
# Distribution-Free Guarantees For Kernel Methods

Balázs Csanád Csáji

Institute for Computer Science and Control (SZTAKI)

Institute of Mathematics, Eötvös Loránd University (ELTE)

Joint research with Krisztián Kis and Bálint Horváth

Mathematical Modelling Seminar, BME, September 6, 2022

# I. Introduction

## Kernels in Machine Learning

# Kernel Methods in Machine Learning

## I. Supervised Learning

Learning from a sample of (typically noisy) input-output data.
Problems, e.g., classification, regression and experiment design.

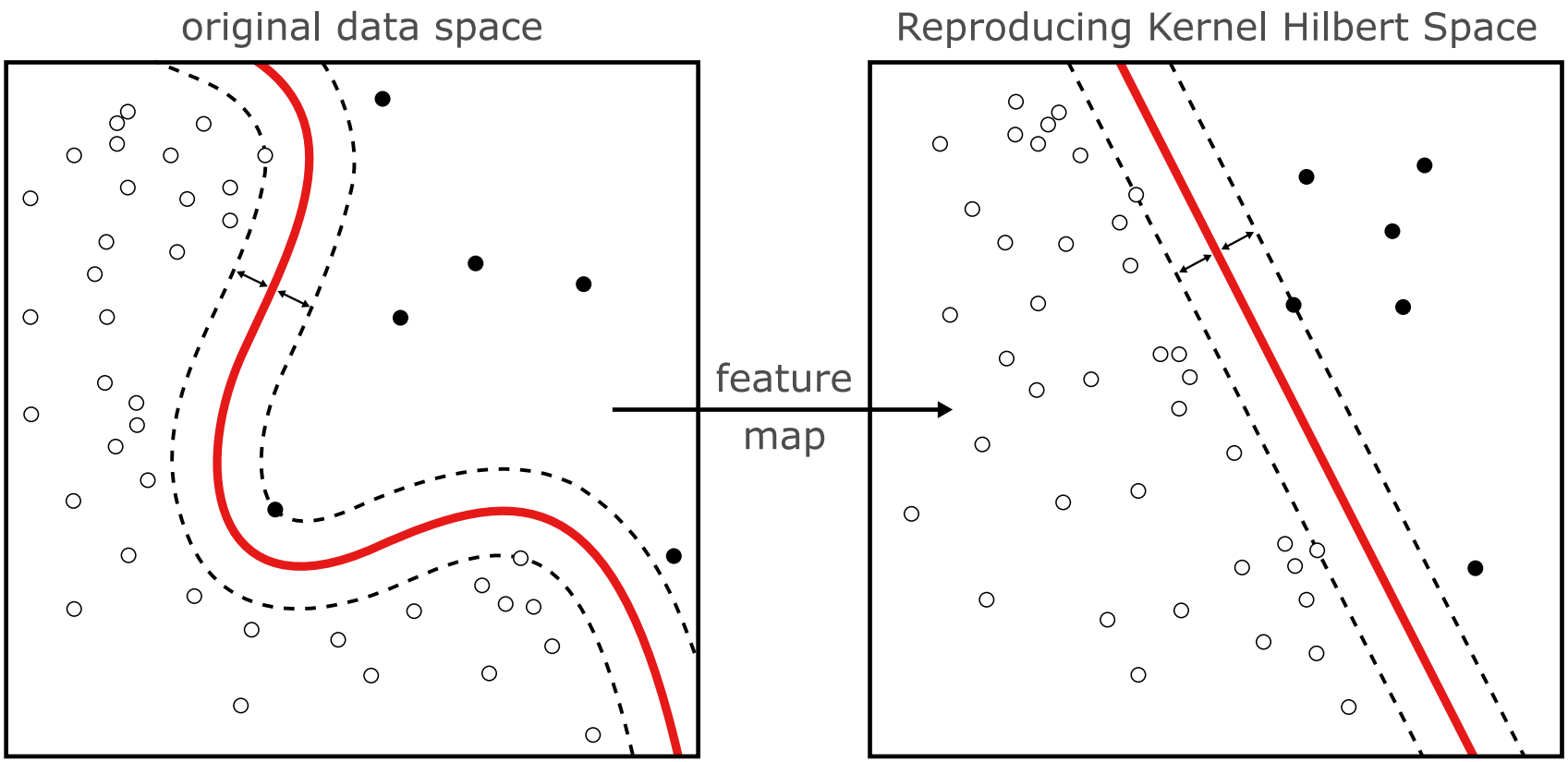
## II. Unsupervised Learning

Learning from a sample of unlabelled data (raw data, no outputs).
Problems, e.g., density estimation, clustering, and dim. reduction.

