2 r(1+r),\end{aligned}$$

where $r = \frac{d(2m+1)}{\lambda n}$. Therefore, choosing the regularization parameter as $\lambda = n^{-2s/(2s+1)}$ and setting the truncation level as $m = n^{1/(2s+1)}/d$ yields the convergence rate

$$\mathbb{E}(\|f_{\hat{\theta}} - f^\star\|^2_{L^2(\mathbb{P}_X)}) = \mathcal{O}(n^{-2s/(2s+1)}).$$

Complexity. In this part, we discuss an efficient way to implement $\hat{\theta}$. As explained in Section 2.3, the terms $n^{-1}\Phi_1^*\mathbb{Y}$ and the diagonal blocks $n^{-1}\Phi_\ell^*\Phi_\ell$ of the matrix $\hat{\Sigma}$ can be efficiently computed using the NUFFT. As a result, the vector $n^{-1}(\Phi_1^*\mathbb{Y}, \ldots, \Phi_d^*\mathbb{Y})^\top$ and the diagonal of $\hat{\Sigma}$ can be computed with overall complexity $\mathcal{O}(dn\log n)$. Moreover, the off-diagonal block $n^{-1}\Phi_{\ell_1}^*\Phi_{\ell_2}$ corresponds to the 2d-NUFFT of the constant function equal to 1, evaluated at the points $(X_{1,\ell_1}, -X_{1,\ell_2}), \ldots, (X_{n,\ell_1}, -X_{n,\ell_2})$. The matrix $n^{-1}\Phi_{\ell_1}^*\Phi_{\ell_2}$ can thus be computed in $\mathcal{O}(n\log n)$. Consequently, the full matrix $\hat{\Sigma}$ can be assembled in $\mathcal{O}(d^2 n\log n)$ operations (since there are d^2 blocks of $\Phi^\star\Phi$-type). The linear system involving $(\hat{\Sigma} + \lambda I)$ can then be solved in $\mathcal{O}(d^3m^3) = \mathcal{O}(n)$ operations, provided that $s \geq 1$.

B. Additional experiments

B.1. Low smoothness experiment

Here, we apply our algorithm to the case $s < d/2$, while still maintaining a $O(n\log n)$ complexity. We start by recalling that the step function $f(x) = 1_{x>0}$ belongs to the fractional Sobolev space $H^s([-1,1])$ for any $s \in [0, 1/2]$, but not for $s > 1/2$.

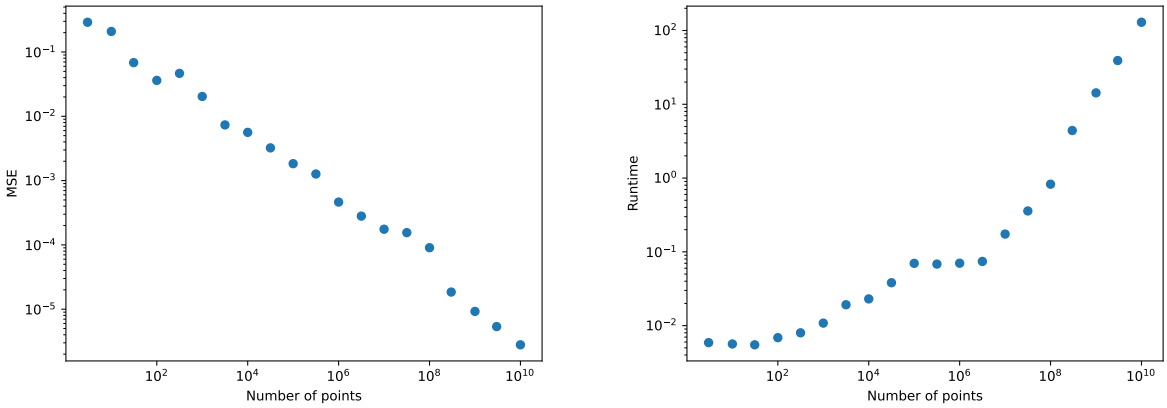

Figure 6. Sobolev regression with $s = 1/2$. Mean square error (Left) and runtime in seconds (Right).

Indeed, the Gagliardo norm of f is $\|f\|^2_{H^s([-1,1])} = \int_{[-1,1]^2} \frac{(1_{x>0}-1_{y>0})^2}{(x-y)^{1+2s}} dxdy$, which is finite only if $0 < s \leq 1/2$ and is equal to $\frac{2(2-2^{1-2s})}{s(1-2s)}$.

We then implement the Sobolev kernel regression with $s = 1/2 \leq d/2$, where X is uniformly distributed over the domain $\Omega =]-1, 1[$, $Y = 1_{X>0} + \varepsilon$, and $\varepsilon \sim \mathcal{N}(0, 1)$. The results in Figure 6 experimentally confirm that the kernel estimator achieves the minimax rate $n^{-1/2}$ and that the runtime grows linearly (after a transition phase up to $n = 10^6$, which can be explained by GPU communication and Python language overheads).

B.2. Dependency in s in the two-dimensional case for the Sobolev regression.

In this experiment, we aim to compare the performance of the Sobolev kernels in dimension 2, for the task of approximating smooth functions. The regression setting is defined as follows:

(i) The input $X = (X^{(1)}, X^{(2)})$ is uniformly distributed over the square domain $\Omega =]0, 1[^2$;

(ii) The response is given by $Y = \exp(X^{(1)}) \cos(X^{(2)}) + \varepsilon$, where the noise term $\varepsilon \sim \mathcal{N}(0, 1)$ is independent of X.

We implement both kernel estimators for several sample sizes n, and evaluate their MSE on a test set of 10^4 points. To ensure a meaningful comparison, we adopt the same regularization and truncation schedules for both methods, setting $\lambda_n = n^{-2s/(2s+2)}$ and $m = n^{1/(2s+2)}$. We consider 40 values of the smoothness parameter s, uniformly spaced between $s = d/2 = 1$ and $s = 10$. Figure 7 reports the MSEs of both techniques, averaged over 10 independent resamples for each value of s. This large-scale experiment, which involves training 400 kernel estimators at $n = 10^8$, completes in approximately 21 minutes on a standard GPU (NVIDIA L4).

