Nonparametric Simultaneous Confidence Bands: The Case of Known Input Distributions

Bálint Horváth*[1] and Balázs Csanád Csáji[2]

[1]*Institute for Computer Science and Control (SZTAKI), Budapest, Hungary and Institute of Mathematics, Budapest University of Technology and Economics (BME)*
[2]*Institute for Computer Science and Control (SZTAKI), Budapest, Hungary and Institute of Mathematics, Eötvös Loránd University (ELTE), Hungary*

October 2, 2023

Abstract: In this paper, we construct nonparametric, nonasymptotic, and simultaneous confidence bands for band-limited regression functions based on the theory of Paley-Wiener kernels. We work with a sample of independent and identically distributed (i.i.d.) input-output pairs, the measurement noises are assumed to have a joint distribution that is invariant with respect to transformations from a compact matrix group (e.g., permutations), and we also assume that the distribution of the inputs is a priori known. The task is divided into two steps: first, we study the case when the outputs are noise-free, then the problem is generalized for measurement noises. The algorithms provide nonasymptotic guarantees for the inclusion of the true regression function in the confidence band, simultaneously for all possible inputs. Finally, we demonstrate our results via numerical experiments.

Keywords: nonasymptotic confidence bands, convex optimization, reproducing kernel Hilbert spaces, Paley-Wiener spaces, randomized algorithms

AMS subject classification: 46E22, 62G15, 68Q32

Acknowledgements: This research was supported by the European Union within the framework of the Artificial Intelligence National Laboratory, RRF-2.3.1-21-2022-00004; and by the TKP2021-NKTA-01 grant of the National Research, Development and Innovation Office, Hungary.

*Corresponding author: horvath.balint@sztaki.hu (should be participant of 23rd EYSM)

1 Introduction

Constructing confidence bands for the regression function from a finite sample of input-output data is a core problem in statistics and machine learning [1]. In a parametric setting, such region estimates are typically induced by confidence sets in the parameter space, however, in a nonparameteric setting this indirect approach is often infeasible, which calls for direct constructions.

The problem comes with a fairly standard setting. We are given a finite i.i.d. sample of input-output pairs, $(x_1, y_1), \dots, (x_n, y_n)$, having an unknown joint distribution $\mathbb{P}_{X,Y}$, where $x_k \in \mathbb{R}^m, y_k \in \mathbb{R}$ and $\mathbb{E}[y_k^2] < \infty$. We assume that $y_k = f_*(x_k) + \varepsilon_k$, for $k \in [n] = \{1, ..., n\}$, where $\{\varepsilon_k\}$ represent the (measurement) noise terms on the true regression function f_* with $\mathbb{E}[\varepsilon_k] = 0$.

Our primary goal is the following: we are looking for an $I : \mathbb{R}^m \to \mathbb{R} \times \mathbb{R}$ function, such that $I(x) = (I_1(x), I_2(x))$ specifies the endpoints of an interval estimate for $f_*(x)$, where $x \in \mathbb{R}^m$. The aim is to construct I with the property:

$$\nu(I) \doteq \mathbb{P}\big(I_1(x) \leq f_*(x) \leq I_2(x), \text{ for } \mathbb{P}_X\text{-a.e. } x \in \mathbb{R}^m \big) \geq 1 - \alpha,$$

where $\alpha \in (0, 1)$ is a (user-chosen) risk probability. Since the confidence band I depends on the sample, typically we have $\mathbb{P} = \mathbb{P}_{X,Y}^{\otimes n}$. We call $\nu(I)$ the *reliability* of the confidence band. Next, we define Paley-Wiener spaces [2].

Definition 1. A *Paley-Wiener space $\mathcal{H}$ is a subspace of $\mathcal{L}^2(\mathbb{R}^m)$, where for each $\varphi \in \mathcal{H}$ the support of the Fourier transform of φ is included in a given hypercube $[-\eta, \eta]^m$, where $\eta > 0$ is a hyper-parameter.*

Paley-Wiener spaces are *Reproducing Kernel Hilbert Spaces* (RKHSs) with the following reproducing kernel function. For all $u, v \in \mathbb{R}^m$:

$$k(u, v) \doteq \pi^{-m} \prod\nolimits_{j=1}^{m} \frac{\sin(\eta(u_j - v_j))}{u_j - v_j},$$

where, for convenience, $\sin(\eta \cdot 0)/0$ is defined to be η. Henceforth, we work with the *Paley-Wiener kernel* defined above and denote our RKHS by $\mathcal{H}$.

