Nonparametric, Nonasymptotic Confidence Bands with Paley-Wiener Kernels for Band-Limited Functions

Balázs Csanád Csáji, Member, IEEE, Bálint Horváth

Abstract—The paper introduces a method to construct confidence bands for bounded, band-limited functions based on a finite sample of input-output pairs. The approach is distribution-free w.r.t. the observation noises and only the knowledge of the input distribution is assumed. It is nonparametric, that is, it does not require a parametric model of the regression function and the regions have non-asymptotic guarantees. The algorithm is based on the theory of Paley-Wiener reproducing kernel Hilbert spaces. The paper first studies the fully observable variant, when there are no noises on the observations and only the inputs are random; then it generalizes the ideas to the noisy case using gradient-perturbation methods. Finally, numerical experiments demonstrating both cases are presented.

Index Terms—statistical learning, stochastic systems, estimation, nonlinear system identification

I. INTRODUCTION

Regression is one of the fundamental problems of statistics, system identification, signal processing and machine learning [1]. Given a finite sample of input-output pairs, the typical aim is to estimate the so-called regression function, which, given an input, encodes the conditional expectation of the corresponding output [2]. There are several well-known (parametric and nonparametric) approaches for regression, from linear regression to neural networks and kernel methods, which provide point-estimates from a given model class [3].

However, sole point-estimates are often not sufficient and region-estimates are also needed, for example, to support robust approaches. These region-estimates have several variants, such as confidence regions for the “true” function generating the observations [4]; for the expected output at a given input [5]; and prediction regions for the next (noisy) observation [6].

In this paper, we focus on building confidence bands for the regression function. These bands have natural connections to filtering and smoothing methods. While in a parametric setting such region-estimates are typically induced by confidence sets in the parameter space, in a nonparametric setting this indirect approach is not feasible. Therefore, nonparametric confidence bands for the expected outputs should be constructed directly.

This research was supported by the National Research, Development and Innovation Office (NRDIO) of Hungary, within the framework of the Artificial Intelligence National Laboratory Program; and by the Thematic Excellence Programme (TKP) of NRDIO, pr. no. TKP2021-NKTA-01. B. Cs. Csáji is with SzTAKI: Institute for Computer Science and Control, Budapest, Hungary; and also with Institute of Mathematics, Eötvös Loránd University (ELTE), Budapest, Hungary; csaji@sztaki.hu.
B. Horváth is with SzTAKI: Institute for Computer Science and Control, Eötvös Loránd Research Network (ELKH), Budapest, Hungary; and also with Institute of Mathematics, Budapest University of Technology and Economics (BME), Hungary; balint.horvath@sztaki.hu.

Regarding prediction intervals for the next observation, promising distribution-free approaches are interval predictor models (IPMs) based on the scenario approach [7], [8], and the conformal prediction framework also offers several nonparametric methods for regression and classification [6].

If the data is jointly Gaussian, a powerful methodology is offered by Gaussian process regression [5] that can provide prediction regions for the outputs, and credible regions for the expected outputs. However, the Gaussianity assumption is sometimes unrealistic that calls for alternative approaches.

In this paper, we suggest a nonparametric approach using Paley-Wiener kernels, to build data-driven simultaneous confidence bands for an unknown bounded, band-limited function, based on an independent and identically distributed (i.i.d.) sample of input-output pairs. The method is distribution-free in the sense that only very mild assumptions are needed about the observation noises, such as they are distributed symmetrically about zero. On the other hand, we assume that the distribution of the inputs is known, particularly, we assume uniformly distributed inputs, as more general cases can often be traced back to this assumption. First, the case without observation noises is studied, then the ideas are extended to the general, noisy case. The results are supported by both non-asymptotic theoretical guarantees and numerical experiments.

II. KERNELS AND BAND-LIMITED FUNCTIONS

Kernel methods have an immense range of applications in machine learning and related fields [9]. In this section, we review some of their fundamental theoretical concepts.

A. Reproducing Kernel Hilbert Spaces

A Hilbert space $\mathcal{H}$ of $f : \mathbb{X} \to \mathbb{R}$ functions with an inner product $\langle \cdot, \cdot \rangle_\mathcal{H}$ is called a Reproducing Kernel Hilbert Space (RKHS), if each Dirac functional, which evaluates functions at a point, $\delta_z : f \mapsto f(z)$, is bounded for all $z \in \mathbb{X}$, that is $\forall z \in \mathbb{X} : \exists k_z > 0$ with $|\delta_z(f)| \leq k_z \|f\|_\mathcal{H}$ for all $f \in \mathcal{H}$.