Once again, the L^2-Fourier kernel consistently outperforms the Sobolev kernel. In this example, the target function satisfies $f^\star \in \cap_{s\in\mathbb{N}} H^s(\Omega)$, meaning that it is infinitely smooth and the theoretically optimal regularity parameter is $s^\star = \infty$. As in the one-dimensional case, the optimal value of s for the Sobolev kernel increases more slowly toward $s^\star$ than for the L^2-Fourier estimator, which benefits more rapidly from the higher smoothness of $f^\star$.

B.3. Additive models with cross terms

The framework developed in this paper can easily be extended to add interaction terms (i.e., allowing models of the form $f^\star(x_1, x_2, x_3, x_4) = g_1^\star(x_1, x_2) + g_2^\star(x_3, x_4)$, where $g_1^\star$ and $g_2^\star$ are bivariate functions) to the additive model regression. Indeed, consider the model $f^\star(x) = \sum_{\ell=1}^{\ell_{max}} g_\ell^\star(x_{2\ell+1}, x_{2\ell+2})$ under the Sobolev assumption $g_\ell^\star \in H^s$. Then, the bivariate functions can be expanded into Fourier series $g_\ell^\star(x_{2\ell+1}, x_{2\ell+2}) = \langle \phi(x_{2\ell+1}, x_{2\ell+2}), \theta_\ell \rangle$ similarly to what is done in the Sobolev regression frameworks. Then, all the arguments derived in Appendix A.2 are valid, except that instead of involving the 1d-NUFFT to compute $n^{-1}\Phi_\ell^\star \Phi_\ell$ and the 2d-NUFFT to compute $n^{-1}\Phi_{\ell_1}^\star \Phi_{\ell_2}$, this would now involve the 2d-NUFFT of $(x_{2\ell+1}, x_{2\ell+2})$ to compute $n^{-1}\Phi_\ell^\star \Phi_\ell$ and the 4d-NUFFT of $(x_{2\ell_1+1}, x_{2\ell_1+2}, x_{2\ell_2+1}, x_{2\ell_2+2})$ to compute $n^{-1}\Phi_{\ell_1}^\star \Phi_{\ell_2}$.

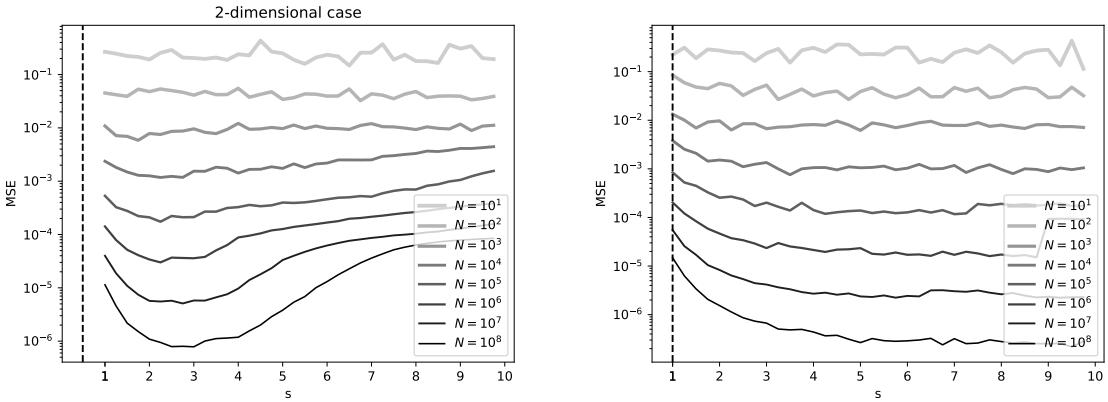

Figure 7. Sobolev regression (Left), and L^2-Fourier regression (Right) in dimension $d = 2$

C. Proofs

C.1. Proof of Proposition 2.1

For $f \in H^s(\Omega)$, let $\theta(f) \in \mathbb{C}^{Z^d}$ be the Fourier coefficients of f, as in Proposition A.2. To simplify notation, we write $\theta = \theta(f)$. Let f_m be the projection of f on H_m, defined by

$$f_m(x) = \sum_{\|k\|_\infty \leq m} \theta_k \exp(i\pi\langle k, x\rangle/2L).$$

Parseval's identity ensures that

$$\mathbb{E}((f(X) - f_m(X))^2) \leq \kappa \int_{[-2L,2L]^d} (f(x) - f_m(x))^2 dx = \kappa(4L)^d \sum_{\|k\|_\infty > m} |\theta_k|^2. \tag{4}$$

Since $f \in H^s(\Omega)$, one has $\sum_{k\in\mathbb{Z}^d} |\theta_k|^2 \|k\|_2^{2s} \leq C_\Omega \|f\|^2_{H^s(\Omega)} < \infty$, where C_Ω is the Sobolev embedding constant in Proposition A.2. Since $\|k\|_\infty > m \Rightarrow \|k\|_2 > m$, we deduce that

$$\sum_{\|k\|_\infty > m} |\theta_k|^2 = \sum_{\|k\|_\infty > m} (|\theta_k|^2 \|k\|_2^{2s}) \|k\|_2^{-2s} \leq \sum_{\|k\|_\infty > m} (|\theta_k|^2 \|k\|_2^{2s}) m^{-2s} \leq C_\Omega \|f\|^2_{H^s(\Omega)} m^{-2s}.$$

Combining this result with (4), we deduce that $\mathbb{E}((f(X) - f_m(X))^2) \leq C\kappa \|f\|^2_{H^s(\Omega)} m^{-2s}$, where $C = C_\Omega (4L)^d$.

C.2. Proof of Propositions 3.1 and 3.3

General kernel regression setting. Let $\theta \in \mathbb{C}^{(2m+1)^d}$ be an arbitrary vector of Fourier coefficients. We are interested in bounding the quantity

$$\begin{aligned}\mathbb{E}(\|f_{\hat\theta} - f^\star\|^2_{L^2(\mathbb{P}_X)}) &= \mathbb{E}(\|(f_{\hat\theta} - f_\theta) + (f_\theta - f^\star)\|^2_{L^2(\mathbb{P}_X)}) \\ &\leq 2\mathbb{E}(\|f_{\hat\theta} - f_\theta\|^2_{L^2(\mathbb{P}_X)}) + 2\mathbb{E}(\|f_\theta - f^\star\|^2_{L^2(\mathbb{P}_X)}).\end{aligned} \tag{5}$$

Let us start by bounding the error term $\mathbb{E}(\|f_{\hat\theta} - f_\theta\|^2_{L^2(\mathbb{P}_X)})$.

