IEEE Control Systems Letters paper presented at
2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Error analysis of regularized trigonometric linear regression with unbounded sampling: a statistical learning viewpoint

Anna Scampicchio, Elena Arcari, Melanie N. Zeilinger

Abstract— **The effectiveness of non-parametric, kernel-based methods for function estimation comes at the price of high computational complexity, which hinders their applicability in adaptive, model-based control. Motivated by approximation techniques based on sparse spectrum Gaussian processes, we focus on models given by regularized trigonometric linear regression. This paper provides an analysis of the performance of such an estimation set-up within the statistical learning framework. In particular, we derive a novel bound for the sample error in finite-dimensional spaces, accounting for noise with potentially unbounded support. Next, we study the approximation error and discuss the bias-variance trade-off as a function of the regularization parameter by combining the two bounds.**

I. INTRODUCTION

Non-parametric approaches for regularized function estimation are a key tool in machine learning, and have been successfully applied to, e.g., system identification [1] and learning-based control [2], [3]. Nevertheless, their applicability in real-time scenarios is hindered by their high computational complexity, which scales as $\mathcal{O}(N^3)$, with N being the data-set cardinality. The strategies proposed to enable fast adaptation of kernel-based methods can be grouped into two main categories: input location selection, and low-rank approximations of the kernel [4]. In this second class of approaches, a vast success was achieved by sparse spectrum Gaussian processes [5], [6], where operations on the (shift-invariant) kernel yield a parametric approximation by means of linear combinations of Fourier features.

In this paper, we draw inspiration from the latter method and consider regularized regression within a finite-dimensional hypothesis space $\mathscr{H}$ defined by a span of $E \ll N$ predefined trigonometric functions. Such a set-up yields an estimator whose computation scales as $\mathcal{O}(E^2 N)$, relaxes the assumption of having shift-invariant kernels in the derivation of random Fourier features, and benefits from the use of regularization without resorting to purely non-parametric approaches. The focus of this work is to provide an error analysis for this parametric regularized estimator. The results derive non-asymptotic, non-conservative bounds in view of obtaining reliable guarantees for data-driven, model-based control schemes that leverage such a model (see, e.g., [7], [8], [9]). Such an analysis has been performed for standard, projection-based estimators (for an overview see, e.g., [10] and references therein); however, the available results do not account for regularization, which acts against potential basis function misspecification in a more robust way compared with discrete model-order selection rules [11]. On the other hand, the available error bounds for non-parametric estimators (which do rely on regularization to ensure well-posedness) hold for infinite-dimensional hypothesis spaces and result therefore conservative in our set-up. Note also that trigonometric functions appear as a spline basis in non-parametric estimation when considering the Sobolev space $\mathscr{W}_1^2 = \{f : f \text{ is absolutely continuous, and } f^{(1)} \in \mathscr{L}^2\}$ if the inputs are equally spaced in the domain [12].

We frame our error analysis in the statistical learning set-up [13], [14]. The function to be estimated (i.e., the *regression function* f_ρ) is defined as the minimizer of the expected risk over a (partially) unknown probability distribution, jointly defined over the input-output spaces, and from which i.i.d. samples are drawn. Consequently, this formulation can also handle fully nonlinear measurement models. Furthermore, f_ρ is generally assumed to belong to the space of square-integrable functions $\mathscr{L}^2$, and the hypothesis space is typically taken as an infinite-dimensional Reproducing Kernel Hilbert Space (RKHS), which is related to $\mathscr{L}^2$ by interpolation spaces arguments ([15, Theorem 2]). Differently from classic non-parametric set-ups, the regression function is not assumed to belong to the hypothesis space. Thus, two objects can be therein defined: the actual data-based estimate f_z and its data-free limit $f_{\mathscr{H}}$. The goal of error analysis consists in quantifying the approximation error, or *bias*, $\|f_\rho - f_{\mathscr{H}}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$, and the sample error, or *variance*, $\|f_{\mathscr{H}} - f_z\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$ [13]. As regards the latter, results abound in the statistical learning literature. Most of them deal with probability measures on the outputs that have bounded support, and thus obtain bounds leveraging concentration inequalities such as Hoeffding's or Bennett's [16], [14, Chapter 3.1]. Works in this direction are, e.g., [17], [14], [18], [19], [20]. Contributions considering unbounded sampling include [21], [22], [23]. The bounds therein derived leverage the so-called moment hypothesis, which relaxes the boundedness assumption, and holds also for (sub-)Gaussian noises. Such results rely on the computation of covering numbers quantifying the capacity of the hypothesis space [24] and showcase optimal rates of convergence; nevertheless, they tend to be of limited practical relevance in the non-asymptotic case due to the large values of the multiplicative coefficients, which are often furthermore difficult to compute. Other non-conservative bounds obtained without using concentration inequalities are given, e.g., in [25]; however, their practical use is limited by the involved constants, which are generally hard to compute.

The Authors are affiliated with the Institute of Dynamic Systems and Control, ETH Zürich. {ascampicc,earcari,mzeilinger}@ethz.ch.

In this work, we perform error analysis for finite-dimensional hypothesis spaces given by trigonometric functions. Our first contribution is a sample error bound, which is less conservative than those available in the literature even if it accounts also for noises with unbounded supports. Our second contribution consists in studying the bias-variance trade-off of the regularized trigonometric regression set-up. To this end, we obtain two bounds on the approximation error, combine them with the sample complexity result and analyze the conditions ensuring the existence of a unique value of the regularization parameter γ returning the optimal trade-off. The differences between the two approaches are investigated in a Monte Carlo study, which shows that one of the two criteria returns a value of γ that captures the oracle behavior (i.e., minimizing the overall error), thus leading to fast estimation schemes that do not need preliminary hyper-parameter selection.