## III. Reinforcement Learning

Learning via interactions with an uncertain, dynamic environment.
Problems, e.g., (partially observable) Markov decision processes.

# Lifting the Data into a Higher Dimensional Space



# Reproducing Kernel Hilbert Spaces

- A Hilbert space, $\mathcal{H}$, of functions $f : \mathcal{X} \to \mathbb{R}$, with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, is called a Reproducing Kernel Hilbert Space (RKHS), if $\forall\, z \in \mathcal{X}$ the point evaluation (Dirac) functional $\delta_z : f \to f(z)$ is bounded (i.e., $\forall z \colon \exists\, \kappa > 0$ with $|\delta_z(f)| \leq \kappa \|f\|_{\mathcal{H}}$ for all $f \in \mathcal{H}$).
- Then, one can construct a kernel, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, having the reproducing property, that is for all $z \in \mathcal{X}$ and $f \in \mathcal{H}$, we have

$$\langle\, k(\cdot, z), f \,\rangle_{\mathcal{H}} \;=\; f(z),$$

  which is ensured by the Riesz-Fréchet representation theorem.
- As a special case, the kernel satisfies $k(z, s) = \langle\, k(\cdot, z), k(\cdot, s) \,\rangle_{\mathcal{H}}$.
- A kernel is therefore a symmetric and positive-definite function.
- Conversely, by the Moore-Aronszajn theorem, for every symmetric and positive definite function, there uniquely exists an RKHS.

# Examples of Kernels

| Kernel | $k(x,y)$ | Domain | U | C |
|---|---|---|---|---|
| Gaussian | $\exp\left(\frac{-\|x-y\|_2^2}{\sigma}\right)$ | $\mathbb{R}^d$ | ✓ | ✓ |
| Linear | $\langle x,y\rangle$ | $\mathbb{R}^d$ | × | × |
| Polynomial | $(\langle x,y\rangle + c)^p$ | $\mathbb{R}^d$ | × | × |
| Laplacian | $\exp\left(\frac{-\|x-y\|_1}{\sigma}\right)$ | $\mathbb{R}^d$ | ✓ | ✓ |
| Rat. quadratic | $\exp(\|x-y\|_2^2 + c^2)^{-\beta}$ | $\mathbb{R}^d$ | ✓ | ✓ |
| Exponential | $\exp(\sigma\langle x,y\rangle)$ | compact | × | ✓ |
| Poisson | $1/(1-2\alpha\cos(x-y)+\alpha^2)$ | $[0,2\pi)$ | ✓ | ✓ |

Table: typical kernels; *U* means "universal" and *C* means "characteristic" (where the hyper-parameters satisfy $\sigma, \beta, c > 0$, $\alpha \in (0,1)$ and $p \in \mathbb{N}$).

# Kernel Norm as Smoothness Measure

– By the reproducing property and the Cauchy-Schwartz inequality:

$$|f(x) - f(x')| = |\langle f, k_x - k_{x'}\rangle_{\mathcal{H}}| \leq \|f\|_{\mathcal{H}} \|k_x - k_{x'}\|_{\mathcal{H}}$$
$$\leq \|f\|_{\mathcal{H}} d(x, x')$$

where $k_x \doteq k(\cdot, x)$ and the (kernel-dependent) distance is

$$d(x, x') = \sqrt{k(x, x) + k(x', x') - 2k(x, x')}$$

– Example: if $\mathcal{H} = \mathbb{R}^d$ and $\langle x, x'\rangle_{\mathcal{H}} = x^{\mathrm{T}} x'$, then $d(x, x') = \|x - x'\|_2$ (choosing the linear kernel yields the standard Euclidean distance)

– Therefore, functions in $\mathcal{H}$ satisfy a Lipschitz-like condition.

– The kernel norm, $\|f\|_{\mathcal{H}}$, acts as a measure of smoothness of $f$.

– The precise notation of smoothness depends on the chosen kernel.

# Regression Function and Gram Matrix

- The data sample, $\mathcal{Z}$, is a finite sequence of input-output data

$$(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathbb{R}$$

  where $\mathcal{X} \neq \emptyset$ and $\mathbb{R}$ are the input and output spaces, respectively.
- We are searching for a model for this data in an RKHS containing $f : \mathcal{X} \to \mathbb{R}$ functions. The kernel of the RKHS is $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.
- We set $x \doteq (x_1, \ldots, x_n)^{\mathrm{T}} \in \mathcal{X}^n$ and $y \doteq (y_1, \ldots, y_n)^{\mathrm{T}} \in \mathbb{R}^n$.
- The Gram matrix of the kernel with respect to inputs $\{x_i\}$ is

$$[K]_{i,j} \doteq k(x_i, x_j).$$

  (a data-dependent symmetric and positive semi-definite matrix)
- A kernel is called strictly positive definite if its Gram matrix, $K$, is (strictly) positive definite for all possible distinct inputs $\{x_i\}$.