Generalization error. In both propositions, $\hat\theta$ can be written as

$$\hat\theta = (\hat\Sigma + \lambda R)^{-1} \Phi^\star \mathbb{Y}/n,$$

where $\hat\Sigma = n^{-1}\Phi^\star\Phi$ is the empirical covariance matrix and R is a regularization matrix: $R = S^2$ in the Sobolev regression, and $R = I$ for the L^2-Fourier regression. Let $\Sigma = \mathbb{E}(\hat\Sigma)$ be the theoretical covariance matrix, defined by

$\langle \theta_1, \Sigma\theta_2 \rangle = \mathbb{E}(\langle \phi(X), \theta_1 \rangle \langle \phi(X), \theta_2 \rangle)$. Since $\mathbb{Y}$ can be decomposed as $\mathbb{Y} = \Phi\theta + \delta\mathbb{Y} + \epsilon$, where $\epsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top$ and $\delta\mathbb{Y}_j = f^\star(X_j) - f_\theta(X_j)$, we can write

$$\begin{aligned}\hat{\theta} &= (\hat{\Sigma} + \lambda R)^{-1}\Phi^\star \mathbb{Y}/n \\ &= (\hat{\Sigma} + \lambda R)^{-1}(\hat{\Sigma}\theta + \Phi^\star(\delta\mathbb{Y} + \epsilon)/n) \\ &= \theta + (\hat{\Sigma} + \lambda R)^{-1}(-\lambda R\theta + \Phi^\star(\delta\mathbb{Y} + \epsilon)/n).\end{aligned}$$

This leads to the following decomposition of the generalization error:

$$\begin{aligned}\mathbb{E}(\|f_{\hat{\theta}} - f_\theta\|^2_{L^2(\mathbb{P}_X)}) &= \mathbb{E}(\|\sqrt{\Sigma}(\hat{\theta} - \theta)\|_2^2) \\ &= \mathbb{E}(\|\sqrt{\Sigma}(\hat{\Sigma} + \lambda R)^{-1}(-\lambda R\theta + \Phi^*(\epsilon + \delta\mathbb{Y})/n)\|_2^2) \\ &\leq 3\mathbb{E}(\|\sqrt{\Sigma}(\hat{\Sigma} + \lambda R)^{-1}\lambda R\theta\|_2^2) + 3\mathbb{E}(\|\sqrt{\Sigma}(\hat{\Sigma} + \lambda R)^{-1}\Phi^*\epsilon/n\|_2^2) \\ &\quad + 3\mathbb{E}(\|\sqrt{\Sigma}(\hat{\Sigma} + \lambda R)^{-1}\Phi^*\delta\mathbb{Y}/n\|_2^2),\end{aligned}$$

where we used the fact that $(x_1 + x_2 + x_3)^2 \leq 3x_1^2 + 3x_2^2 + 3x_3^2$. The first term $\mathbb{E}(\|\sqrt{\Sigma}(\hat{\Sigma} + \lambda R)^{-1}\lambda R\theta\|_2^2)$ is a bias term depending on the regularization λR. The second term $\mathbb{E}(\|\sqrt{\Sigma}(\hat{\Sigma} + \lambda R)^{-1}\Phi^*\epsilon/n\|_2^2)$ is a variance term depending on the noise ϵ. The last term $\mathbb{E}(\|\sqrt{\Sigma}(\hat{\Sigma} + \lambda R)^{-1}\Phi^*\delta\mathbb{Y}/n\|_2^2)$ is another approximation error term, which measures the impact of the interaction between the regularization R and the approximation errors $\delta\mathbb{Y}$. In what follows, we will successively bound these three errors terms.

Bias term. According to Bach (2024, Lemma 7.1), one has

$$\begin{aligned}\mathbb{E}(\|\sqrt{\Sigma}(\hat{\Sigma} + \lambda R)^{-1}\lambda R\theta\|_2^2) &= \lambda^2\mathbb{E}(\|\sqrt{\Sigma}R^{-1/2}(R^{-1/2}\hat{\Sigma}R^{-1/2} + \lambda)^{-1}R^{1/2}\theta\|_2^2) \\ &\leq \lambda(1 + \frac{\alpha^2}{\lambda n})^2\langle R^{1/2}\theta, R^{-1/2}\Sigma R^{-1/2}(R^{-1/2}\Sigma R^{-1/2} + \lambda)^{-1}R^{1/2}\theta\rangle \\ &\leq \lambda(1 + \frac{\alpha^2}{\lambda n})^2\|R^{1/2}\theta\|_2^2,\end{aligned}$$

where $\alpha = \max_{x\in\mathbb{R}^d}\|R^{-1/2}\phi(x)\|_2$. Observe that $R^{1/2} = S$ for the Sobolev regression and $R^{1/2} = I$ for the L^2-Fourier regression. Thus, $\alpha^2 = \sum_{k\in\mathbb{Z}^d}\frac{1}{1+\|k\|_2^{2s}} < \infty$ for the Sobolev regression, while $\alpha^2 = (2m+1)^d$ for the L^2-Fourier regression. Remark that $\|R^{1/2}\theta\|_2^2$ is the kernel norm.

Variance term. By expanding $\epsilon = \sum_{j=1}^n \varepsilon_j e_j$ on the canonical basis $(e_1, \dots, e_n)$, we have that

$$\begin{aligned}&\mathbb{E}(\|\sqrt{\Sigma}(\hat{\Sigma} + \lambda R)^{-1}\Phi^*\epsilon/n)\|_2^2) \\ &= n^{-2}\sum_{j_1,j_2=1}^n \mathbb{E}(e_{j_1}^\star\Phi(\hat{\Sigma} + \lambda R)^{-1}\Sigma(\hat{\Sigma} + \lambda R)^{-1}\Phi^* e_{j_2}\mathbb{E}(\varepsilon_{j_1}\varepsilon_{j_2} \mid X_1, \dots X_n)).\end{aligned}$$