II. PROBLEM SET-UP

Let the metric space of inputs $\mathscr{X}$ be a compact subset of $\mathbb{R}$: without loss of generality, we take $\mathscr{X} = [-X/2,\, X/2]$ for some $X \in \mathbb{R}_+$ (the scalar case is presented just for ease of visualization: the multi-dimensional is a straightforward extension). The output space is assumed to be $\mathscr{Y} = \mathbb{R}$. There is a probability measure ρ defined over $\mathscr{Z} \doteq \mathscr{X} \times \mathscr{Y}$ that decomposes into $\rho_{\mathscr{X}}(x)$ and $\rho(y|x)$ according to Fubini's Theorem. In the considered set-up, the probability measure defined on $\mathscr{X}$ is the standard uniform: denoting with μ the Lebesgue measure, we have that $\rho_{\mathscr{X}}(A) = \mu(A \cap \mathscr{X})/\mu(\mathscr{X}) = \mu(A \cap \mathscr{X})/X$ for any set A in the $\sigma-$algebra of interest. In this way, $\rho_{\mathscr{X}}$ is a Borel non-degenerate, $\sigma-$finite measure. As regards $\rho(y|\cdot)$, we assume it is unknown and defined over $\mathbb{R}$.

Having N independent samples drawn from ρ collected in the data-set $\mathcal{D} \doteq \{(x_t,\, y_t)\}_{t=1}^N$, the goal is to estimate the regression function

$$f_\rho(x) \doteq \int_{\mathscr{Y}} y d\rho(y|x). \tag{1}$$

We make use of the following Assumption.

Assumption 1. *The regression function f_ρ belongs to the space of square-integrable functions on $\mathscr{X}$, denoted by $\mathscr{L}^2_{\rho_{\mathscr{X}}}$, and is such that $\|f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} = \sqrt{\int_{\mathscr{X}} f^2(x) d\rho_{\mathscr{X}}(x)} = \sqrt{\int_{\mathscr{X}} f^2(x) d\mu(x)/X} \leq B_f$. Moreover, we also have that $\sigma_\rho^2 \doteq \int_{\mathscr{X}} \sigma_\rho^2(x) d\rho_{\mathscr{X}}(x) = \int_{\mathscr{Z}} (y - f_\rho(x))^2 d\rho \leq B_\sigma^2$.*

In other words, we assume to have access to bounds on the energy of the unknown function to be estimated, and on the variance of the additive noises.

The space $\mathscr{L}^2_{\rho_{\mathscr{X}}}$ is a separable Hilbert space whose complete orthonormal basis by means of trigonometric functions [26] is given by

$$\left\{ \sqrt{2} \sin\left(\frac{2\pi q}{X} x\right), \sqrt{2} \cos\left(\frac{2\pi q}{X} x\right) \right\}_{q \in \mathbb{N}} \tag{2}$$

$$\doteq \{\bar{\varphi}_q^s(x),\, \bar{\varphi}_q^c(x)\}_{q \in \mathbb{N}} \quad \text{with } x \in \mathscr{X}. \tag{3}$$

Accordingly, any function $f \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$ can be expressed as $f(\cdot) = \sum_{q \in \mathbb{N}} (\alpha_q^s \bar{\varphi}_q^s(\cdot) + \alpha_q^c \bar{\varphi}_q^c(\cdot))$, which will be also compactly written as $f(\cdot) = \sum_{q \in \mathbb{N}} \alpha_q \bar{\varphi}_q(\cdot)$, with $\sum_{q \in \mathbb{N}} \alpha_q^2 < \infty$. Within this representation, we denote the target function as $f_\rho(\cdot) = \sum_{q \in \mathbb{N}} (\bar{\alpha}_q^s \bar{\varphi}_q^s(\cdot) + \bar{\alpha}_q^c \bar{\varphi}_q^c(\cdot)) = \sum_{q \in \mathbb{N}} \bar{\alpha}_q \bar{\varphi}_q(\cdot)$.

Function estimation in $\mathscr{L}^2_{\rho_{\mathscr{X}}}$ cannot be performed, because pointwise evaluation is not well defined. Therefore, we perform such a task within a hypothesis space having the structure of an RKHS. Specifically, we consider the RKHS obtained from a subset of functions in (3) with cardinality E, where E is chosen according to the computational capacity. Denote by Q the set of selected frequencies, i.e., $Q = \{q_j\}_{j=1}^{E/2} \subset \mathbb{N}$, and consider the following functions extracted from (3) using Q defined, for $j = 1, ..., E/2$, as

$$\varphi_i(x) \doteq \begin{cases} \bar{\varphi}_{q_j}^s(x), \; i = j \\ \bar{\varphi}_{q_j}^c(x), \; i = j + \frac{E}{2}. \end{cases} \tag{4}$$

Then, the RKHS of interest is the one induced by the following kernel:

$$\mathscr{K}(x_a, x_b) = \phi^\top(x_a) \Sigma_\alpha \phi(x_b), \tag{5}$$

where $\Sigma_\alpha \doteq \mathrm{diag}(\lambda_1, ..., \lambda_E)$ is a positive definite matrix, and the vector $\phi(\cdot) \in \mathbb{R}^E$ is such that $\phi^\top(x) = [\varphi_1(x) \; \dots \; \varphi_E(x)]$. Clearly, (5) is a Mercer kernel ([27, (6), p.346]; it satisfies Mercer's condition $\int_{\mathscr{X}} \int_{\mathscr{X}} \mathscr{K}(x, x')^2 d\rho_{\mathscr{X}}(x) d\rho_{\mathscr{X}}(x') = \sum_{i=1}^E \lambda_i^2$, and it is non-stationary if and only if $\lambda_i \neq \lambda_{i+E/2}$ for all $i = 1, \dots, E/2$. Furthermore, using the argument in [28, Chapter 4.3]), it holds that

$$\begin{aligned} C_{\mathscr{X}} &\doteq \sup_{x_a, x_b \in \mathscr{X}} \sqrt{\mathscr{K}(x_a, x_b)} \\ &\leq \sqrt{\sum_{i=1}^{E/2} \max\{\lambda_i, , \lambda_{i+E/2}\}} < +\infty. \end{aligned} \tag{6}$$