# Minimum Norm Interpolation with Kernels

For a (finite) dataset $\{(x_k, y_k)\}$, where inputs $\{x_k\}$ are distinct, the element from $\mathcal{H}$ that has the minimum norm and interpolates each output $y_k$ at the corresponding input $x_k$, that is

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

takes the following (finite dimensional) form, for all input $x \in \mathbb{X}$ :

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

where (assuming $K$ is invertible) the optimal coefficients are

$$\bar{\alpha} = K^{-1} y,$$

with $y \doteq (y_1, \ldots, y_n)^{\mathrm{T}} \in \mathbb{R}^n$ and $\bar{\alpha} \doteq (\bar{\alpha}_1, \ldots, \bar{\alpha}_n)^{\mathrm{T}} \in \mathbb{R}^n$.

# Regression and Classification

(1) The data sample, $\mathcal{Z}$, is a finite sequence of input-output data

$$(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are the input and output spaces, respectively.

If $|\mathcal{Y}| < \infty$, it is called "classification", otherwise "regression".

(2) The model class, $\mathcal{F}$, is a space of $f : \mathcal{X} \to \mathcal{Y}$ functions.

(3) A criterion or objective function is $\mathcal{V} : \mathcal{F} \times \mathcal{D} \to [0, \infty)$, where $\mathcal{D}$ is the space of possible data samples.

## Regression Model (Point Estimate)

$$\widehat{f} \doteq \operatorname*{arg\,min}_{f \in \mathcal{F}} \mathcal{V}(f, \mathcal{Z}) = \widehat{f}(\mathcal{V}, \mathcal{F}, \mathcal{Z})$$

# Regularized Optimization Criterion

## Regularized Criterion

$$g(f, \mathcal{Z}) = \mathcal{L}(x_1, y_1, f(x_1), \ldots, x_n, y_n, f(x_n)) + \Omega(f)$$

- The loss function, $\mathcal{L}$, measures how well the model fits the data, while the regularizer, $\Omega$, controls other properties of the solution.
- Regularization can help in several issues, for example:
  - To convert an ill-posed problem to a well-posed problem.
  - To make an ill-conditioned approach better conditioned.
  - To reduce over-fitting and thus to help the generalization.
  - To force the sparsity of the solution.
  - Or in general to control shape and smoothness.

# Representer Theorem

We are given a sample, $\mathcal{Z}$, a positive-definite kernel $k(\cdot,\cdot)$, an associated RKHS with a norm $\|\cdot\|_{\mathcal{H}}$ induced by $\langle\cdot,\cdot\rangle_{\mathcal{H}}$, and a class

$$\mathcal{F} \doteq \left\{ f \mid f(z) = \sum_{i=1}^{\infty} \beta_i k(z, z_i),\ \beta_i \in \mathbb{R},\ z_i \in \mathcal{X},\ \|f\|_{\mathcal{H}} < \infty \right\},$$

then, for any mon. increasing regularizer, $\Omega : [0,\infty) \to [0,\infty)$, and an arbitrary loss function $\mathcal{L} : (\mathcal{X} \times \mathbb{R}^2)^n \to \mathbb{R} \cup \{\infty\}$, the criterion

$$g(f, \mathcal{Z}) \doteq \mathcal{L}\big( (x_1, y_1, f(x_1)), \ldots, (x_n, y_n, f(x_n)) \big) + \Omega( \|f\|_{\mathcal{H}} )$$

has a minimizer admitting the following representation

$$f_\alpha(z) = \sum_{i=1}^{n} \alpha_i k(z, x_i),$$

where $\alpha \doteq (\alpha_1, \ldots, \alpha_n)^{\mathrm{T}} \in \mathbb{R}^n$ is a finite vector of coefficients.