Since ε_{j_1} and ε_{j_2} are independent when $j_1 \neq j_2$, and since $\mathbb{E}(\varepsilon_j \mid X_1, \dots X_n)) = 0$, we deduce that

$$\begin{aligned}&\mathbb{E}(\|\sqrt{\Sigma}(\hat{\Sigma} + \lambda R)^{-1}\Phi^*\varepsilon/n\|_2^2) \\ &= \mathbb{E}(\varepsilon_1^2)n^{-2}\mathrm{tr}(\mathbb{E}(\Phi(\hat{\Sigma} + \lambda R)^{-1}\Sigma(\hat{\Sigma} + \lambda R)^{-1}\Phi^*)) \\ &\leq \sigma^2 n^{-2}\mathrm{tr}(\mathbb{E}(\Phi R^{-1/2}(R^{-1/2}\hat{\Sigma}R^{-1/2} + \lambda)^{-1}R^{-1/2}\Sigma R^{-1/2}(R^{-1/2}\hat{\Sigma}R^{-1/2} + \lambda)^{-1}R^{-1/2}\Phi^*)) \\ &= \sigma^2 n^{-1}\mathrm{tr}(\mathbb{E}((R^{-1/2}\hat{\Sigma}R^{-1/2} + \lambda)^{-1}R^{-1/2}\Sigma R^{-1/2}(R^{-1/2}\hat{\Sigma}R^{-1/2} + \lambda)^{-1}R^{-1/2}\hat{\Sigma}R^{-1/2})) \\ &\leq \sigma^2 n^{-1}\mathrm{tr}(\mathbb{E}((R^{-1/2}\hat{\Sigma}R^{-1/2} + \lambda)^{-1}R^{-1/2}\Sigma R^{-1/2})).\end{aligned}$$

Using Bach (2024, Lemma 7.1), we conclude that

$$\begin{aligned}\mathbb{E}(\|\sqrt{\Sigma}(\hat{\Sigma} + \lambda R)^{-1}\Phi^*\varepsilon/n\|_2^2) &\leq \sigma^2 n^{-1}(1 + \frac{\alpha^2}{\lambda n})\mathrm{tr}(\mathbb{E}((R^{-1/2}\Sigma R^{-1/2} + \lambda)^{-1}R^{-1/2}\Sigma R^{-1/2})) \\ &\leq \sigma^2 n^{-1}(1 + \frac{\alpha^2}{\lambda n})(2m+1)^d.\end{aligned}$$

Approximation term. Let (X_{n+1}, Y_{n+1}) be a new sample, drawn independently from $(X_1, Y_1), \dots, (X_n, Y_n)$. Let

$$C = \sum_{j=1}^{n+1} R^{-1/2}\phi(X_j)\phi(X_j)^* R^{-1/2},$$

so that $C = nR^{-1/2}\hat{\Sigma}R^{-1/2} + R^{-1/2}\phi(X_{n+1})\phi(X_{n+1})^* R^{-1/2}$. The rationale behind this approach is that (X_{n+1}, Y_{n+1}) relates to Σ by $R^{-1/2}\Sigma R^{-1/2} = \mathbb{E}(R^{-1/2}\phi(X_{n+1})\phi(X_{n+1})^* R^{-1/2})$. Moreover, $\hat{\Sigma}$ and C are related by Bach (2024, Equation (7.26)), so that

$$\begin{aligned}
&(C + \lambda n)^{-1} R^{-1/2}\phi(X_{n+1}) \\
&\quad = (1 + \langle R^{-1/2}\phi(X_{n+1}), (nR^{-1/2}\hat{\Sigma}R^{-1/2} + n\lambda)^{-1} R^{-1/2}\phi(X_{n+1})\rangle)^{-1} \\
&\qquad \times (nR^{-1/2}\hat{\Sigma}R^{-1/2} + n\lambda)^{-1} R^{-1/2}\phi(X_{n+1}),
\end{aligned}$$

where the scaling coefficient is such that

$$(1 + \langle R^{-1/2}\phi(X_{n+1}), (nR^{-1/2}\hat{\Sigma}R^{-1/2} + n\lambda)^{-1} R^{-1/2}\phi(X_{n+1})\rangle)^{-1} \geq (1 + \frac{\alpha^2}{\lambda n})^{-1}.$$

Thus,

$$\begin{aligned}
&\mathbb{E}(\|\sqrt{\Sigma}(\hat{\Sigma} + \lambda R)^{-1}\Phi^*\delta\mathbb{Y}/n\|_2^2) \\
&\quad = n^{-2}\mathbb{E}((\delta\mathbb{Y})^\star \Phi(\hat{\Sigma} + \lambda R)^{-1}\phi(X_{n+1})\phi(X_{n+1})^*(\hat{\Sigma} + \lambda R)^{-1}\Phi^*\delta\mathbb{Y}) \\
&\quad = n^{-2}\mathbb{E}(|\langle \Phi(\hat{\Sigma} + \lambda R)^{-1}\phi(X_{n+1}), \delta\mathbb{Y}\rangle|^2) \\
&\quad = n^{-2}\mathbb{E}(|\langle \Phi R^{-1/2}(R^{-1/2}\hat{\Sigma}R^{-1/2} + \lambda)^{-1} R^{-1/2}\phi(X_{n+1}), \delta\mathbb{Y}\rangle|^2) \\
&\quad \leq (1 + \frac{\alpha^2}{\lambda n})^2 \mathbb{E}(|\langle \Phi R^{-1/2}(C + \lambda n)^{-1} R^{-1/2}\phi(X_{n+1}), \delta\mathbb{Y}\rangle|^2) \\
&\quad \leq (1 + \frac{\alpha^2}{\lambda n})^2 \mathbb{E}(\|\Phi R^{-1/2}(C + \lambda n)^{-1} R^{-1/2}\phi(X_{n+1})\|_2^2 \, \|\delta\mathbb{Y}\|^2).
\end{aligned}$$