Being a Mercer kernel, we have from Moore-Aronszajn Theorem [27] that $\mathscr{K}$ is in one-to-one correspondence with the Hilbert space of functions $(\mathscr{H}, \langle \cdot, \cdot \rangle_{\mathscr{H}})$, which is

$$\mathscr{H} = \{f \in \mathscr{L}^2_{\rho_{\mathscr{X}}} : f(\cdot) = \phi^\top(\cdot)\alpha,\; \alpha \in \mathbb{R}^E\} \tag{7}$$

with inner product given, for $f^{(\natural)}(\cdot) = \phi^\top(\cdot)\alpha^{(\natural)}$ and $\natural = a,\, b$:

$$\langle f^{(a)},\, f^{(b)} \rangle_{\mathscr{H}} = \langle \Sigma_\alpha^{-1/2} \alpha^{(a)}, \Sigma_\alpha^{-1/2} \alpha^{(b)} \rangle_2. \tag{8}$$

Within the hypothesis space, we can compute the estimate from the data set $\mathcal{D}$ as follows. Consider the sampling operator $\mathcal{S}_{\mathscr{X}} : \mathscr{H} \to \mathbb{R}^N$ such that $\mathcal{S}_{\mathscr{X}}(f) = [f(x_1) \; \dots \; f(x_N)]^\top$, together with its adjoint $\mathcal{S}_{\mathscr{X}}^\top : \mathbb{R}^N \to \mathscr{H}$ yielding $\mathcal{S}_{\mathscr{X}}^\top c = \sum_{t=1}^N c_t \mathscr{K}(x_t, \cdot)$. Thus, considering $Y = [y_1, ..., y_N]^\top$ and regularization parameter $\gamma > 0$, we have

$$f_z \doteq \arg\min_{f \in \mathscr{H}} \frac{1}{N} \sum_{t=1}^N (y_t - f(x_t))^2 + \gamma \|f\|_{\mathscr{H}}^2 \tag{9}$$

$$= \left(\frac{1}{N} \mathcal{S}_{\mathscr{X}}^\top \mathcal{S}_{\mathscr{X}} + \gamma I\right)^{-1} \frac{1}{N} \mathcal{S}_{\mathscr{X}}^\top Y. \tag{10}$$

The aim of error analysis is to quantify the discrepancy between f_z and f_ρ. To this end, we additionally consider the data-free limit of (9) as

$$f_{\mathscr{H}} \doteq \arg\min_{f \in \mathscr{H}} \int_{\mathscr{X}} (f(x) - f_\rho(x))^2 d\rho_{\mathscr{X}}(x) + \gamma \|f\|^2_{\mathscr{H}} \tag{11}$$

$$= (L_{\mathscr{K}} + \gamma I)^{-1} L_{\mathscr{K}} f_\rho, \tag{12}$$

where $L_{\mathscr{K}}(f)(\bar{x}) \doteq \int_{\mathscr{X}} \mathscr{K}(\bar{x}, x) f(x) d\rho_{\mathscr{X}}(x)$ is an integral operator which, thanks to the properties of $\mathscr{K}$, is (a) is self-adjoint and strictly positive, (b) continuous and compact, (c) satisfies the Spectral Theorem [14, Theorem 4.3] with eigenpairs $\{(\varphi_i(\cdot), \lambda_i)\}_{i=1}^E$. Thanks to these properties, given an arbitrary $\mathscr{L}^2_{\rho_{\mathscr{X}}}$ function $f(x) = \sum_{q \in \mathbb{N}} \alpha_q \bar{\varphi}_q(x)$, using linearity and orthonormality of the basis, we have

$$L_{\mathscr{K}}(f)(\bar{x}) = \sum_{i=1}^E \lambda_i \alpha_i^\pi \varphi_i(\bar{x}), \tag{13}$$

where we define the $i-$th component of the vector α^π for $i = 1, ..., E$, along the lines of (4), as follows:

$$\text{For } j = 1, ..., \frac{E}{2}, \quad \alpha_i^\pi \doteq \begin{cases} \alpha^s_{q_j}, & i = j \\ \alpha^c_{q_j}, & i = j + E/2. \end{cases} \tag{14}$$

Moreover, thanks to property (a), we can also define the r-th power of the integral operator[1] as [13]:

$$L^r_{\mathscr{K}}(f)(\bar{x}) = \sum_{i=1}^E \lambda_i^r \alpha_i^\pi \varphi_i(\bar{x}). \tag{15}$$

In the following, we study the sample error $\|f_z - f_{\mathscr{H}}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$ introduced by the finiteness of the data-set $\mathcal{D}$, and the approximation error $\|f_{\mathscr{H}} - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$ determined by the choice of the hypothesis space. The two bound the overall error as $\|f_z - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} \leq \|f_z - f_{\mathscr{H}}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} + \|f_{\mathscr{H}} - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$, which is to be minimized as a function of the regularization parameter γ.

III. SAMPLE ERROR

In this Section, we provide the novel result concerning the error between f_z and $f_{\mathscr{H}}$ introduced in (10) and (12).