Then, by building on the Riesz representation theorem, a unique kernel, $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$, can be constructed encoding the Dirac functionals satisfying $(k(\cdot, z), f)_\mathcal{H} = f(z)$, for all $z \in \mathbb{X}$ and $f \in \mathcal{H}$, which formula is called the reproducing property. As a special case of this property, we also have for all $z, s \in \mathbb{X}$ that $k(z, s) = \langle k(\cdot, z), k(\cdot, s) \rangle_\mathcal{H}$. Therefore, the kernel of an RKHS is a symmetric and positive-definite function.

Furthermore, the Moore-Aronszajn theorem asserts that the converse statement holds true, as well: for every symmetric and positive-definite function $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$, there exists a unique RKHS for which $k$ is its reproducing kernel [10].
The Gram or kernel matrix of a given kernel \( k \) w.r.t. (input) points \( x_1, \ldots, x_n \) is \( K_{i,j} = k(x_i, x_j) \), for all \( i, j \in [n] \). Observe that \( K \in \mathbb{R}^{n \times n} \) is always positive semi-definite. A kernel is called strictly positive-definite, if its Gram matrix is positive-definite for all distinct inputs \( \{x_i\} \).

Archetypal kernels include the Gaussian kernel \( k(z,s) = \exp\left(-||z-s||^2/(2\sigma^2)\right) \), where \( \sigma > 0 \); the polynomial kernel \( k(z,s) = (z^T s + c)^p \), where \( c \geq 0, p \in \mathbb{N} \); and the sigmoidal kernel \( k(z,s) = \tanh(\langle z, s \rangle + b) \), for some \( a, b \geq 0 \).

**B. Paley-Wiener Spaces**

Let \( \mathcal{H} \) be the space of \( f \in \mathcal{L}^2(\mathbb{R}, \lambda) \) functions, where \( \lambda \) is the Lebesgue measure, such that the support of the Fourier transform of \( f \) is included in \([-\eta, \eta] \), where \( \eta > 0 \). It is a subspace of \( \mathcal{L}^2 \) and thus we use the \( \mathcal{L}^2 \) inner product:

\[
(f,g)_\mathcal{H} = \int_{\mathbb{R}} f(x) g(x) d\lambda(x).
\]

This space of band-limited functions, called the Paley-Wiener space \([10]\), is an RKHS. Its reproducing kernel is

\[
k(z,s) = \frac{\sin(\eta(z - s))}{\pi(z - s)},
\]

for \( z \neq s \), where \( z, s \in \mathbb{R} \); and \( k(z,z) = \eta/\pi \). Henceforth, we will work with the above defined Paley-Wiener kernel.

**Remark 1:** Paley-Wiener spaces can also be defined on \( \mathbb{R}^d \) [11], but for simplicity we focus on the scalar input case.

**III. Nonparametric Confidence Bands**

Let \( (x_1, y_1), \ldots, (x_n, y_n) \) be a finite sample of i.i.d. pairs of random variables with unknown joint distribution \( P_{x,y} \), where \( x_k \) and \( y_k \) are \( \mathbb{R} \)-valued, and \( \mathbb{E}[y_k^2] < \infty \). We assume that

\[
y_k = f_*(x_k) + \varepsilon_k,
\]

for \( k \in [n] \), where \( \mathbb{E}[\varepsilon_k] = 0 \). Variables \( \{\varepsilon_k\} \) represent the measurement or observation noises on the “true” \( f_* \).

We call \( f_* \) the regression function \([1]\), as on the support of \( \{x_k\} \) it can also be written as \( f_*(x) = \mathbb{E}[y|X = x] \), where \((X, Y)\) is a random vector with distribution \( P_{x,y} \).

**A. Objectives and Reliability**

Our aim is to build a (simultaneous) confidence band for \( f_* \), i.e., a function \( I : \mathcal{D} \to \mathbb{R} \times \mathbb{R} \), where \( \mathcal{D} \) is the support of the input distribution, such that \( I(x) = (I_1(x), I_2(x)) \) specifies the endpoints of an interval estimate for \( f_*(x) \), for all \( x \in \mathcal{D} \). More precisely, we would like to construct \( I \) with

\[
\nu(I) = \mathbb{P}(\forall x \in \mathcal{D} : I_1(x) \leq f_*(x) \leq I_2(x)) \geq 1 - \alpha,
\]

where \( \alpha \in (0,1) \) is a user-chosen risk probability, and \( \nu(I) \) is the reliability of the confidence band. Let us introduce

\[
I = \{ (x,y) \in \mathcal{D} \times \mathbb{R} : y \in [I_1(x), I_2(x)] \}.
\]

Based on this, the reliability is \( \nu(I) = \mathbb{P}(\text{graph}_D(f_*) \subseteq I) \), where we define \( \text{graph}_D(f_*) = \{ (x,f_*(x)) : x \in \mathcal{D} \} \).