Since $R^{-1/2}\Phi^\star\Phi R^{-1/2} = nR^{-1/2}\hat{\Sigma}R^{-1/2} \leq C$, we deduce that

$$\begin{aligned}
&\|\Phi R^{-1/2}(C + \lambda n)^{-1} R^{-1/2}\phi(X_{n+1})\|_2^2 \\
&\quad \leq \phi(X_{n+1})^* R^{-1/2}(C + \lambda n)^{-1} C (C + \lambda n)^{-1} R^{-1/2}\phi(X_{n+1}) \\
&\quad \leq \phi(X_{n+1})^* R^{-1/2}(C + \lambda n)^{-1} R^{-1/2}\phi(X_{n+1}).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\mathbb{E}(\|\sqrt{\Sigma}(\hat{\Sigma} + \lambda R)^{-1}\Phi^*\delta\mathbb{Y}/n\|_2^2) \\
&\quad \leq (1 + \frac{\alpha^2}{\lambda n})^2 \mathbb{E}(\|(C + \lambda n)^{-1/2} R^{-1/2}\phi(X_{n+1})\|_2^2 \, \|\delta\mathbb{Y}\|^2) \\
&\quad \leq (1 + \frac{\alpha^2}{\lambda n})^2 \mathbb{E}(\|(C + \lambda n)^{-1/2} R^{-1/2}\phi(X_{n+1})\|_2^2 \, (\|\delta\mathbb{Y}\|^2 + (\delta Y_{n+1})^2)).
\end{aligned}$$

Since the observations $(X_1, Y_1), \dots, (X_{n+1}, Y_{n+1})$ are assumed i.i.d., the random variables $\|(C + \lambda n)^{-1/2} R^{-1/2}\phi(X_j)\|_2^2 \; (\|\delta\mathbb{Y}\|^2 + (\delta Y_{n+1})^2)$ have the same distribution for all $1 \leq j \leq n+1$. Therefore, they

also share the same expectation. Thus,

$$\begin{aligned}
&\mathbb{E}(\|(C+\lambda n)^{-1/2}R^{-1/2}\phi(X_{n+1})\|_2^2\,(\|\delta\mathbb{Y}\|^2+(\delta Y_{n+1})^2))\\
&=\frac{1}{n+1}\mathbb{E}\Big(\sum_{j=1}^{n+1}\|(C+\lambda n)^{-1/2}R^{-1/2}\phi(X_j)\|_2^2\,(\|\delta\mathbb{Y}\|^2+(\delta Y_{n+1})^2)\Big)\\
&=\frac{1}{n+1}\mathbb{E}\Big(\sum_{j=1}^{n+1}\mathrm{tr}(\phi(X_j)^*R^{-1/2}(C+\lambda n)^{-1}R^{-1/2}\phi(X_j))\,(\|\delta\mathbb{Y}\|^2+(\delta Y_{n+1})^2)\Big)\\
&=\frac{1}{n+1}\mathbb{E}\Big(\mathrm{tr}((C+\lambda n)^{-1}\sum_{j=1}^{n+1}R^{-1/2}\phi(X_j)\phi(X_j)^*R^{-1/2})\,(\|\delta\mathbb{Y}\|^2+(\delta Y_{n+1})^2)\Big)\\
&=\frac{1}{n+1}\mathbb{E}\Big(\mathrm{tr}((C+\lambda n)^{-1}C)\,(\|\delta\mathbb{Y}\|^2+(\delta Y_{n+1})^2)\Big)\\
&\le\frac{1}{n+1}\mathbb{E}\Big((\|\delta\mathbb{Y}\|^2+(\delta Y_{n+1})^2)\Big)=\mathbb{E}((\delta Y^2))=\mathbb{E}(\|f^\star-f_\theta\|^2_{L^2(\mathbb{P}_X)}),
\end{aligned}$$

where we use the fact $x=\mathrm{tr}(x)$ for any real number x, and $\mathrm{tr}(M_1M_2)=\mathrm{tr}(M_2M_1)$. All in all, we have

$$\mathbb{E}(\|\sqrt{\Sigma}(\hat{\Sigma}+\lambda R)^{-1}\Phi^*\delta\mathbb{Y}/n\|_2^2)\le(1+\frac{\alpha^2}{\lambda n})^2\,\mathbb{E}(\|f^\star-f_\theta\|^2_{L^2(\mathbb{P}_X)}).$$

Conclusion. Putting everything together, we deduce from inequality (5) that

$$\begin{aligned}
\mathbb{E}(\|f_{\hat\theta}-f^\star\|^2_{L^2(\mathbb{P}_X)})\le\inf_{\theta\in H_m}&\Big((2+6(1+\frac{\alpha^2}{\lambda n})^2)\mathbb{E}(\|f^\star-f_\theta\|^2_{L^2(\mathbb{P}_X)})+\lambda(1+\frac{\alpha^2}{\lambda n})^2\|R^{1/2}\theta\|_2^2\Big)\\
&+\sigma^2n^{-1}(1+\frac{\alpha^2}{\lambda n})(2m+1)^d,
\end{aligned}$$

where

- $\alpha^2=\sum_{k\in\mathbb{Z}^d}\frac{1}{1+\|k\|_2^{2s}}<\infty$ for the Sobolev regression, while $\alpha^2=(2m+1)^d$ for the L^2-Fourier regression;
- $\|R^{1/2}\theta\|_2^2=\|S\theta\|_2^2\le C_\Omega\|f_\theta\|^2_{H^s(\Omega)}$ for the Sobolev regression, and $\|R^{1/2}\theta\|_2^2=\|\theta\|_2^2$ for the L^2-Fourier regression.

Convergence rates. According to Proposition A.2, there is a parameter $\theta(f^\star)$ such that $f^\star(x)=\langle\phi(x),\theta(f^\star)\rangle$. Let $\theta^\star\in\mathbb{C}^{(2m+1)^d}$ be the truncation of $\theta(f^\star)$, i.e., $\theta^\star_k=\theta(f^\star)_k$. Proposition 2.1 states that the approximation error term $\mathbb{E}(\|f_{\theta^\star}-f^\star\|^2_{L^2(\mathbb{P}_X)})=\mathcal{O}(n^{-\gamma})$, where $\gamma=2s/(2s+d)$. Moreover, the RKHS norm of $\theta^\star$ is bounded by $\max(\|\theta^\star\|_2^2,\|S\theta^\star\|_2^2)\le\sum_{k\in\mathbb{C}^{\mathbb{Z}^d}}(1+\|k\|_2^{2s})|\theta(f^\star)_k|^2\le C_\Omega\|f^\star\|^2_{H^s(\Omega)}$. Therefore, in the Sobolev regression, setting $\lambda=\mathcal{O}(n^{-2s/(2s+d)})$ such that $\lambda\ge n^{-1}$, and taking $m=n^{1/(2s+d)}$ leads to Sobolev minimax rate. Moreover, in the L^2-Fourier regression, setting $\lambda=\Theta(n^{-2s/(2s+d)})$ and taking $m=n^{1/(2s+d)}$ leads to Sobolev minimax rate.