Theorem 1. *Let Assumption 1 hold. Consider $C_{\mathscr{K}}$ as introduced in* (6)*, and define $\check{\lambda} \doteq \min_{i=1,...,E} \lambda_i$. Then, with confidence at least $1 - \delta$, it holds that*

$$\|f_z - f_{\mathscr{H}}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} \leq \frac{C^3_{\mathscr{K}}}{\gamma} \sqrt{\frac{B_f^2 + B_\sigma^2}{\check{\lambda} N \delta}}. \tag{16}$$

Proof. Defining $\xi_t : \mathscr{X} \to \mathscr{H}$ such that $\xi_t(\cdot) \doteq (y_t - f_{\mathscr{H}}(x_t))\mathscr{K}(x_t, \cdot)$, it holds that $\mathbb{E}_{\mathscr{X}}[\xi_t](\cdot) = L_{\mathscr{K}}(f_\rho - f_{\mathscr{H}})(\cdot) = \gamma f_{\mathscr{H}}(\cdot)$. From this, and recalling the definition of sampling operator, it follows that $f_z(x) - f_{\mathscr{H}}(x)$ equals [15]

$$\left(\frac{1}{N} \mathcal{S}^\top_{\mathscr{X}} \mathcal{S}_{\mathscr{X}} + \gamma I\right)^{-1} \left[\frac{1}{N} \sum_{t=1}^N \xi_t(x) - \mathbb{E}_{\mathscr{X}}[\xi](x)\right].$$

We can now study the $\mathscr{L}^2_{\rho_{\mathscr{X}}}-$ norm of the expression above. Since $\mathscr{X}$ is compact and the measure $\rho_{\mathscr{X}}$ on it defined is a probability measure, $\|f\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} \leq \|f\|_\infty$ for any function $f \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$: therefore, $\|f_z - f_{\mathscr{H}}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$ is upper bounded by

$$\left\| \left(\frac{1}{N} \mathcal{S}^\top_{\mathscr{X}} \mathcal{S}_{\mathscr{X}} + \gamma I\right)^{-1} \right\|_\infty \left\| \frac{1}{N} \sum_{t=1}^n \xi_t - \mathbb{E}_{\mathscr{X}}[\xi] \right\|_\infty.$$

Since the operator norm can be bounded by $\frac{C_{\mathscr{K}}}{\gamma \sqrt{\check{\lambda}}}$ (the proof is reported at the end of this subsection), we can now study an upper bound for $\rho_N(\|f_z - f_{\mathscr{H}}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} > \epsilon)$ which, for an arbitrary $\epsilon > 0$, is

$$\rho_N \left(\left\| \frac{1}{N} \sum_{t=1}^n \xi_t - \mathbb{E}_{\mathscr{X}}[\xi] \right\|_\infty > \frac{\epsilon \gamma \sqrt{\check{\lambda}}}{C_{\mathscr{K}}} \right). \tag{17}$$

At an arbitrary input location $x \in \mathscr{X}$ and a given $\bar{\epsilon} \in (0, 1)$, Chebychev's inequality yields

$$\rho_N \left(\left| \frac{1}{N} \sum_{t=1}^N \xi_t(x) - \mathbb{E}_{\mathscr{X}}[\xi](x) \right| > \bar{\epsilon} \right) \leq \frac{\text{var}(\xi)(x)}{N \bar{\epsilon}^2},$$

noting that $\{\xi_t\}_{t=1}^N$ are independent and identically distributed. Using this result, we can further bound (17) as

$$\rho_N(\|f_z - f_{\mathscr{H}}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} > \epsilon) \leq \frac{C^2_{\mathscr{K}}}{\gamma^2 \check{\lambda}} \frac{\|\text{var}(\xi)\|_\infty}{N \epsilon^2}. \tag{18}$$

The variance term can be bounded as

$$\begin{aligned} \sup_{\bar{x} \in \mathscr{X}} \text{var}(\xi)(\bar{x}) &\leq \sup_{\bar{x} \in \mathscr{X}} \int_{\mathscr{X}} \mathscr{K}(\bar{x}, x)^2 (y - f_{\mathscr{H}}(x))^2 d\rho \\ &\leq C^4_{\mathscr{K}} \int_{\mathscr{X}} (y - f_{\mathscr{H}}(x))^2 d\rho \leq B_f^2 + B_\sigma^2, \end{aligned}$$

where the last inequality follows from the fact that $\int_{\mathscr{X}} (f(x) - y)^2 d\rho - \int_{\mathscr{X}} (f_\rho(x) - y)^2 d\rho = \|f - f_\rho\|^2_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$ for any $f : \mathscr{X} \to \mathscr{Y}$ [15], and that $\|f_{\mathscr{H}} - f_\rho\|^2_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} + \gamma \|f_{\mathscr{H}}\|^2_{\mathscr{H}} = \mathcal{J}(f_{\mathscr{H}}) \leq \mathcal{J}(0) = \|f_\rho\|^2_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} \leq B_f^2$. Coming back to (18), we have that

$$\rho_N \left(\|f_z - f_{\mathscr{H}}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} > \epsilon \right) \leq \frac{C^6_{\mathscr{K}}}{\gamma^2 \check{\lambda}} \frac{(B_f^2 + B_\sigma^2)}{N \epsilon^2} = \delta. \tag{19}$$

The proof is concluded by retrieving the expression for ϵ from δ in the equality (19).

Proof for operator norm bound: By definition, we look for a constant $\mathfrak{C}_\infty$ is such that, for any $u \in \mathscr{H}$, $\|(\mathcal{S}^\top_{\mathscr{X}} \mathcal{S}_{\mathscr{X}}/N + \gamma I)^{-1} u\|_\infty \leq \mathfrak{C}_\infty \|u\|_\infty$. By the reproducing property, $\left\| \left(\frac{1}{N} \mathcal{S}^\top_{\mathscr{X}} \mathcal{S}_{\mathscr{X}} + \gamma I\right)^{-1} u \right\|_\infty = \sup_{\bar{x} \in \mathscr{X}} \left| \left\langle \left(\frac{1}{N} \mathcal{S}^\top_{\mathscr{X}} \mathcal{S}_{\mathscr{X}} + \gamma I\right)^{-1} u(\cdot), \mathscr{K}(\bar{x}, \cdot) \right\rangle_{\mathscr{H}} \right|$, which is further upper bounded by $C_{\mathscr{K}} \left\| \left(\frac{1}{N} \mathcal{S}^\top_{\mathscr{X}} \mathcal{S}_{\mathscr{X}} + \gamma I\right)^{-1} \right\|_{\mathscr{H}} \|u\|_{\mathscr{H}}$ by Cauchy-Schwartz inequality and