# Kernel Methods as a General Framework

By choosing the kernel and the criterion function, several machine learning approaches can be recovered as special cases, such as

- Polynomial regression
- Logistic regression (kernelized)
- Support vector classification and regression
- Multi-layer perceptrons
  (feedforward neural networks with one hidden layer)
- Radial basis function networks
  (e.g., Gaussian, multiquadric, inverse multiquadric)
- Gaussian process regression
- Thin plate splines
- Principal component analysis (kernelized)

# Uncertainty Quantification

- In practice often some quality tag is needed to judge the estimate.
- Safety, stability, or quality requirements? $\Rightarrow$ confidence regions

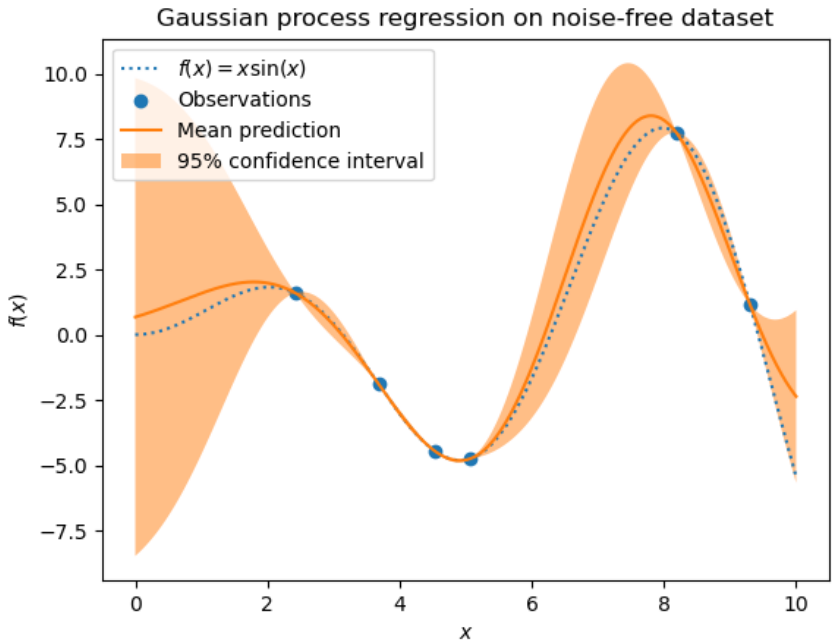
**Confidence Region (Level $\mu$)**

$$\mathbb{P}\left(f_0 \in \widehat{\Theta}_{\mathcal{Z},\mu}\right) \geq 1 - \mu$$

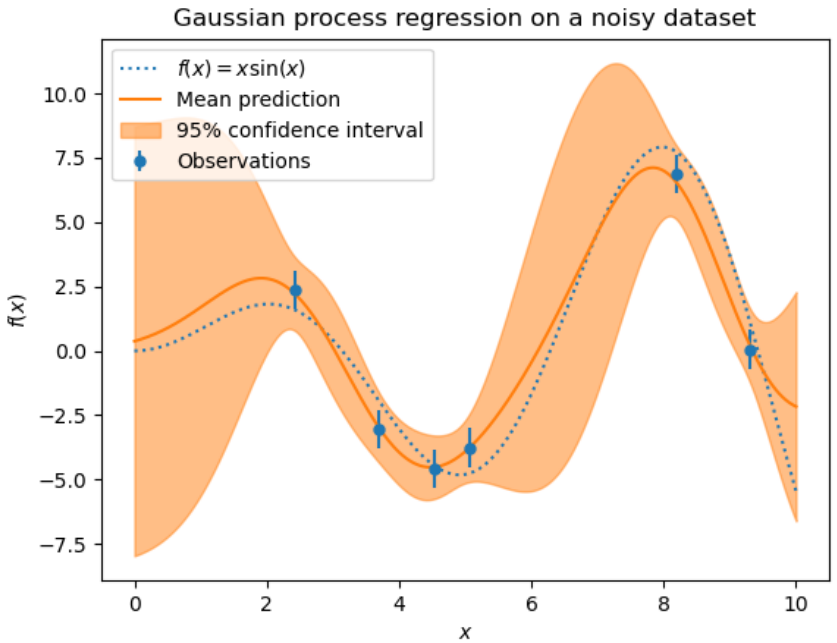
for some risk probability $\mu \in (0,1)$, where $f_0$ is a target function, e.g., the "true" regression function generating the data or some "good" representation of it in the model space.

- Needed for robust decisions, risk management, active learning, etc.
- Typically the level sets of the (scaled) limiting distribution is used.
- Issues with using asymptotic distributions: only approximately correct for finite samples, requires the existence of a (known) limit.

# Region Estimation with Gaussian Process Regression




(a) GPR: Noise-Free Observations
(source: scikit-learn website)




(b) GPR: Noisy Observations
(source: scikit-learn website)

– Issues with GPR: assumes that the data is jointly Gaussian (which is sometimes unrealistic), therefore, it is not distribution-free.