C.3. Proof of Propositions 3.2 and 3.4

As in the previous subsection, we let $\hat\Sigma=n^{-1}\Phi^\star\Phi$ be the empirical covariance matrix and R be the regularization matrix: $R=S^2$ in the Sobolev regression, and $R=I$ for the L^2-Fourier regression.

Conjugate gradient convergence speed. Let $A=\hat\Sigma+\lambda R$, and $\hat\theta=A^{-1}\Phi^*\mathbb{Y}/n$, it is well known (see, e.g., Nocedal & Wright, 1999, Equation 5.35) that the conjugate gradient method estimates x_k converge at the speed

$$\|A^{1/2}(\theta_k-\hat\theta)\|_2\le\|A^{1/2}(\theta_0-\hat\theta)\|_2\left(\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1}\right)^{2k}\le\|A^{1/2}(\theta_0-\hat\theta)\|_2\exp(-2k/\sqrt{\kappa(A)}),$$

where $\kappa(A)=\lambda_{\max}(A)/\lambda_{\min}(A)$ is the conditioning number of A (i.e., the ratio of its highest and its lowest eigenvalues) and $\theta_0=\Phi^*\mathbb{Y}/n$. To apply this result to our setting, we thus need to bound the spectrum of the matrix $A=\hat\Sigma+\lambda R$.

Lower bounding the eigenvalues of $\hat{\Sigma} + \lambda R$. Since $R \geq I$, we have $\lambda_{\min}(\hat{\Sigma} + \lambda R) \geq \lambda = n^{-2s/(2s+d)}$.

Upper bounding the eigenvalues of $\hat{\Sigma} + \lambda R$. On the other hand, the empirical covariance matrix can be written as $\hat{\Sigma} = \Phi^* \Phi / n = \sum_{j=1}^n \phi(X_j)\phi(X_j)^*/n$. Thus, we have that $\hat{\Sigma} - \Sigma = \sum_{j=1}^n (\phi(X_j)\phi(X_j)^* - \Sigma)/n$ satisfies

- $(\phi(X_j)\phi(X_j)^* - \Sigma)/n$ is symmetric
- $\mathbb{E}((\phi(X_j)\phi(X_j)^* - \Sigma)/n) = 0$
- the highest eigenvalue of $\lambda_{\max}((\phi(X_j)\phi(X_j)^* - \Sigma)) \leq \lambda_{\max}(\phi(X_j)\phi(X_j)^*) = \max_{\|\theta\|_2=1} |\langle \phi(X_i), \theta \rangle|^2 \leq \|\phi(X_i)\|_2^2 = (2m+1)^d$
- the highest eigenvalue of $\lambda_{\max}(\Sigma - (\phi(X_j)\phi(X_j)^*)) \leq \lambda_{\max}(\Sigma) = \max_{\|\theta\|_2=1} \theta^* \Sigma \theta = \max_{\|\theta\|_2=1} \mathbb{E}(|\langle \phi(X), \theta \rangle|^2) \leq \kappa \int_{[-2L,2L]^d} |\langle \phi(x), \theta \rangle|^2 dx \leq \kappa (4L)^d$, where κ is the constant in Assumption (v)
- $\mathbb{E}(((\phi(X_j)\phi(X_j)^* - \Sigma)/n)^2) = n^{-2}(\mathbb{E}((\phi(X_j)\phi(X_j)^*)^2) - \Sigma^2) = n^{-2}(\mathbb{E}((2m+1)^d(\phi(X_j)\phi(X_j)^*) - \Sigma^2)$. Thus, $\lambda_{\max}(\sum_{j=1}^n \mathbb{E}(((\phi(X_j)\phi(X_j)^* - \Sigma)/n)^2)) \leq n^{-2} \sum_{j=1}^n \lambda_{\max}(\mathbb{E}((2m+1)^d(\phi(X_j)\phi(X_j)^*)^2)) = (2m+1)^{2d}/n$.

Thus, the Bernstein inequality (see, e.g., Bach, 2024, Proposition 1.7) applied to $\hat{\Sigma} - \Sigma$ states that

$$\mathbb{P}\big(\lambda_{\max}(\hat{\Sigma} - \Sigma) \geq t\big) \leq (2m+1)^d \exp(-\frac{nt^2}{2(2m+1)^{2d}/n + 2(2m+1)^d t/3}).$$

Since $m^d \leq n$, we deduce that,

$$\mathbb{P}\big(\lambda_{\max}(\hat{\Sigma} - \Sigma) \geq t\big) \leq n \exp\big(-\frac{t^2}{3^{d+1}(1+t/3)}\big)$$

and so for $t \geq 3$, $t \geq 1 + t/3$ and

$$\mathbb{P}\big(\lambda_{\max}(\hat{\Sigma} - \Sigma) \geq t\big) \leq n \exp(-3^{-d-1} t).$$

We deduce that

$$\mathbb{P}\big(\lambda_{\max}(\hat{\Sigma} - \Sigma) \geq 3^{d+1}(\log n - \log \varepsilon)\big) \leq \varepsilon.$$

Overall, we deduce that, with probability at least $1 - \varepsilon$, we have that

$$\lambda_{\max}(\hat{\Sigma} + \lambda R) \leq \lambda_{\max}(\hat{\Sigma}) + \lambda \lambda_{\max}(R) \leq \lambda_{\max}(\Sigma) + \lambda_{\max}(\hat{\Sigma} - \Sigma) + \lambda \lambda_{\max}(R) \leq \kappa (4L)^d + 3^{d+1}(\log n - \log \varepsilon) + \lambda \lambda_{\max}(R).$$

Since $\lambda = n^{-2s/(2s+d)}$ and $\lambda_{\max}(R) = 1$ if $R = I$ or $\lambda_{\max}(R) = \|(m, \ldots, m)\|_2^{2s} = d^s n^{2s/(2s+d)}$ if $R = S^2$, we deduce that with probability at least $1 - \varepsilon$, we have that

$$\lambda_{\max}(\hat{\Sigma} + \lambda R) \leq \kappa (4L)^d + 3^{d+1}(\log n - \log \varepsilon) + d^s.$$