For notational simplicity, we will use \( I(x) = \emptyset \) to denote \( I(x) = (1,-1) \), i.e., the endpoints of an empty interval.

Hence, we aim at building a confidence band that contains the graph (w.r.t. domain \( \mathcal{D} \)) of the “true” \( f_* \) with a user-chosen probability level. Moreover, we would like to have a distribution-free method (w.r.t. the noises) and the region should have finite-sample guarantees without a parametric model of \( f_* \), namely, we take a nonparametric approach.

**Remark 2:** We note here, as well, that in the IPMs \([7]\) \([8]\) and in the conformal prediction framework \([6]\), the aim is to build a guaranteed prediction region for the next observation, while here we aim at predicting the value of the regression function instead. In this sense, our objective is similar to that of the region estimates of Gaussian process regression \([5]\), however, without the assumption of joint Gaussianity.

**B. Main Assumptions**

Our core assumptions can be summarized as follows:

**A0:** The dataset, \((x_1,y_1), \ldots, (x_n,y_n) \) \( \in \mathbb{R} \times \mathbb{R} \), is an i.i.d. sample of input-output pairs; and \( \mathbb{E}[y_k^2] < \infty \), for \( k \in [n] \).

**A1:** Each (measurement) noise, \( \varepsilon_k = y_k - f_*(x_k) \), for \( k \in [n] \), has a symmetric probability distribution about zero.

**A2:** The inputs, \( \{x_k\} \), are distributed uniformly on \([0,1]\).

**A3:** Function \( f_* \) is from a Paley-Wiener space \( \mathcal{H} \); \( \forall x \in [0,1] : |f_*(x)| \leq 1 \); and \( f_* \) is almost time-limited to \([0,1] : \int_{\mathbb{R}} f_*(x) I(x \notin [0,1]) \, d\lambda(x) \leq \delta_0 \), where \( I(\cdot) \) is an indicator and \( \delta_0 > 0 \) is a universal constant.

Now, let us briefly discuss these assumptions. The i.i.d. requirement of A0 is standard in mathematical statistics and supervised learning \([12]\). The square-integrability of the outputs is needed to estimate the \( \mathcal{L}^2 \) norm of \( f_* \), based on the sample and to have a well-defined regression function. The assumption on the noises, A1, is very mild, as most standard distributions (e.g., Gauss, Laplace and uniform) satisfy this.

Our strongest assumption is certainly A2, which basically amounts to the assumption that we know the distribution of the inputs and it is absolutely continuous. The more general case when the inputs, \( \{x_k\} \), have a known, strictly monotone increasing and continuous cumulative distribution function \( F \), could be traced back to assumption A2, since it is well-known that \( x_k \sim F(x_k^\prime) \) is distributed uniformly on \([0,1]\).

Assumption A3, especially limiting the frequency domain of \( f_* \), is needed to restrict the model class and to ensure that we can effectively generalize to unknown data points. We allow the “true” function to be defined outside the support of the inputs, cf. the Fourier uncertainty principle \([13]\), but the part of \( f_* \) outside of \( \mathcal{D} = [0,1] \) should be “negligible”: i.e., its norm cannot exceed a (known) small constant, \( \delta_0 \).

A crucial property of Paley-Wiener spaces is that their norms coincide with the standard \( \mathcal{L}^2 \) norm, which will allow us to efficiently upper bound \( \|f_*\|_{\mathcal{H}}^2 \) based on the sample.

**IV. Confidence Bands: Noise-Free Case**

In order to motivate our solution, we start with a simplified problem, in which we observe the regression function perfectly at random inputs. In this noise-free case, we can recall the
celebrated Nyquist–Shannon sampling theorem, which states that
a band-limited function can be fully reconstructed from the
samples, assuming the sampling rate exceeds twice the
maximum frequency. On the other hand, if we only have a
small number of observations, we cannot apply this result.
Nevertheless, we still would like to have at least a region
estimate. In this section we provide such an algorithm.