Our fundamental assumptions are as follows:

A1. *The sample $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^m \times \mathbb{R}$ is i.i.d., and $\mathbb{E}[y_1^2] < \infty$.*

A2. *For $k \in [n]$, $\mathbb{E}\big[\varepsilon_k\big] = 0$, variables $\{x_k\}$ and $\{\varepsilon_k\}$ are independent, and there is a compact matrix group, $\mathcal{G} \subseteq \mathbb{R}^{n \times n}$, such that $\forall\, G \in \mathcal{G} : G\,\varepsilon \overset{\mathrm{d}}{=} \varepsilon$.*

A3. *The probability distribution of the inputs, $\{x_k\}$, is a priori known, it is absolutely continuous, and its density, h_*, satisfies $h_*(x) > 0$, $\forall x \in \mathbb{R}^m$.*

A4. *The regression function f_* is from a Paley-Wiener space and there is a (universal) constant $\rho > 0$, such that for all $x \in \mathbb{R}^m$, $f_*^2(x) \leq \rho\, h_*(x)$.*

The third part of A2 can be easily satisfied, e.g., by the group of permutation matrices [3], as $\{\varepsilon_k\}$ are i.i.d.; A4 ensures that the observations are informative.

2 Noise-Free Outputs

We start by studying a simplified problem: when the true regression function, f_*, is observed perfectly at random inputs, that is $\forall k \in [n] : y_k = f_*(x_k)$.

Since A1 and A3 guarantee that the inputs $\{x_k\}$ are almost surely distinct, the element from $\mathcal{H}$, which interpolates every y_k output and the corresponding x_k input, and which has the smallest possible kernel norm, that is

$$\bar{f} \doteq \arg\min \left\{ \| f \|_{\mathcal{H}} : f \in \mathcal{H} \ \& \ \forall k \in [n] : f(x_k) = y_k \right\},$$

exists and it takes the following form for all possible inputs $x \in \mathbb{R}^m$:

$$\bar{f}(x) = \sum_{k=1}^{n} \hat{\alpha}_k k(x, x_k),$$

where the weights are $\hat{\alpha} = K^{-1} y$ with $y \doteq (y_1, ..., y_n)^{\mathrm{T}}$ and $\hat{\alpha} := (\hat{\alpha}_1, ..., \hat{\alpha}_n)$, and $K_{i,j} = k(x_i, x_j)$ is the kernel or Gram matrix. Note that under A1, A3 and A4, the Gram matrix is almost surely invertible.

The main idea behind our approach is as follows. First, we need to estimate how "smooth" the function is, which is measured by $\|f_*\|_{\mathcal{H}} = \|f_*\|_2$.

Lemma 1. *Assuming A1, A3, A4 and that $y_k = f_*(x_k)$ for $k \in [n]$, for any risk probability $\alpha \in (0,1)$, we have $\mathbb{P}(\|f_*\|_{\mathcal{H}}^2 \leq \kappa) \geq 1 - \alpha$, with*

$$\kappa \doteq \frac{1}{n} \sum_{k=1}^{n} \frac{y_k^2}{h_*(x_k)} + \rho \sqrt{\frac{\ln \alpha}{-2n}}.$$

This statement can be proved analogously to the similar Lemma 1 in [3].

To test a "candidate" $(x_0, y_0) \in \mathbb{R}^m \times \mathbb{R}$ input-output pair, we can compute the minimum norm needed to interpolate the original $\{(x_k, y_k)\}, k \in [n]$ combined with $(x_0, y_0) \in \mathbb{R}^m \times \mathbb{R}$. The *minimum norm interpolation* of $(x_0, y_0), \ldots, (x_n, y_n)$ is now $\tilde{f}(x) = \sum_{k=0}^{n} \tilde{\alpha}_k k(x, x_k)$, where the weights are $\tilde{\alpha} = K_0^{-1} \tilde{y}$ with $\tilde{y} \doteq (y_0, y_1, \ldots, y_n)^{\mathrm{T}}$, $\tilde{\alpha} \doteq (\tilde{\alpha}_0, \ldots, \tilde{\alpha}_n)^{\mathrm{T}}$, and $K_0(i+1, j+1) = k(x_i, x_j)$ is the extended kernel matrix. Since $\mathcal{H}$ is an RKHS, we have

$$\| \tilde{f} \|_{\mathcal{H}}^2 = \tilde{\alpha}^{\mathrm{T}} K_0 \tilde{\alpha} = \tilde{y}^{\mathrm{T}} K_0^{-1} K_0 K_0^{-1} \tilde{y} = \tilde{y}^{\mathrm{T}} K_0^{-1} \tilde{y}.$$

For a candidate (x_0, y_0) input-output pair, we first calculate the norm square of the minimum norm interpolation of $\{(x_0, y_0)\} \cup \{(x_k, y_k)\}$. Then, if this norm square is less than or equal to our estimate (denoted by κ), we include (x_0, y_0) in our confidence band, otherwise, (x_0, y_0) is not included in the band.