Conjugate gradient stopping. Overall, we deduce that, there is a constant $C > 0$ depending only on Ω and κ such that, with probability at least $1 - \varepsilon$, $\lambda_{\max}(A) \leq C(\log n - \log \varepsilon)$ and

$$\kappa(\hat{\Sigma} + \lambda R) \leq C n^{2s/(2s+d)}(\log n - \log \varepsilon).$$

Thus, with probability at least $1 - \varepsilon$, after

$$k_{\text{stop}} = 2\sqrt{C n^{2s/(2s+d)}(\log n - \log \varepsilon)}(\log n + 0.25 \log \|A^{1/2}(\Phi^* \mathbb{Y}/n - \hat{\theta})\|_2)$$

steps, the conjugate gradient estimate θ_k is such that

$$\|A^{1/2}(\theta_k - \hat{\theta})\|_2 \leq \|A^{1/2}(\Phi^* \mathbb{Y}/n - \hat{\theta})\|_2 \exp(-4(\log n + 0.25 \log \|A^{1/2}(\Phi^* \mathbb{Y}/n - \hat{\theta})\|_2)) = 1/n^4.$$

Setting $\varepsilon = \exp(-n^{2s/(2s+d)})$ leads to

$$k_{\text{stop}} = 2\sqrt{2C n^{4s/(2s+d)}}(\log n + 0.25 \log \|A^{1/2}(\hat{\theta} - \Phi^* \mathbb{Y}/n)\|_2)$$

Then, notice that $\|A^{1/2}\hat{\theta}\|_2^2 = \|A^{-1/2}\Phi^*\mathbb{Y}/n\|_2^2 \leq \|\hat{\Sigma}^{1/2}\Phi^*\mathbb{Y}/n\|_2^2 = \mathbb{Y}^*\Phi(\Phi^*\Phi)^{-1}\Phi^*\mathbb{Y}/n \leq \|\mathbb{Y}\|_2^2/n$, since $\Phi(\Phi^*\Phi)^{-1}\Phi^*$ is the orthogonal projection on $Im(\Phi)$.

Moreover, we have placed ourselves in the event where $\lambda_{\max}(\hat{\Sigma}) \leq \lambda_{\max}(A) \leq C(\log n - \log \varepsilon) \leq 2Cn^{2s/(2s+d)}$. Thus, we deduce that

$$\|A^{1/2}\Phi^*\mathbb{Y}/n\|_2^2 \leq \lambda_{\max}(A)\|\Phi^*\mathbb{Y}/n\|_2^2 \leq \lambda_{\max}(A)\lambda_{\max}(\hat{\Sigma})\|n^{-1/2}\mathbb{Y}\|_2^2 \leq (2Cn^{2s/(2s+d)})^2\|Y\|_2^2/n \leq 4C^2n\|Y\|_2^2.$$

Combining the last two results, we deduce that $\log\|A^{1/2}(\hat{\theta} - \Phi^*\mathbb{Y}/n))\|_2 \leq \log(\|A^{1/2}\hat{\theta}\|_2 + \|A^{1/2}\Phi^*\mathbb{Y}/n\|_2) \leq 0.5\log(8n\|\mathbb{Y}\|_2^2C^2)$. We deduce that there is a constant C_2, depending only on Ω and κ such that

$$k_{\text{stop}} \leq C_2 n^{2s/(2s+d)}(\log n + \log(\|\mathbb{Y}\|_2)).$$

Thus, after k_{stop} conjuguate gradient steps, we have that $\|A^{1/2}(\theta_k - \hat{\theta})\|_2 \leq 1/n^4$ and so

$$\|A(\theta_k - \hat{\theta})\|_2 \leq \lambda_{\max}(A)/n^4 \leq C(\log n - \log \varepsilon)/n^4 \leq 2Cn^{2s/(2s+d)-4}.$$

If $n \geq 2C^{1+d/2s}$, this means that the criterion $\|A(\theta_k - \hat{\theta})\|_2 \leq 1/n^3$ is reached before k_{stop} steps.

Since $R \geq I$, we deduce that $A \geq \lambda I$ and that $\|\theta_k - \hat{\theta}\|_2 \leq \|A(\theta_k - \hat{\theta})\|_2/\lambda^2 \leq 1/n^3/\lambda^2 \leq 1/n$. Since each step of the conjuguate gradient descent has a complexity of $O(n^{d/(2s+d)})$, the algorithm has a complexity $O(k_{\text{stop}}n^{d/(2s+d)}) = O(n^{(d+2s)/(2s+d)}\log(n)^{3/2}) = O(n\log(n))$ and reaches a precision $\|\theta_k - \hat{\theta}\|_2 \leq 1/n^2/\lambda \leq 1/n$.

Minimax rate. Finally, we have that $\mathbb{E}(\|f_{\theta_k} - f^\star\|_{L^2(\mathbb{P}_X)}^2) \leq 2\mathbb{E}(\|f_{\theta_k} - f_{\hat{\theta}}\|_{L^2(\mathbb{P}_X)}^2) + 2\mathbb{E}(\|f_{\hat{\theta}} - f^\star\|_{L^2(\mathbb{P}_X)}^2)$. Since

(i) $\mathbb{E}(\|f_{\hat{\theta}} - f^\star\|_{L^2(\mathbb{P}_X)}^2) = \mathcal{O}(n^{-2s/(2s+d)})$ from Propositions 3.1 and 3.3,

(ii) $\mathbb{E}(\|f_{\theta_k} - f_{\hat{\theta}}\|_{L^2(\mathbb{P}_X)}^2) = \mathbb{E}(\|\sqrt{\Sigma}(\theta_k - \hat{\theta})\|_2^2) \leq \kappa(4L)^d\mathbb{E}(\|\theta_k - \hat{\theta}\|_2^2) \leq \kappa(4L)^d/n^2$ from Assumption (v) and since $\|\theta_k - \hat{\theta}\|_2 \leq 1/n$,

we deduce that $\mathbb{E}(\|f_{\theta_k} - f^\star\|_{L^2(\mathbb{P}_X)}^2) = \mathcal{O}(n^{-2s/(2s+d)})$.