# II. Guarantees for Ideal Representations

## Distribution-Free Confidence Sets for Ideal Interpolants

Joint work with: Krisztián Balázs Kis

# Confidence Sets for Ideal Representations

- Kernel methods are widely used in machine learning and related fields (such as signal processing and system identification).
- Besides how to construct a models from empirical data, it is also a fundamental issue how to quantify the uncertainty of the model.
- Standard solutions either use strong distributional assumptions (e.g., Gaussian processes) or heavily rely on asymptotic results.
- Here, a new construction for non-asymptotic and distribution-free confidence sets for models built by kernel methods are proposed.
- We target the ideal representation of the underlying true function.
- The constructed regions have exact coverage probabilities and only require a mild regularity (e.g., symmetry or exchangeability).
- The quadratic case with symmetric noises has special importance.
- Several examples are discussed, such as support vector machines.

# Ideal Representations

– Sample $\mathcal{Z}$ is generated by an underlying true function $f_*$

$$y_i \doteq f_*(x_i) + \varepsilon_i,$$

for $i = 1, \ldots, n$, where $\{x_i\}$ inputs and $\{\varepsilon_i\}$ are the noise terms.

– The vector of noises is denoted by $\varepsilon \doteq (\varepsilon_1, \ldots, \varepsilon_n)$.

– In an RKHS, we can focus on, $f_\alpha(z) = \sum_{i=1}^n \alpha_i k(z, x_i)$ functions.

– Function $f_\alpha \in \mathcal{F}$ is called an ideal representation of $f_*$ w.r.t. $\mathcal{Z}$, if

$$f_\alpha(x_i) = f_*(x_i), \qquad \text{for all} \qquad x_1, \ldots, x_n$$

the corresponding ideal coefficients are denoted by $\alpha^* \in \mathbb{R}^n$.

– Gram matrix is positive-definite $\Rightarrow$ exactly one ideal represent.

– We aim at building confidence regions for ideal representations, instead of the true function (which may not be in the RKHS).

# Distributional Invariance

– Our approach does not need strong distributional assumption on the noises (such as Gaussianity). The needed property is:

An $\mathbb{R}^n$-valued random vector $\varepsilon$ is distributionally invariant w.r.t. a compact group of transformations, $(\mathcal{G}, \circ)$, where "$\circ$" denotes the function composition and each $G \in \mathcal{G}$ maps $\mathbb{R}^n$ to itself, if for all $G \in \mathcal{G}$, vectors $\varepsilon$ and $G(\varepsilon)$ have the same distribution.

– Two arch-typical examples having this property are

(1) If $\{\varepsilon_i\}$ are exchangeable (for example: i.i.d.), then we can use the (finite) group of permutations on the noise vector.

(2) If $\{\varepsilon_i\}$ independent and symmetric, then we can apply the group consisting sign-changes for any subsets of the noises.

# Main Assumptions

**A1** The kernel is strictly positive definite and $\{x_i\}$ are a.s. distinct.

**A2** The input vector $x$ and the noise vector $\varepsilon$ are independent.

**A3** The noises, $\{\varepsilon_i\}$, are distributionally invariant with respect to a known group of transformations, $(\mathcal{G}, \circ)$.

**A4** The gradient, or a subgradient, of the objective w.r.t. $\alpha$ exists and it only depends on $y$ through the residuals, i.e., there is $\bar{g}$,

$$\nabla_\alpha g(f_\alpha, \mathcal{Z}) \;=\; \bar{g}(x, \alpha, \widehat{\varepsilon}(x, y, \alpha)),$$

where the residuals are defined as $\widehat{\varepsilon}(x, y, \alpha) \;\doteq\; y - K\,\alpha$.

(A1 $\Rightarrow$ the ideal representation is unique with prob. one; A2 $\Rightarrow$ no autoregression; A3 $\Rightarrow$ $\varepsilon$ can be perturbed; A4 holds in most cases.)

# Perturbed Gradients

– Let us define a reference "evaluation" function, $Z_0 : \mathbb{R}^n \to \mathbb{R}$, and $m-1$ perturbed "evaluation" functions, $\{Z_i\}$, with $Z_i : \mathbb{R}^n \to \mathbb{R}$,

$$Z_0(\alpha) \doteq \| \Psi(x)\, \bar{g}(x, \alpha, \widehat{\varepsilon}(x, y, \alpha)) \|^2,$$

$$Z_i(\alpha) \doteq \| \Psi(x)\, \bar{g}(x, \alpha, G_i(\widehat{\varepsilon}(x, y, \alpha))) \|^2,$$

for $i = 1, \ldots, m-1$, where $m$ is a hyper-parameter, $\Psi(x)$ is an (optional, possibly input dependent) weighting matrix, and $\{G_i\}$ are (random) uniformly sampled i.i.d. transformations from $\mathcal{G}$.