Recall that for a dataset \( \{(x_k, y_k)\} \), where inputs \( \{x_k\} \) are
distinct (which has probability one under A2), the element
from \( \mathcal{H} \) that has the minimum norm and interpolates each
output \( y_k \) at the corresponding input \( x_k \), that is
\[
\tilde{f} = \arg \min \{ \| f \|_{\mathcal{H}} : f \in \mathcal{H} & \forall k \in [n] : f(x_k) = y_k \},
\]
takes the following form [10] for all input \( x \in \mathcal{X} \):
\[
\tilde{f}(x) = \sum_{k=1}^{n} \tilde{\alpha}_k k(x, x_k),
\]
where the weights are \( \tilde{\alpha} = K^{-1}y \) with \( y \equiv (y_1, \ldots, y_n)^T \)
and \( \tilde{\alpha} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_n)^T \); we also used that the Paley-Wiener
kernel is strictly positive-definite, thus matrix \( K \) is invertible.

We will exploit, as well, that the norm square of \( \tilde{f} \) is
\[
\| \tilde{f} \|_{\mathcal{H}}^2 = \tilde{\alpha}^T K \tilde{\alpha},
\]
which is a direct consequence of the reproducing property.

Assuming we have a stochastic upper bound for the norm
square of the regression function, denoted by \( \kappa \), the idea
of our construction is as follows. We include those \( (x_0, y_0) \)
pairs in the confidence band, for which the minimum norm
interpolation of \( \{(x_k, y_k)\} \cup \{(x_0, y_0)\} \), namely, which simulta-
nuously interpolates the original dataset and \( (x_0, y_0) \), has a
norm square which is less than or equal to \( \kappa \). In order to make

this approach practical, we need (1) a guaranteed upper bound
for the norm square of the “true” data-generating function;
and (2) an efficient method to decide the endpoints of the
confidence interval for each potential input \( x_0 \in \mathcal{D} \).

\section{Bounding the Norm: Noise-Free Case}

It is easy to see that in the noise-free case, if \( y_k = f_*(x_k) \), for
\( k \in [n] \), the norm square of \( f_* \) can be estimated by
\[
\frac{1}{n} \sum_{k=1}^{n} y_k^2 = \frac{1}{n} \sum_{k=1}^{n} f_*^2(x_k) \approx \mathbb{E}[f^2(X)] \approx \| f_* \|^2 = \| f_* \|^2_{\mathcal{H}},
\]
since in the Paley-Wiener space the norm is the \( L^2 \) norm,
and we also used that \( \{x_k\} \) are uniform on domain \( \mathcal{D} = [0, 1] \).

As the next lemma demonstrates, we can construct such a
guaranteed upper bound using the Hoeffding inequality.

\textbf{Lemma 1: Assuming A0, A2, A3 and that} \( y_k = f_*(x_k) \), for
\( k \in [n] \), \textbf{we have for any risk probability} \( \alpha \in (0, 1) \),
\[
\mathbb{P}(\| f_* \|^2_{\mathcal{H}} \leq \kappa) \geq 1 - \alpha,
\]
\textbf{with the following choice of the upper bound} \( \kappa \):
\[
\kappa = \frac{1}{n} \sum_{k=1}^{n} y_k^2 + \sqrt{\frac{\ln(\alpha)}{-2n}} + \delta_0.
\]

\textbf{Proof: By using the notation} \( R = \frac{1}{n} \sum_{k=1}^{n} y_k^2 \), we have
\[
\mathbb{E}[R] = \| f_* \cdot \mathbb{I}_\mathcal{D} \|^2_{\mathcal{H}} \geq \| f_* \|^2_{\mathcal{H}} - \delta_0,
\]
where \( \mathbb{I}_\mathcal{D} \) is the indicator function of \( \mathcal{D} = [0, 1] \). That is, \( R \)
is a Monte Carlo estimate of the integral of this \( L^2 \) norm.

Then, from the Hoeffding inequality, for all \( t > 0 \):
\[
\mathbb{P}(R - \mathbb{E}[R] \leq -t) \leq \exp(-2nt^2).
\]
According to the complement rule, we also have
\[
\mathbb{P}(\mathbb{E}[R] < R + t) \geq 1 - \exp(-2nt^2).
\]
We would like choose a threshold \( t > 0 \) such that
\[
1 - \alpha \leq \mathbb{P}(\mathbb{E}[R] < R + t).
\]
This inequality is satisfied if we choose a \( t > 0 \) with
\[
1 - \alpha \leq 1 - \exp(-2nt^2) \implies \exp(-2nt^2) \leq \alpha,
\]
After taking the natural logarithm, we get \( -2nt^2 \leq \ln(\alpha) \), hence,
the choice of \( t^* = \sqrt{\ln(\alpha)/(-2n)} \) guarantees
\[
\mathbb{P}(\| f_* \|^2_{\mathcal{H}} \geq R + t^* + \delta_0) \leq \alpha,
\]
which completes the proof of the lemma.