To obtain the interval endpoints for a given query input x_0, we have to calculate the highest and lowest y_0 values which can be interpolated with a function from $\mathcal{H}$ having at most norm square κ. This leads to the problems:

$$\begin{aligned} &\min / \max \quad y_0 \\ &\text{subject to} \quad (y_0, y^{\mathrm{T}}) K_0^{-1} (y_0, y^{\mathrm{T}})^{\mathrm{T}} \leq \kappa, \end{aligned} \tag{1}$$

where "min / max" means that we have to solve the problem as a minimization and also as a maximization (separately). The problems in (1) are convex and their solutions can be calculated analytically [3]. The optimal values, denoted by $y_{\min}$ and $y_{\max}$, respectively, determine the *endpoints* of the confidence interval for $f_*(x_0)$, that is $I_1(x_0) \doteq y_{\min}$ and $I_2(x_0) \doteq y_{\max}$. If there is no solution, we return $I(x_0) = \emptyset$. We conclude with the following theorem.

Theorem 1. *Assume A1, A3, A4 and that $y_k = f_*(x_k)$ for all k. Then, for any risk probability $\alpha \in (0,1)$ and for any finite sample size n, the constructed confidence band is guaranteed to have the reliability $\nu(I) \geq 1 - \alpha$.*

Proof sketch. According to Lemma 1, $\mathbb{P}(\|f_*\|_{\mathcal{H}}^2 \leq \kappa) \geq 1 - \alpha$. If $\|f_*\|_{\mathcal{H}}^2 \leq \kappa$, then for all x_0, the value $f_*(x_0)$ is in the confidence band, since f_* interpolates the sample extended with $(x_0, f_*(x_0))$, and its norm is $\leq \kappa$, thus the minimum norm interpolant of this extended sample inherits this norm bound. □

3 Noisy Outputs

Next, we provide our solution for the case when the outputs are affected by measurement noises which satisfy A2. A new problem is that we do not observe the true function values at the sample inputs. However, we can apply the KGP method [5] to construct an ellipsoid $\mathcal{Z}$ with the guarantee $\mathbb{P}\big((f_*(x_1), \dots, f_*(x_d))^{\mathrm{T}} \in \mathcal{Z} \big) \geq 1 - \beta$, for a given $\beta \in (0,1)$ and $d \leq n$ [4].

In order to get an estimate of $\|f_*\|_{\mathcal{H}}^2$, we can solve the following problem

$$\text{maximize} \quad \frac{1}{d} \sum_{k=1}^{d} \frac{z_k^2}{h_*(x_k)} \qquad \text{subject to} \;\; z \in \mathcal{Z}. \tag{2}$$

This problem is not convex, but due to strong duality, we can solve its convex dual instead. The construction is analogous to the one in [4, Section 6.1].

Lemma 2. *Under A1, A2, A3, A4 and for any $\alpha, \beta \in (0,1)$ risk probabilities,*

$$\mathbb{P}\big((f_*(x_1), \dots, f_*(x_d))^{\mathrm{T}} \in \mathcal{Z} \,\wedge\, \|f_*\|_{\mathcal{H}}^2 \leq \tau \big) \geq 1 - \alpha - \beta,$$

where $\tau \doteq \xi + \rho\sqrt{\ln(\alpha)/(-2d)}$ and ξ is the optimal value of problem (2).

Given ellipsoid $\mathcal{Z}$ which contains with high probability the true outputs of f_* at the *sample* inputs $\{x_k\}_{k=1}^d$, we can construct a confidence interval for $f_*(x_0)$ at *any* query input x_0 by computing the maximum and the minimum potential output $z_0 \in \mathbb{R}$ at x_0, for which there is an interpolant that interpolates the sample $\{(x_0, z_0)\} \cup \{(x_k, z_k)\}_{k=1}^d$, for a $z \in \mathcal{Z}$, and has a norm square $\leq \tau$, i.e., our upper bound for $\|f_*\|_{\mathcal{H}}^2$. This leads to the (convex) problems

$$\begin{aligned} \min / \max \quad & z_0 \\ \text{subject to} \quad & (z_0, z_1, \dots, z_d) K_0^{-1} (z_0, z_1, \dots, z_d)^{\mathrm{T}} \leq \tau \\ & (z_1, ..., z_d) \in \mathcal{Z}. \end{aligned} \tag{3}$$

Let $z_{\min}$ and $z_{\max}$ be the optimal values of (3). The confidence interval for $f_*(x_0)$ is given by $[z_{\min}, z_{\max}]$. We return $I(x_0) = \emptyset$ if (3) is infeasible.