– If $\alpha = \alpha^* \Rightarrow Z_0(\alpha^*) \overset{d}{=} Z_i(\alpha^*)$, for all $i = 1, \ldots, m-1$ ("$\overset{d}{=}$" denotes equality in distribution; observe that $\widehat{\varepsilon}(x, y, \alpha^*) = \varepsilon$).

– If $\alpha \neq \alpha^*$, this distributional equivalence does not hold, and if $\| \alpha - \alpha^* \|$ is large enough, $Z_0(\alpha)$ will dominate $\{Z_i(\alpha)\}_{i=1}^{m-1}$.

# Confidence Regions

– The normalized rank of $\|Z_0(\alpha)\|^2$ in the ordering of $\{\|Z_i(\alpha)\|^2\}$ is

$$\mathcal{R}(\alpha) \doteq \frac{1}{m}\left[1 + \sum_{i=1}^{m-1} \mathbb{I}\left(\|Z_i(\alpha)\|^2 \prec \|Z_0(\alpha)\|^2\right)\right],$$

where $\mathbb{I}(\cdot)$ is an indicator function, and binary relation "$\prec$" is the standard "$<$" ordering with random tie-breaking (pre-generated).

– Given any $p \in (0,1)$ with $p = 1 - q/m$, a confidence regions is

**Confidence Region for the Ideal Coefficient Vector**

$$A_p \doteq \left\{\alpha \in \mathbb{R}^n : \mathcal{R}(\alpha) \leq 1 - \frac{q}{m}\right\}$$

where $0 < q < m$ are user-chosen integers (hyper-parameters).

# Main Theoretical Result: Exact Coverage

Theorem: Under assumptions A1, A2, A3 and A4, the coverage probability of $A_p$ with respect to the ideal coefficient vector $\alpha^*$ is

$$\mathbb{P}\big(\alpha^* \in A_p\big) = p = 1 - \frac{q}{m},$$

for any choice of the integer hyper-parameters, $0 < q < m$.

- The coverage probabiltiy is exact (it is non-conservative), and as $m$ and $q$ are user-chosen, probability $p$ is under our control.
- The result is non-asymptotic, as it is valid for any finite sample.
- Furthermore, no particular distribution is assumed for the noises affecting measurements, hence the ideas are distribution-free.
- The needed statistical assumptions are very mild, for example, the noises can be non-stationary, heavy-tailed, and skewed.

# Quadratic Objectives and Symmetric Noises

– Assume the noises are independent and symmetric and the objective is convex quadratic taking the (canonical) form

$$g(\alpha) \doteq \| z - \Phi\alpha \|^2$$

where $z$ is the vector of outputs, and $\Phi$ is the regressor matrix.

**Evaluation Function of Sign-Perturbed Sums (SPS)**

$$Z_i(\alpha) \doteq \left\| (\Phi^{\mathrm{T}}\Phi)^{-1/2} \Phi^{\mathrm{T}} G_i (z - \Phi\alpha) \right\|^2$$

where $G_i = \mathrm{diag}(\sigma_{i,1}, \ldots, \sigma_{i,n})$, for $i \neq 0$, where $\{\sigma_{i,j}\}$ are i.i.d. Rademacher variables, they take $+1$ and $-1$ with probability $1/2$.

– The SPS confidence regions are star convex with the least-squares estimate as a center, and have ellipsoidal outer approximations.

# Least-Squares Support Vector Classification

– The primal form of (soft-margin) LS-SVM classification is

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} w^{\mathrm{T}} w + \lambda \sum_{k=1}^{n} \xi_k^2 \\ \text{subject to} \quad & y_k (w^{\mathrm{T}} x_k + b) = 1 - \xi_k \end{aligned}$$

for $k = 1, \ldots, n$, where $\lambda > 0$ is fixed. This convex quadratic optimization problem can be rewritten, with $\alpha \doteq (b, w^{\mathrm{T}})^{\mathrm{T}}$, as

$$g(\alpha) = \frac{1}{2} \| B\alpha \|^2 + \lambda \| \mathbb{1}_n - y \odot (X\alpha) \|^2,$$