[1] Note that the case $r = -1/2$ plays a crucial role in connecting the norms in $\mathscr{L}^2_{\rho_{\mathscr{X}}}$ and $\mathscr{H}$ for functions in the hypothesis space. Indeed, considering $f(\cdot) = \sum_{i=1}^E \alpha_i \varphi_i(\cdot)$, one has by definition of $\mathscr{H}$ that $\|f\|^2_{\mathscr{H}} = \|\Sigma_\alpha^{-1/2} \alpha\|_2^2 = \sum_{i=1}^E \alpha_i^2/\lambda_i$. On the other hand, we have that $L^{-1/2}_{\mathscr{K}}(f)(\cdot) = \sum_{i=1}^E \alpha_i/\sqrt{\lambda_i} \varphi_i(\cdot)$, and its $\mathscr{L}^2_{\rho_{\mathscr{X}}}$-norm is equal, by Parseval's Theorem, to $\sum_{i=1}^E \alpha_i^2/\lambda_i$. Therefore, $\|f\|^2_{\mathscr{H}} = \|L^{-1/2}_{\mathscr{K}} f\|^2_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$.

(6). Now, by the bound on the operator norm in $\mathscr{H}$ provided in [15, Equation 3.5], we have $\left\|\left(\frac{1}{N}\mathcal{S}_{\mathscr{X}}^{\top}\mathcal{S}_{\mathscr{X}}+\gamma I\right)^{-1}u\right\|_{\infty} \leq \frac{C_{\mathscr{K}}}{\gamma}\|L_{\mathscr{K}}^{-1/2}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}\|u\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} \leq \frac{C_{\mathscr{K}}}{\gamma}\|L_{\mathscr{K}}^{-1/2}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}\|u\|_{\infty}$. The proof is concluded by deriving the operator norm for $\|L_{\mathscr{K}}^{-1/2}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$, which is $\|L_{\mathscr{K}}^{-1/2}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} \leq 1/\sqrt{\breve{\lambda}}$ because, for an arbitrary $f \in \mathscr{L}^2_{\rho_{\mathscr{X}}}$, $\|L_{\mathscr{K}}^{-1/2}f\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} = \sqrt{\sum_{i=1}^{E}\frac{\alpha_i^2}{\lambda_i}} \leq \sqrt{\frac{1}{\breve{\lambda}}\sum_{i=1}^{E}\alpha_i^2} \leq \frac{1}{\sqrt{\breve{\lambda}}}\|f\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$. □

Remark 1. *We did not study bounds for $\mathbb{E}_{\mathscr{Z}}[\rho_N(\|f_z - f_{\mathscr{H}}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}]$, because they typically return conservative values. A result for unbounded sampling is given, e.g., in [29, Proposition 20]. Note also that our probabilistic guarantees fall in the category of "honest" bounds rather than "exact" bounds [30]: this means that, for a user-chosen confidence level δ, the result holds with confidence "at least $1-\delta$" and not with "exact probability $1-\delta$".*

IV. APPROXIMATION ERROR

We now study the error due to the choice of the hypothesis space $\mathscr{H}$, i.e., the $\mathscr{L}^2_{\rho_{\mathscr{X}}}$-distance between the solution $f_{\mathscr{H}}$ introduced in (12) and the regression function f_ρ defined in (1). We first provide an expression for $f_{\mathscr{H}}$: letting the regression function be expressed through the basis functions of $\mathscr{L}^2_{\rho_{\mathscr{X}}}$ as $f_\rho(\cdot) = \sum_{q\in\mathbb{N}}\bar{\alpha}_q\bar{\varphi}_q(\cdot)$, and recalling the definition of the RKHS basis functions $\varphi_i(\cdot)$ in (4) and of the coefficients α_i^π in (14), we have

$$f_{\mathscr{H}}(\cdot) = \sum_{i=1}^{E}\frac{\lambda_i}{\lambda_i+\gamma}\bar{\alpha}_i^\pi\varphi_i(\cdot). \tag{20}$$

Thanks to this result, we derive two bounds on the approximation error depending on different norms of the vector $\bar{\alpha}^\pi$ defined in (14). The discussion of their performance is deferred to Section VI-C. We present the result in the following Proposition.

Proposition 1. *In the trigonometric linear regression framework presented in Section II, the approximation error $\|f_{\mathscr{H}} - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$ admits the following upper bounds:*

$$\text{(a)} \qquad \frac{\gamma}{\breve{\lambda}+\gamma}\|\bar{\alpha}^\pi\|_2 + \sqrt{\sum_{q\in\mathbb{N}\setminus Q}\bar{\alpha}_q^2} \tag{21}$$

$$\text{(b)} \qquad \|\bar{\alpha}^\pi\|_\infty\gamma\sum_{i=1}^{E}\frac{1}{\lambda_i} + \sqrt{\sum_{q\in\mathbb{N}\setminus Q}\bar{\alpha}_q^2}. \tag{22}$$

Proof. Expressing the regression function as $f_\rho = \sum_{q\in\mathbb{N}}\bar{\alpha}_q\bar{\varphi}_q$ and $f_{\mathscr{H}}$ as in (20), we apply the triangle inequality and Parseval's Theorem on $\|f_{\mathscr{H}} - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$ and obtain

$$\begin{aligned}\|f_{\mathscr{H}} - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} &= \Big\|\sum_{i=1}^{E}\frac{\lambda_i}{\lambda_i+\gamma}\bar{\alpha}_i^\pi\varphi_i - \sum_{q\in\mathbb{N}}\bar{\alpha}_q\bar{\varphi}_q\Big\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} \\ &\leq \sqrt{\sum_{i=1}^{E}\left(\frac{\gamma}{\lambda_i+\gamma}\right)^2(\bar{\alpha}_i^\pi)^2} + \sqrt{\sum_{q\in\mathbb{N}\setminus Q}\bar{\alpha}_q^2}.\end{aligned}$$

Let us now focus on the first term on the right-hand side. The first bound (21) is obtained by considering $(\lambda_i+\gamma)^{-1} \leq (\breve{\lambda}+\gamma)^{-1}$. As for the second, we take $\bar{\alpha}_i^\pi \leq \|\bar{\alpha}^\pi\|_\infty$, bound the square root of the sum as the sum of the square roots, and take $(\lambda_i+\gamma)^{-1} \leq (\lambda_i)^{-1}$. □

V. BIAS-VARIANCE TRADE-OFF

In this section, we combine the bounds on the sample and approximation errors derived in Theorem 1 and Proposition 1, respectively, and study the estimated overall error $\|f_z - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$ as a function of γ and fixing $\mathscr{H}$. The main result is presented in the following Proposition.