\section{Interval Endpoints: Noise-Free Case}

Now, we construct a confidence interval for a given input
query point \( x_0 \in \mathcal{D} \), for which \( x_0 \neq x_k \), for \( k \in [n] \). That
is, we build an interval \( [I_1(x_0), I_2(x_0)] \) that contains \( f_*(x_0) \)
with probability at least \( 1 - \alpha \), where \( \alpha \in (0, 1) \) is given.

First, we extend the Gram matrix with query point \( x_0 \),
\[
K_0(i + 1, j + 1) = k(x_i, x_j),
\]
for \( i, j = 0, 1, \ldots, n \). As \( \{x_k\}_{k=0}^{n} \) are distinct (a.s.), this
Gramian can be inverted. Hence, for any \( y_0 \), the minimum
norm interpolation of \( (x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n) \) is
\[
\tilde{f}(x) = \sum_{k=0}^{n} \tilde{\alpha}_k k(x, x_k),
\]
where the weights are \( \tilde{\alpha} = K_0^{-1} \tilde{y} \) with \( \tilde{y} \equiv (y_0, y_1, \ldots, y_n)^T \)
and \( \tilde{\alpha} = (\tilde{\alpha}_0, \ldots, \tilde{\alpha}_n)^T \). The norm square of \( \tilde{f} \) is
\[
\| \tilde{f} \|^2_{\mathcal{H}} = \tilde{\alpha}^T K_0 \tilde{\alpha} = \tilde{y}^T K_0^{-1} \tilde{y} = \tilde{y}^T K_0^{-1} \tilde{y}.
\]
Since the output query point \( y_0 \) in \( \tilde{y} = (y_0, y^T)^T \) is arbitrary,
we can compute the minimum norm needed to interpolate the
original dataset extended by \( (x_0, y_0) \) for any candidate \( y_0 \).

Therefore, having a bound \( \kappa \) on the norm square (which is
guaranteed with probability \( \geq 1 - \alpha \) ), we can compute the
highest and the lowest \( y_0 \) values which can be interpolated
with a function from \( \mathcal{H} \) having at most norm square \( \kappa \).

This leads to the following two optimization problems:
\[
\min_{y_0} \max_{\mathcal{Y}} \mathbb{I}_{\mathcal{Y}}(y_0, y^T)^T K_0^{-1} (y_0, y^T)^T \leq \kappa
\]
where “min/max” means that we have to solve the problem
as a minimization and also as a maximization (separately).

The optimal values of these problems, denoted by \( y_{\text{min}} \) and
\( y_{\text{max}} \), respectively, determine the endpoints of the confidence interval for \( f_*(x_0) \), that is \( I_1(x_0) = y_{\text{min}} \) and \( I_2(x_0) = y_{\text{max}} \).

Problems (1) are convex, moreover, as we will show, their
optimal values can be calculated analytically. First, note that
the only decision variable of these problems is $y_0$, everything else is constant (including the input $x_0$, which is also given).

Let us partition the inverse Gramian, $K_0^{-1}$, as

$$
\begin{bmatrix}
    c & b^T \\
    b & A
\end{bmatrix} = K_0^{-1},
$$

where $c \in \mathbb{R}$, $b \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$; after which

$$(y_0, y^T)K_0^{-1}(y_0, y^T) = c y_0^2 + 2b^T y + y^T A y.$$

Then, introducing $a_0 = c$, $b_0 = 2b^T y$ and $c_0 = y^T A y - \kappa$, the two optimization problems (1) can be written as

$$
\begin{align*}
\min & \quad y_0 \\
\text{subject to} & \quad a_0 y_0^2 + b_0 y_0 + c_0 \leq 0
\end{align*}
$$

in which $a_0$, $b_0$ and $c_0$ are constants (w.r.t. the optimization).

Since these are (convex) quadratic programming problems (with linear objectives), their optimal solutions must be on the boundary of the constraint. This can be easily verified directly, for example, by the technique of Lagrange multipliers. There are at most two solutions of the quadratic equation $a_0 y_0^2 + b_0 y_0 + c_0 = 0$. The smaller one will be denoted by $y_{\text{min}}$ and the larger one by $y_{\text{max}}$ (they are allowed to be the same, if there is only one solution). Then, we set $I_1(x_0) \equiv y_{\text{min}}$, and $I_2(x_0) \equiv y_{\text{max}}$; or $I(x_0) \equiv \emptyset$, in case there is no solution. Finally, we define $I_1(x_k) = I_2(x_k) = y_k$, for all $k \in [n]$, as the outputs are noise-free, that is $y_k = f^*(x_k)$, for $k \in [n]$.