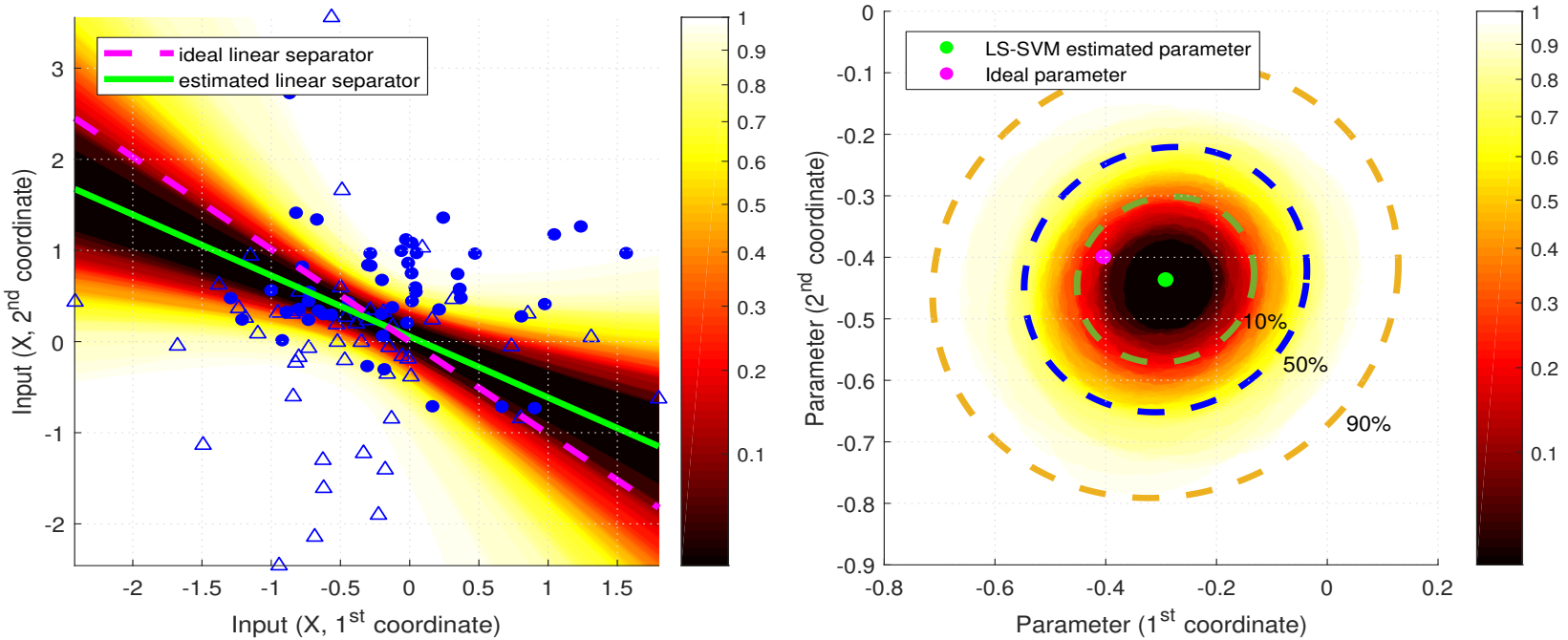
where $\mathbb{1}_n \in \mathbb{R}^n$ is the all-one vector, "$\odot$" denotes the Hadamard (entrywise) product, $X \doteq [\tilde{x}_1, \ldots, \tilde{x}_n]^{\mathrm{T}}$ with $\tilde{x}_k \doteq [1, x_k^{\mathrm{T}}]^{\mathrm{T}}$ and $B \doteq \mathrm{diag}(0, 1, \ldots, 1)$, the role of matrix $B$ is to remove bias $b$.

# Experiment: Confidence Sets for LS-SVC

– This can be further reformulated to have the form $\| z - \Phi\alpha \|^2$,

$$\Phi = \begin{bmatrix} \sqrt{\lambda}\,(y\mathbb{1}_d^{\mathrm{T}}) \odot X \\ (1/\sqrt{2})\,B \end{bmatrix}, \quad \text{and} \quad z = \begin{bmatrix} \sqrt{\lambda}\,\mathbb{1}_n \\ 0_d \end{bmatrix}.$$

– Then, under a symmetry assumption, SPS can be applied.



# Confidence Sets for Kernel Ridge Regression

– The kernelized version of RR, Kernel Ridge Regression (KRR) is

$$g(f) \doteq \frac{1}{2}\sum_{i=1}^{n}(f(x_i)-y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2$$

where $f$ may come from an infinite dimensional RKHS.