Proposition 2. *(a) Consider the approximation error bound as in* (21)*. Then, if the number of data N and the confidence parameter δ are such that*

$$\sqrt{N\delta} > \frac{C_{\mathscr{K}}^3}{\breve{\lambda}^{3/2}}\sqrt{\frac{B_f^2+B_\sigma^2}{\sum_{i=1}^{E}(\bar{\alpha}_i^\pi)^2}}, \tag{23}$$

there exists a unique $\gamma = \hat{\gamma}_{(a)}$ minimizing the estimated error $\|f_z - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$.

(b) Take now the approximation error bound as in (22)*. Then, there always exist a unique $\gamma = \hat{\gamma}_{(b)}$ minimizing the estimated error $\|f_z - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$.*

Proof. For both cases (a) and (b), we derive sufficient conditions ensuring that the bound admits a unique minimum by studying the first and second derivatives with respect to γ.

(a) Consider the sample and approximation errors as obtained in (16) and (21), respectively. Introducing the following notation:

$$\begin{cases} A = C_{\mathscr{K}}^3\sqrt{\dfrac{B_f^2+B_\sigma^2}{N\delta\breve{\lambda}}}, \quad b = \breve{\lambda}, & \text{(24a)} \\ B = \sqrt{\displaystyle\sum_{i=1}^{E}(\bar{\alpha}_i^\pi)^2}, \quad C = \sqrt{\displaystyle\sum_{q\in\mathbb{N}\setminus Q}\bar{\alpha}_q^2}, & \text{(24b)}\end{cases}$$

we have that the overall error can be bounded as follows:

$$\|f_z - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}} \leq \frac{A}{\gamma} + \frac{B\gamma}{b+\gamma} + C = F(\gamma). \tag{25}$$

The function $F(\gamma)$ is always positive for $\gamma > 0$. Studying the first derivative, we obtain that the condition ensuring a unique root on the positive real axis is $Bb - A > 0$, which is (23). Such a condition implies also the existence of a unique flexus on $\gamma > 0$. The optimal γ has the expression $\hat{\gamma}_{(a)} = b(A+\sqrt{ABb})/(Bb-A)$.

(b) We proceed along the lines of the preceding argument but consider the approximation error bound as in (22). Considering the following coefficients:

$$A \text{ as in (24a)}, \qquad D = \sum_{i=1}^{E}\frac{\|\bar{\alpha}^\pi\|_\infty}{\lambda_i}, \tag{26}$$

the claim follows by proving that the function $F(\gamma) = \frac{A}{\gamma} + D\gamma$ has a unique minimum for $\gamma > 0$. The resulting optimal γ always exists and takes the value $\hat{\gamma}_{(b)} = \sqrt{A/D}$. □

VI. DISCUSSION

We first study the performance of the sample error bound provided in Section III by comparing it with other bounds given in [15] and [29]. Next, we discuss the result of Proposition 2, especially showcasing the capability of $\gamma^{(b)}$ to capture the behavior of the oracle γ minimizing the overall error.

A. Comparison with sample error bound in [15, Theorem 5]

In the numerical set-up, we assume that a uniformly distributed random noise with a Signal-to-Noise Ratio (SNR) of 150 affects the measurements of the regression function $f_\rho(x) = \sum_{q\in\mathbb{N}} \bar{\varphi}_q(x)\bar{\alpha}_q$ with $x \in [-1250, 1250]$. Such a function is assumed to be characterized by 20 sine/cosine couples $\{\bar{\varphi}_q\}$, where q is randomly drawn without repetitions from the set $\{1, ..., 30\}$. The hypothesis space $\mathscr{H}$ is characterized by a subset of $E/2 = 10$ sine/cosine couples randomly selected among those that define the regression function.

We perform a Monte Carlo study of 500 runs, where at each step we draw a new set of basis functions defining the regression function and the hypothesis space. Coefficients $\bar{\alpha}_q$ of the regression function are drawn from a Gaussian distribution $\mathcal{N}(0, \lambda)$, where λ is sampled from a uniform distribution on $(0, 5)$, and also enters the definition of the hypothesis space as in (8) as $\lambda_i = \lambda$ for all $i = 1, ..., E$. At each run, the number of data-points N is randomly sampled from the set $\{100, 101, ..., 1000\}$. We consider a confidence level of $\delta = 0.1$. Then, we evaluate the sample error bounds corresponding to the minimum value of γ satisfying the bound in [15, Theorem 5], and evaluate their relative difference with respect to the true sample error attained with such a γ. The results are displayed in Figure 1. Both bounds decay as $1/\sqrt{N}$, but (16) evidences a more favorable behavior in terms of the confidence level, at least for values of δ smaller than the solution of $1/\sqrt{\delta} = \log(4/\delta)$ in $(0, 1]$, that is ≈ 0.0539. Conservatism in the bound in [15, Theorem 5] is mostly due to the linear dependence on the output values bound, M. The explicit condition on M ensuring bound (16) to be more conservative is the following:

$$M \leq \frac{C_{\mathscr{K}}^2}{12}\sqrt{\frac{B_f^2 + B_\sigma^2}{\check{\lambda}\gamma}}\frac{1}{\sqrt{\delta}\log(4/\delta)}. \qquad (27)$$

Such a value tends to be very small: e.g., in the Monte Carlo test, the bound (27) returned a mean value of 3.50 ± 2.02, while the true value M emerging from the (quite favorable) SNR attained a mean value of 39.02 ± 14.66.