Table I summarizes the proposed algorithm for the case without measurement noise. By observing that if $\kappa$ satisfies $\|f\|_H^2 \leq \kappa$, which has probability at least $1 - \alpha$, then the construction guarantees that $\text{graph}_H(f^*_\kappa) \subseteq \mathcal{I}$, as the region contains all outputs that can be interpolated with a function from $\mathcal{H}$ which also interpolates the original dataset and has norm square at most $\kappa$. Hence, we can conclude that

**Theorem 1:** Assume that $A_0$, $A_2$, $A_3$ and $y_k = f^*_\kappa(x_k)$, for $k \in [n]$, are satisfied. Let $\alpha \in (0, 1)$ be a risk probability. Then, the confidence band of Algorithm I guarantees

$$
\mathbb{P}(\text{graph}_H(f^*_\kappa) \subseteq \mathcal{I}) \geq 1 - \alpha.
$$

### V. Confidence Bands with Measurement Noise

Now, we turn to the general case, when the observations of $f_\kappa$ are affected by noises, $y_k = f_\kappa(x_k) + \varepsilon_k$, for $k \in [n]$.

Since now we do not have exact knowledge of the function values at the sample inputs, we cannot directly apply our previous approach. The main idea in this case is that first we need to construct interval estimates of $f_\kappa$ at some observed inputs, $\{x_k\}_k$, which can then be used to bound the norm and to build confidence intervals for the unobserved inputs.

#### A. Confidence Intervals at the Observed Inputs

We employ the kernel gradient perturbation (KGP) method, proposed in [14], to build non-asymptotically guaranteed, distribution-free confidence intervals for $f_\kappa$ at some of the observed inputs. The KGP algorithm is based on ideas from finite-sample system identification [4], particularly, it is an extension of the Sign-Perturbed Sums (SPS) method [15].

The KGP method can build non-asymptotically guaranteed distribution-free confidence regions for the RKHS coefficients of the ideal representation (w.r.t. given input points) of $f_\kappa$. A representation $f \in \mathcal{H}$ is called ideal w.r.t. $\{x_k\}_k$, if it has the property that $f(x_k) = f_\kappa(x_k)$, for all $k \in [d]$.

The KGP construction guarantees [14, Theorem 2] that the confidence set contains the coefficients of an ideal representation w.r.t. $\{x_k\}_k$ exactly with a user-chosen confidence probability, assuming the noises satisfy regularity conditions, e.g., they are symmetric and independent (cf. $A_0$ and $A_1$).

Note that KGP regions are only guaranteed at the observed inputs. KGP cannot provide confidence bands directly.

The KGP approach can be used together with a number of kernel methods, such as support vector regression and kernelized LASSO. Here, we use it with kernel ridge regression (KRR) which is the kernelized version of Tikhonov regularized least squares (LS). It solves the following problem:

$$
\hat{x}_{\text{KRR}} = \arg\min_{x \in \mathcal{H}} \frac{1}{n} \sum_{k=1}^{n} w_k (y_k - x(x_k))^2 + \lambda \|x\|_H^2,
$$

where $\lambda > 0$, $w_k > 0$, $i \in [n]$, are given (constant) weights.

Using the representer theorem [16] and the reproducing property, the objective of (3) can be rewritten as [14]

$$
\frac{1}{n} (y - K \theta)^T W (y - K \theta) + \lambda \theta^T K \theta,
$$

where $W \equiv \text{diag}(w_1, \ldots, w_n)$, $K$ is the Gramian matrix, and $\theta = (\theta_1, \ldots, \theta_n)$ are the coefficients of the solution.

Minimizing (4) can be further reformulated as a canonical ordinary least squares (OLS) problem, $\|v - \Phi \theta\|^2$, by using

$$
\Phi = \begin{bmatrix}
    (1/\sqrt{\pi}) W^{1/2} K \\
    \sqrt{\lambda} K \sqrt{\pi}
\end{bmatrix},
\quad v = \begin{bmatrix}
    (1/\sqrt{\pi}) W^{1/2} y \\
    0_n
\end{bmatrix},
$$

where $W^{1/2}$ and $K^{1/2}$ denote the principal, non-negative square roots of matrices $W$ and $K$, respectively. Note that the square roots exist as these matrices are positive semi-definite.

For convex quadratic problems (such as KRR) and symmetric noises (cf. $A_1$), the KGP confidence regions coincide with SPS regions. They are star convex with the LS estimate.
\( \hat{\theta} \), as a star center. Furthermore, they have *ellipsoidal outer approximations*, that is there are regions of the form
\[
\hat{\Theta}_\beta \doteq \left\{ \theta \in \mathbb{R}^n : (\theta - \hat{\theta})^T \frac{1}{n} \Phi^T \Phi (\theta - \hat{\theta}) \leq r \right\},
\]
where \( 1 - \beta \in (0, 1) \) is a given confidence probability \([15]\). The radius of this confidence ellipsoid, \( r \), can be computed by *semi-definite programming*: see \([15, Section VI.B]\).