Theorem 2. *Assume that A1, A2, A3 and A4 are satisfied. Then, for any risk probabilities $\alpha, \beta \in (0,1)$ and for any finite sample size n, the constructed confidence band is guaranteed to have the reliability $\nu(I) \geq 1 - \alpha - \beta$.*

Proof sketch. The core idea of the proof is very similar to that of Theorem 1. According to Lemma 2, the event A that both $(f_*(x_1), \dots, f_*(x_d))^{\mathrm{T}} \in \mathcal{Z}$ as well as $\|f_*\|_{\mathcal{H}}^2 \leq \tau$ happen has probability at least $1 - \alpha - \beta$.

Conditioning on event A, for all query input point x_0 such that K_0 is invertible, which holds a.e., we can guarantee that there is a $z \in \mathcal{Z}$, namely $z = (f_*(x_1), \dots, f_*(x_d))^{\mathrm{T}}$, and z_0, name $z_0 = f_*(x_0)$, such that the *minimum norm* interpolant of $\{(x_0, z_0)\} \cup \{(x_k, z_k)\}_{k=1}^n$ has a norm square $\leq \tau$, since f_* intself is an interpolant of this dataset. Thus, we have that $z_{\min} \leq z_0 \leq z_{\max}$, for $z_0 = f_*(x_0)$. This property is always guaranteed, hence, we *simultaneously* have for $\mathbb{P}_X$-a.e. $x_0 \in \mathbb{R}^m$ that $z_{\min}(x_0) \leq f_*(x_0) \leq z_{\max}(x_0)$, under A. □

4 Numerical Experiments

The methods were also tested and implemented numerically. The Paley-Wiener RKHS was used with parameter $\eta = 30$ and the original data-generating function was created as follows: 20 random input points $\{\bar{x}_k\}_{k=1}^{20}$ were generated, with uniform distribution on $[-1, 1]$. Then we created $f_*(x) = \sum_{k=1}^{20} w_k k(x, \bar{x}_k)$, where each w_k had a uniform distribution on $[-1, 1]$. The function was normalized, in case its maximum value exceeded 1.

A sample with $n = 300$ random noisy observations from f_* was generated. The inputs $\{x_k\}$ followed Laplace distribution with location $\mu = 0$ and scale $b = 0.3$ parameters, while the measurement noise $\{\varepsilon_k\}$ had the following distribution: first, we experimented with a non-symmetric dase, where $\varepsilon \sim \exp(\lambda) - 1/\lambda$, where $\lambda = 0.3$, then we implemented an experiment with a symmetrically distributed noise, namely ε had Laplace distribution with $\mu = 0$ and $b = 0.3$ parameters. Both of these statistical setups satisfy A2.

Figure 1 demonstrates that the proposed approach leads to feasible and informative simultaneous (nonparametric, nonasymptotic) confidence bands.

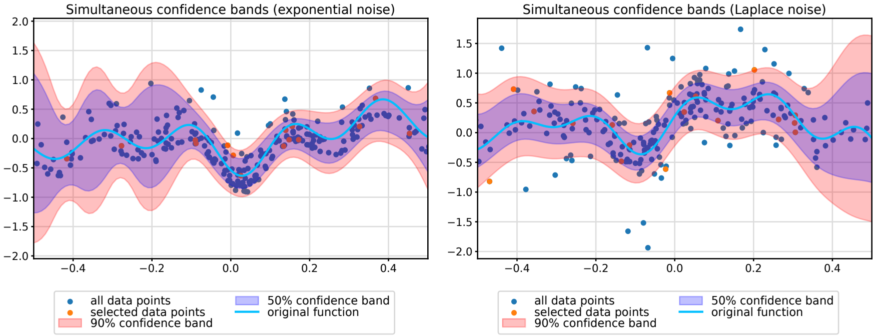

Figure 1: Nonparametric, nonasymptotic, simultaneous confidence bands with Laplace distributed inputs and symmetric and non-symmetric noises.

References

[1] Knafl, G., Sacks, J., Ylvisaker, D. (1985). Confidence bands for regression functions. *Journal of the American Statistical Association*, 80, 683-691.

[2] Yang, J., Sindhwani, V., Avron, H., Mahoney, M. (2014). Quasi-Monte Carlo feature maps for shift-invariant kernels. *In International Conference on Machine Learning* (pp. 485-493). PMLR.

[3] Csáji, B. Cs., Horváth, B. (2022). Nonparametric, Nonasymptotic Confidence Bands With Paley-Wiener Kernels for Band-Limited Functions, *IEEE Control Systems Letters*, IEEE, 6, 3355-3360.

[4] Csáji, B. Cs., Horváth, B. (2023) Improving Kernel-Based Nonasymptotic Simultaneous Confidence Bands, *22nd IFAC World Congress*, Yokohama, Japan, July 9-14, 2023.

[5] Csáji, B. Cs., Kis, K. B. (2019). Distribution-Free Uncertainty Quantification for Kernel Methods by Gradient Perturbations. *Machine Learning*, Springer, 108(8-9), 1677-1699.