B. Comparison with sample error bound in [29, Proposition 20]

We consider the same numerical set-up as the previous section, and we translate the bound of [29, Proposition 20] into a statement of the same type as Theorem 1 by using Markov's inequality. To further adapt to the context given in Section II, we set $p = 2$ and $\mathcal{N}(\gamma) = \sum_{i=1}^{E} \lambda_i/(\lambda_i + \gamma)$. The bound of [29, Proposition 20] shows a slower behavior in the number of data N with respect to (16); moreover, it depends on the approximation error, which is generally not known. We performed the Monte Carlo study by setting $\|f_{\mathscr{H}} - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$ to its true value, and the results are very conservative, as displayed in Figure 1.

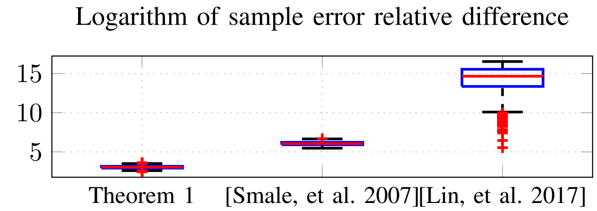

Fig. 1: Behaviour of the sample error bounds in the Monte Carlo trials in Sections VI-A and VI-B. The adopted score is the difference between the bound and true sample error, normalized by the true sample error. For Theorem 1, such an error attains a mean value of 21.44 ± 4.093, while for the bound in [15] it is 440.03 ± 99.33, and $3.40 \times 10^6 \pm 3.32 \times 10^6$ for the one in [29]. We display the values on a logarithmic scale to facilitate visualization.

C. On the choice of γ in view of the bias-variance trade-off

We now perform a Monte Carlo study to discuss the results given in Proposition 2. Consider $\mathscr{X} = [-5 \times 10^5, 5 \times 10^5]$ as input domain. The regression function is characterized by 30 sine/cosine pairs $\{\bar{\varphi}\}_{q=1}^{30}$, where each q is randomly selected without repetitions from the set $\{1, ..., 100\}$, and each component of the vector of coefficients $\bar{\alpha}$ is drawn from a Gaussian random variable with zero mean and variance $\lambda = 1$. The latter hyper-parameter also enters the definition of the RKHS $\mathscr{H}$. The set of frequencies Q is selected as a random subset with cardinality 10 from the set of those characterizing the regression function. Fixing an SNR equal to 50, we draw 50 random regression functions and select the basis functions for the hypothesis space. The number of input/output pairs for each run is $N = 2500$, and we consider a confidence parameter $\delta = 0.5$. For each run, we compute $\gamma^{(a)}$ and $\gamma^{(b)}$ as in Proposition 2, compute the sample- and approximation error bounds as in Theorem 1 and Proposition 1, and compare their values to the true errors yielded by $\gamma^{(a)}$ and $\gamma^{(b)}$. We observe that the bounds computed with $\gamma^{(a)}$ are closer to the true values. We display the values in Table I.

$\gamma_{(a)}$	True value	Bound
$\|f_z - f_{\mathscr{H}}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$	0.036 ± 0.007	0.419 ± 0.033
$\|f_{\mathscr{H}} - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$ (a)	9.313 ± 0.077	10.05 ± 0.992

$\gamma_{(b)}$	True value	Bound
$\|f_z - f_{\mathscr{H}}\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$	0.387 ± 0.076	14.04 ± 1.942
$\|f_{\mathscr{H}} - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$ (b)	7.351 ± 0.575	20.41 ± 2.193

TABLE I: Overall values (mean ± standard deviation) of sample and approximation error bounds compared to the true errors.

The test above described was performed by fixing the regularization parameter and focused on the single errors. If we instead consider the overall error $\|f_z - f_\rho\|_{\mathscr{L}^2_{\rho_{\mathscr{X}}}}$ and compare values of $\gamma^{(a)}$ and $\gamma^{(b)}$ with the oracle value γ^* (obtained via grid search) minimizing it, we observe that $\gamma^{(b)}$

is the one that performs best. The poor performance of $\gamma^{(a)}$ is due to the fact that the condition in (23) needs a large number of data to be satisfied, and this leads to an overestimation of the regularization parameter. In this specific test, γ^* was located at the minimum value of the grid, i.e. $\gamma^* = 0.1$; the mean values for $\gamma^{(a)}$ and $\gamma^{(b)}$ were 7.7703 ± 0.2115 and 0.2308 ± 0.0230, respectively.

VII. CONCLUSIONS

In this paper, we analyzed the estimation errors occurring in regularized trigonometric regression within the statistical learning set-up. To the best of the Authors' knowledge, such a study was missing in the literature, that mostly focused on non-parametric methods or non-regularized trigonometric regression. We derived a novel bound on the sample error that does not require the support of the output distribution to be finite; numerical tests showed it to be less conservative than classical bounds, at least in the non-asymptotic regime. Next, we computed two bounds for the approximation error and combined them with the sample error bound to retrieve a practical selection criterion for the regularization parameter γ, optimizing the trade-off between estimated bias and variance. In particular, we showed that one of the two criteria yields a value of the regularization parameter that is close to the oracle, and thus can in principle be used to speed up hyper-parameter selection. We stress that such an analysis can be extended to any other orthogonal basis of $\mathscr{L}^2_{\rho_{\mathscr{X}}}$. Moreover, we foresee that the generality of such a set-up can have an impact on an abstract treatment of bias learning, which is a planned extension of the present work. Forthcoming research will also focus on complementing the presented error analysis with hyper-parameter estimation for the choice of $\mathscr{H}$ (i.e., of Q and $\{\lambda_i\}_{i=1}^{E}$), possibly leveraging the Bayesian interpretation of the problem as done, e.g., in [8]; moreover, we plan to consider different risk functions (e.g., with the conditional value-at-risk [31]), and study the asymptotic behavior in terms of number of data N and of the basis functions E.