Hence, the construction guarantees \( P(\theta \in \hat{\Theta}_\beta) \geq 1 - \beta \), where \( \theta \) is the coefficient vector of an *ideal* representation:
\[
\sum_{i=1}^n \theta_i k(x_i, x_k) = f_\ast(x_k),
\]
for \( k \in [n] \). By defining \( \varphi_k \doteq (k(x_1, x_k), \ldots, k(x_n, x_k))^T \), we know that \( f_\ast(x_k) = \varphi_k^T \hat{\theta} \), but of course \( \theta \) is unknown.

Since \( \hat{\theta} \) is inside the ellipsoid \( \hat{\Theta}_\beta \) with probability \( 1 - \beta \), we could construct (probabilistic) upper and lower bounds of \( f_\ast(x_k) \) by maximizing and minimizing \( \varphi_k^T \theta \), for \( \theta \in \hat{\Theta}_\beta \).

These problems (linear objective and ellipsoid constraint) have known solutions: the minimum and the maximum are
\[
\hat{\nu}_k = \varphi_k^T \hat{\theta} - (\varphi_k^T P \varphi_k)^{\frac{1}{2}}, \quad \hat{\mu}_k = \varphi_k^T \hat{\theta} + (\varphi_k^T P \varphi_k)^{\frac{1}{2}},
\]
where \( P = (nr)^{-1} \Phi^T \Phi \), and \( \hat{\theta} \) is the center of the ellipsoid, i.e., the solution of the OLS formulation \( \| v - \Phi \theta \|_2^2 \).

Due to the construction of KGP confidence regions, there is a (extremely small, but nonzero) probability of getting an empty region. In this case, we define \( \nu_k = 1 \) and \( \mu_k = -1 \), for all \( k \in [n] \). That is, we give an *empty interval* for each \( f(x_k) \), using a similar representation as in Section III-A.

Finally, we introduced a slight modification to this construction. We can also construct confidence intervals just for the first \( d \leq n \) observations by redefining objective (4) as
\[
\frac{1}{n} \sum_{i=1}^n (y_i - K_1 \theta)^T W (y_i - K_1 \theta) + \lambda \theta^T K_2 \theta,
\]
where \( K_1 \in \mathbb{R}^{n \times d} \) is \( K \) having the last \( n - d \) columns removed, and \( K_2 \in \mathbb{R}^{d \times d} \) is \( K_1 \) having the last \( n - d \) rows removed. Hence, we search for \( \hat{\theta} \in \mathbb{R}^d \) ideal vector, such that for \( k \in [d] \), we have \( (K_1 \hat{\theta})(k) = f_\ast(x_k) \). For the error computation we still use all measurements (\( K_1 \) still has \( n \) rows). It is important that in this case only the first \( d \) residuals are perturbed in the construction of the KGP ellipsoid. This usually considerably reduces the sizes of the intervals, but then we only have guarantees at \( d \leq n \) observed inputs.

**B. Bounding the Norm with Measurement Noise**

In the previous section, we built simultaneous confidence intervals at the sample inputs for the first \( d \leq n \) observations, \([\nu_k, \mu_k] \), for \( k \in [d] \); that is, they have the property
\[
P( \forall k \in [d] : f_\ast(x_k) \in [\nu_k, \mu_k] ) \geq 1 - \beta,
\]
for some (user-chosen) risk probability \( \beta \in (0, 1) \).

Recall that by Lemma 1, for any \( n \), the variable
\[
\kappa \doteq \frac{1}{n} \sum_{k=1}^n f_\ast^2(x_k) + \sqrt{\frac{\ln(n)}{2n}} + \delta_0,
\]
is an upper bound of \( \| f_\ast \|_H^2 \) with probability at least \( 1 - \alpha \).
where “min/max” again means that the problem have to be solved as a minimization and as a maximization (separately).

These problems are convex, therefore, they can be solved efficiently. The optimal values, denoted by $z_{\min}$ and $z_{\max}$, are the endpoints of the confidence interval: $I_1(x_0) \doteq z_{\min}$, and $I_2(x_0) \doteq z_{\max}$. If (9) is infeasible, e.g., we get an empty KGP ellipsoid, we set $I(x_0) = \emptyset$, i.e., we use $I(x_0) = (1, -1)$.