C.4. Proof of Proposition 5.1 (and its extended version Proposition A.3)

We rely on the same tools as in the proof of Propositions 3.1 and 3.3. The key is to remark that the empirical risk

$$R(\theta) = \|\Phi_1\theta_1 + \cdots + \Phi_d\theta_d - \mathbb{Y}\|_2^2 + \lambda\|\theta\|_2^2$$

can be written as

$$R(\theta) = \|\Phi\theta - \mathbb{Y}\|_2^2 + \lambda\|\theta\|_2^2$$

where $\Phi = (\Phi_1 \mid \cdots \mid \Phi_d)$ and $\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{C}^{d(2m+1)}$. By setting $\varphi(x) = (\phi(x_1), \ldots, \phi(x_d)) \in \mathbb{C}^{d(2m+1)}$, we have that $f_\theta(x) = \langle\varphi(x), \theta\rangle$. The empirical covariance matrix is then $\hat{\Sigma} = \Phi^*\Phi/n = n^{-1}\sum_{j=1}^n \varphi(X_j)\varphi(X_j)^*$.

Risk decomposition. As in the proof of Propositions 3.1 and 3.3, we have, for any $\theta \in \mathbb{C}^{d(2m+1)}$,

$$\mathbb{E}(\|f_{\hat{\theta}} - f^\star\|_{L^2(\mathbb{P}_X)}^2) \leq 2\mathbb{E}(\|f_{\hat{\theta}} - f_\theta\|_{L^2(\mathbb{P}_X)}^2) + 2\mathbb{E}(\|f_\theta - f^\star\|_{L^2(\mathbb{P}_X)}^2),$$

and

$$\begin{aligned}\mathbb{E}(\|f_{\hat{\theta}} - f_\theta\|_{L^2(\mathbb{P}_X)}^2) \leq{}& 3\mathbb{E}(\|\sqrt{\Sigma}(\hat{\Sigma} + \lambda)^{-1}\lambda\theta\|_2^2) + 3\mathbb{E}(\|\sqrt{\Sigma}(\hat{\Sigma} + \lambda)^{-1}\Phi^*\epsilon/n\|_2^2) \\ &+ 3\mathbb{E}(\|\sqrt{\Sigma}(\hat{\Sigma} + \lambda)^{-1}\Phi^*\delta\mathbb{Y}/n\|_2^2),\end{aligned}$$

where $\Sigma = \mathbb{E}(\phi(X)\phi(X)^*)$. Similarly, the bias term is bounded by

$$\mathbb{E}(\|\sqrt{\Sigma}(\hat{\Sigma} + \lambda)^{-1}\lambda\theta\|_2^2) \leq \lambda(1 + \frac{\alpha^2}{\lambda n})^2\|\theta\|_2^2,$$

where $\alpha = \max_{x\in\mathbb{R}^d} \|\phi(x)\|_2 = \sqrt{d(2m+1)}$. As for the variance term, it is bounded by

$$\mathbb{E}(\|\sqrt{\Sigma}(\hat{\Sigma}+\lambda)^{-1}\Phi^*\varepsilon/n\|_2^2) \leq \sigma^2 n^{-1}(1+\frac{\alpha^2}{\lambda n})\mathrm{tr}(\mathbb{E}((\Sigma+\lambda)^{-1}\Sigma)) \leq \sigma^2\frac{d(2m+1)}{n}(1+\frac{\alpha^2}{\lambda n}).$$

Finally, the approximation term is bounded by

$$\mathbb{E}(\|\sqrt{\Sigma}(\hat{\Sigma}+\lambda R)^{-1}\Phi^*\delta\mathbb{Y}/n\|_2^2) \leq (1+\frac{\alpha^2}{\lambda n})^2\,\mathbb{E}(\|f^\star - f_\theta\|^2_{L^2(\mathbb{P}_X)}).$$

Putting everything together, we obtain

$$\begin{aligned}\mathbb{E}(\|f_{\hat{\theta}} - f^\star\|^2_{L^2(\mathbb{P}_X)}) \leq \inf_{\theta\in\mathbb{C}^{d(2m+1)}} &\Big((2+6(1+\frac{\alpha^2}{\lambda n})^2)\mathbb{E}(\|f^\star - f_\theta\|^2_{L^2(\mathbb{P}_X)}) + \lambda(1+\frac{\alpha^2}{\lambda n})^2\|\theta\|_2^2\Big)\\ &+\sigma^2\frac{\alpha^2}{n}(1+\frac{\alpha^2}{\lambda n}).\end{aligned}$$

Convergence rate. The convexity of the function $x \mapsto x^2$ leads to

$$\mathbb{E}(\|f^\star - f_\theta\|^2_{L^2(\mathbb{P}_X)}) = \mathbb{E}(\|\sum_{\ell=1}^d g_\ell^\star - g_{\theta_\ell}\|^2_{L^2(\mathbb{P}_X)}) \leq d\sum_{\ell=1}^d \mathbb{E}(\|g_\ell^\star - g_{\theta_\ell}\|^2_{L^2(\mathbb{P}_X)}),$$

where $g_{\theta_\ell}(x) = \langle\phi_\ell(x), \theta_\ell\rangle$. According to Proposition A.2, taking $m = n^{1/(2s+1)}/d$ ensures that

$$\inf_{\theta_\ell\in\mathbb{C}^{2m+1}}\Big(\mathbb{E}(\|g_\ell^\star - g_{\theta_\ell}\|^2_{L^2(\mathbb{P}_X)}) + n^{-2s/(2s+1)}\|\theta_\ell\|_2^2\Big) = \mathcal{O}(n^{-2s/(2s+1)}).$$

Thus, choosing $\lambda = n^{-2s/(2s+1)}$ and $m = n^{1/(2s+1)}/d$, we are led to

$$\inf_{\theta\in\mathbb{C}^{d(2m+1)}}\Big((2+6(1+\frac{\alpha^2}{\lambda n})^2)\mathbb{E}(\|f^\star - f_\theta\|^2_{L^2(\mathbb{P}_X)}) + \lambda(1+\frac{\alpha^2}{\lambda n})^2\|\theta\|_2^2\Big) = \mathcal{O}(n^{-2s/(2s+1)}).$$

We conclude that

$$\mathbb{E}(\|f_{\hat{\theta}} - f^\star\|^2_{L^2(\mathbb{P}_X)}) = \mathcal{O}(n^{-2s/(2s+1)}).$$