Note: An extended version of the paper can be found on `https://doi.org/10.48550/arXiv.2303.09206`.

REFERENCES

[1] G. Pillonetto, F. Dinuzzo, T. Chen, G. De Nicolao, and L. Ljung, "Kernel methods in system identification, machine learning and function estimation: A survey," *Automatica*, vol. 50, no. 3, pp. 657–682, 2014.

[2] L. Hewing, K. P. Wabersich, M. Menner, and M. N. Zeilinger, "Learning-based model predictive control: Toward safe learning in control," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, no. 1, pp. 269–296, 2020.

[3] T. Koller, F. Berkenkamp, M. Turchetta, and A. Krause, "Learning-based model predictive control for safe exploration," in *2018 IEEE Conference on Decision and Control (CDC)*, 2018, pp. 6059–6066.

[4] H. Liu, Y.-S. Ong, X. Shen, and J. Cai, "When Gaussian Process meets big data: A review of scalable GPs," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 11, pp. 4405–4423, 2020.

[5] M. Lázaro-Gredilla, J. Quiñonero-Candela, C. E. Rasmussen, and A. R. Figueiras-Vidal, "Sparse spectrum Gaussian process regression," *JMLR*, vol. 11, pp. 1865–1881, 2010.

[6] A. Rudi and L. Rosasco, "Generalization properties of learning with random features," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[7] Y. Pan, X. Yan, E. A. Theodorou, and B. Boots, "Prediction under uncertainty in sparse spectrum Gaussian processes with applications to filtering and control," in *Proceedings of the 34th International Conference on Machine Learning*, ser. PMLR, vol. 70, August 2017, pp. 2760–2768.

[8] E. Arcari, A. Scampicchio, A. Carron, and M. N. Zeilinger, "Bayesian multi-task learning using finite-dimensional models: A comparative study," in *2021 60th IEEE Conference on Decision and Control (CDC)*, 2021, pp. 2218–2225.

[9] E. Arcari, M. V. Minniti, A. Scampicchio, A. Carron, F. Farshidian, M. Hutter, and M. N. Zeilinger, "Bayesian multi-task learning MPC for robotic mobile manipulation," *IEEE Robotics and Automation Letters*, vol. 8, pp. 3222–3229, 2022.

[10] A. B. Tsybakov, *Introduction to Nonparametric Estimation*, ser. Springer Series in Statistics. New York, NY: Springer, 2009.

[11] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[12] A. L. Demmler and C. H. Reinsch, "Oscillation matrices with spline smoothing," *Numerische Mathematik*, vol. 24, pp. 375–382, 1975.

[13] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bulletin of the American Mathematical Society*, vol. 39, pp. 1–49, 2002.

[14] F. Cucker and D. X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*, ser. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2007.

[15] S. Smale and D.-X. Zhou, "Learning theory estimates via integral operators and their approximations," *Constructive Approximation*, vol. 26, pp. 153–172, 2007.

[16] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[17] Q. Wu, Y. Ying, and D.-X. Zhou, "Learning rates of least-square regularized regression," *Foundations of Computational Mathematics*, pp. 171–192, 2006.

[18] F. Cucker and S. Smale, "Best choices for regularization parameters in learning theory: On the bias—variance problem," *Foundations of Computational Mathematics*, vol. 2, pp. 413–428, March 2008.

[19] S. Mendelson and J. Neeman, "Regularization in kernel learning," *The Annals of Statistics*, vol. 38, no. 1, pp. 526 – 565, 2010.

[20] C. Wang and D.-X. Zhou, "Optimal learning rates for least squares regularized regression with unbounded sampling," *Journal of Complexity*, vol. 27, no. 1, pp. 55–67, 2011.

[21] A. Caponnetto and E. de Vito, "Optimal rates for the regularized least-squares algorithm," *Foundations of Computational Mathematics*, vol. 7, pp. 331–368, 2007.

[22] C. Wang and D.-X. Zhou, "Optimal learning rates for least squares regularized regression with unbounded sampling," *Journal of Complexity*, vol. 27, no. 1, pp. 55–67, 2011.

[23] Z.-C. Guo and D.-X. Zhou, "Concentration estimates for learning with unbounded sampling," *Advances in Computational Mathematics*, vol. 38, pp. 207–223, 2013.

[24] D.-X. Zhou, "Capacity of reproducing kernel spaces in learning theory," *IEEE Transactions on Information Theory*, vol. 49, pp. 1743 – 1752, August 2003.

[25] S. Mendelson, "Learning without concentration," in *Proceedings of The 27th Conference on Learning Theory*, ser. PMLR, vol. 35, Barcelona, Spain, 2014, pp. 25–39.

[26] N. Akhiezer and I. Glazman, *Theory of Linear Operators in Hilbert Space*, ser. Dover Books on Mathematics. Dover Publications, 2013.

[27] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, no. 3, pp. 337–404, 1950.

[28] I. Steinwart and A. Christmann, *Support Vector Machines*, 1st ed. Springer Publishing Company, Incorporated, 2008.

[29] S.-B. Lin, X. Guo, and D.-X. Zhou, "Distributed learning with regularized least squares," *Journal of Machine Learning Research*, vol. 18, no. 92, pp. 1–31, 2017.

[30] P. L. Davies, A. Kovac, and M. Meise, "Nonparametric regression, confidence regions and regularization," *Annals of Statistics*, vol. 37, pp. 2597–2625, 2007.

[31] A. Koppel, K. Zhang, H. Zhu, and T. Başar, "Projected stochastic primal-dual method for constrained online learning with kernels," *IEEE Transactions on Signal Processing*, vol. 67, no. 10, pp. 2528–2542, 2019.