Table II summarizes the algorithm to construct the endpoints of a confidence interval at a given query point, in case of having measurement noises. Its theoretical guarantee is:

**Theorem 2:** Assume that A0, A1, A2, A3 are satisfied. Let $\alpha, \beta \in (0, 1)$ be given risk probabilities. Then, the confidence band built by Algorithm II described above guarantees

$$ \mathbb{P}(\text{graph}_D(f_x) \subseteq I) \geq 1 - \alpha - \beta. $$

**Remark 4:** Applying the KGP approach in Algorithm II is optional. One could use any other construction that provides simultaneous confidence intervals for a subset of $\{f_x(x_k)\}$, cf. (6). Another approach could be to assume sub-Gaussian or sub-exponential noises and use their tail bounds to ensure (6).

**VII. NUMERICAL EXPERIMENTS**

The algorithms were also tested numerically. We used a Paley-Wiener RKHS with $\eta = 30$. The “true” function was constructed as follows: first, 20 random input points $\{x_k\}_{k=1}^{20}$ were generated, with uniform distribution on $[0, 1]$. Then $f_\ast(x) = \sum_{k=1}^{20} w_k k(x, x_k)$ was created, where each $w_k$ had a uniform distribution on $[-1, 1]$. The function was normalized, in case its maximum exceeded 1. Then, $n$ random observations were generated about $f_\ast$. In the noisy case, $\{x_k\}$ had Laplace distribution with location $\mu = 0$ and scale $b = 0.4$ parameters.

In the noise-free case, we used $n = 10$ observations, and created confidence bands with risk $\alpha = 0.1$ and $0.5$. Figure 1 demonstrates that in the noise-free setting a very small sample size can lead to informative nonparametric confidence bands.

In case of measurement noises, $n = 100$ sample size was used with $d = 20$ (orange points). Confidence bands with risk $\alpha + \beta = 0.1$ and $0.5$ are illustrated in Figure 2. We simply used $\alpha = \beta$ in these cases. The results indicated that even with limited information, adequate regions can be created.

**REFERENCES**

[1] F. Cucker and D. X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*, vol. 24. Cambridge University Press, 2007.

[2] L. Ljung, “Perspectives on System Identification,” *Annual Reviews in Control*, vol. 34, no. 1, pp. 1–12, 2010.

[3] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

[4] A. Carè, B. Cs. Csáji, M. Campi, and E. Weyer, “Finite-Sample System Identification: An Overview and a New Correlation Method,” *IEEE Control Systems Letters*, vol. 2, no. 1, pp. 61–66, 2018.

[5] J. Quinonero-Candela and C. E. Rasmussen, “A Unifying View of Sparse Approximate Gaussian Process Regression,” *Journal of Machine Learning Research*, vol. 6, pp. 1939–1959, 2005.

[6] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. Springer Science & Business Media, 2005.

[7] M. C. Campi, G. Calafiore, and S. Garatti, “Interval Predictor Models: Identification and Reliability,” *Automatica*, vol. 45, pp. 382–392, 2009.

[8] S. Garatti, M. Campi, and A. Care, “On a Class of Interval Predictor Models with Universal Reliability,” *Automatica*, vol. 110, 2019.

[9] G. Pilloletto, F. Dinuzzo, T. Chen, G. De Nicolao, and L. Ljung, “Kernel Methods in System Identification, Machine Learning and Function Estimation: A Survey,” *Automatica*, pp. 657–682, 2014.

[10] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer, 2004.

[11] A. Iosevich and A. Mayeli, “Exponential Bases, Paley-Wiener Spaces and Applications,” *Journal of Functional Analysis*, pp. 363–375, 2015.

[12] V. Vapnik, *Statistical Learning Theory*. Wiley-Interscience, 1998.

[13] M. A. Pinsky, *Introduction to Fourier Analysis and Wavelets*, vol. 102. American Mathematical Society, 2008.

[14] B. Cs. Csáji and K. B. Kis, “Distribution-Free Uncertainty Quantification for Kernel Methods by Gradient Perturbations,” *Machine Learning*, vol. 108, no. 8, pp. 1677–1699, 2019.

[15] B. Cs. Csáji, M. C. Campi, and E. Weyer, “Sign-Perturbed Sums: A New System Identification Approach for Constructing Exact Non–Asymptotic Confidence Regions in Linear Regression models,” *IEEE Transactions on Signal Processing*, vol. 63, no. 1, pp. 109–181, 2014.

[16] T. Hofmann, B. Schölkopf, and A. J. Smola, “Kernel Methods in Machine Learning,” *Annals of Statistics*, vol. 36, pp. 1171–1220, 2008.