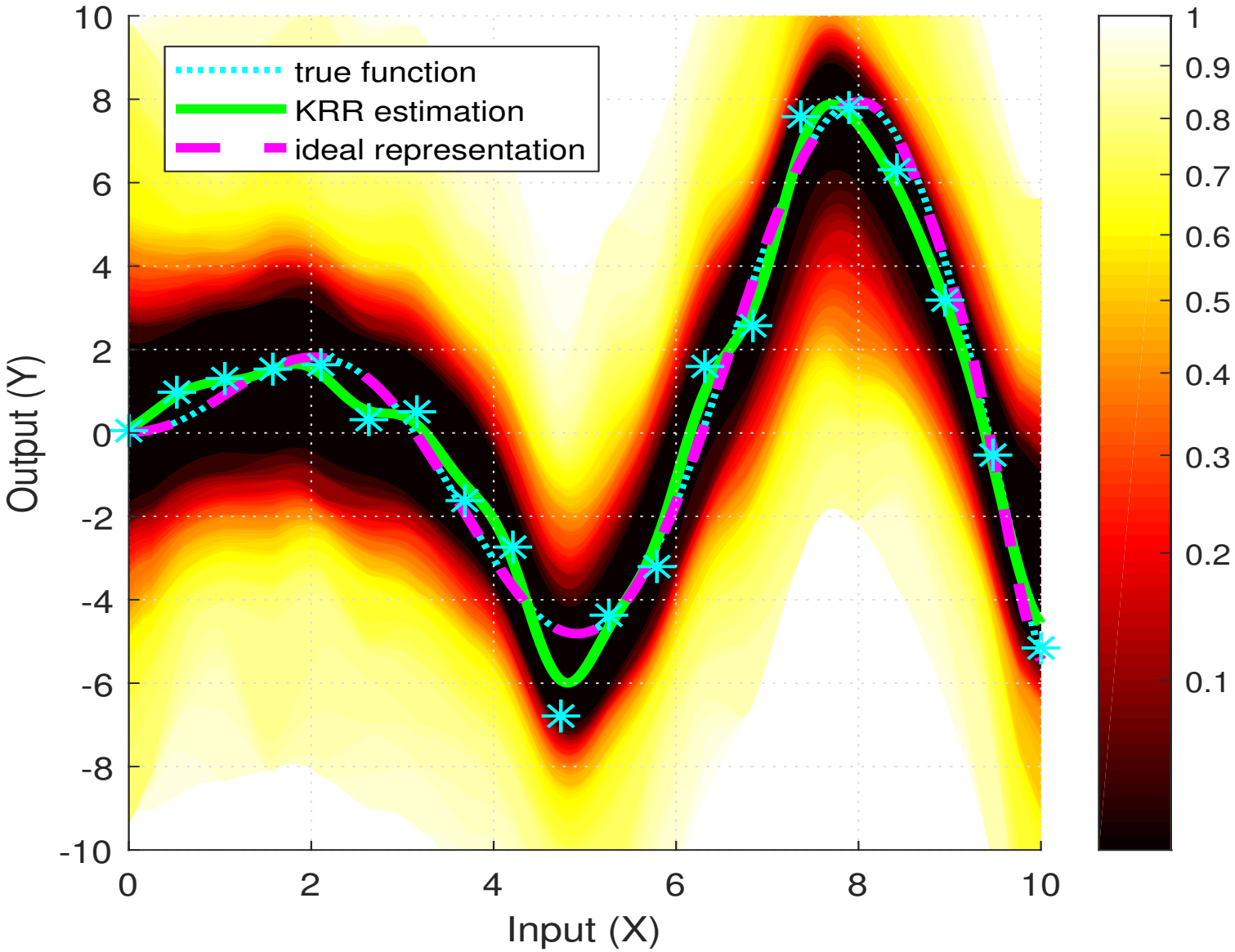
– Using the representer theorem and the reproducing property,

$$g(\alpha) = \frac{1}{2}\|y - K\alpha\|^2 + \lambda \alpha^{\mathrm{T}} K \alpha$$

**SPS Evaluation Function for Kernel Ridge Regression**

$$Z_i(\alpha) \doteq \left\| (K^2 + 2\lambda K^{1/2})^{-1/2} \left[ K G_i (y - K\alpha) + 2\lambda K^{1/2}\alpha \right] \right\|^2$$

# Experiment: SPS for Kernel Ridge Regression



# Confidence Sets for Support Vector Regression

– Criterion of Support Vector Regression, for $c > 0$ and $\bar{\varepsilon} > 0$, is

$$g(f) \doteq \frac{1}{2}\| f \|_{\mathcal{H}}^2 + \frac{c}{n}\sum_{k=1}^{n} \max\{ 0, |f(x_k) - y_k| - \bar{\varepsilon} \}$$

– Using the representer theorem, Lagrangian duality and the Karush–Kuhn–Tucker (KKT) conditions, we arrive at the dual

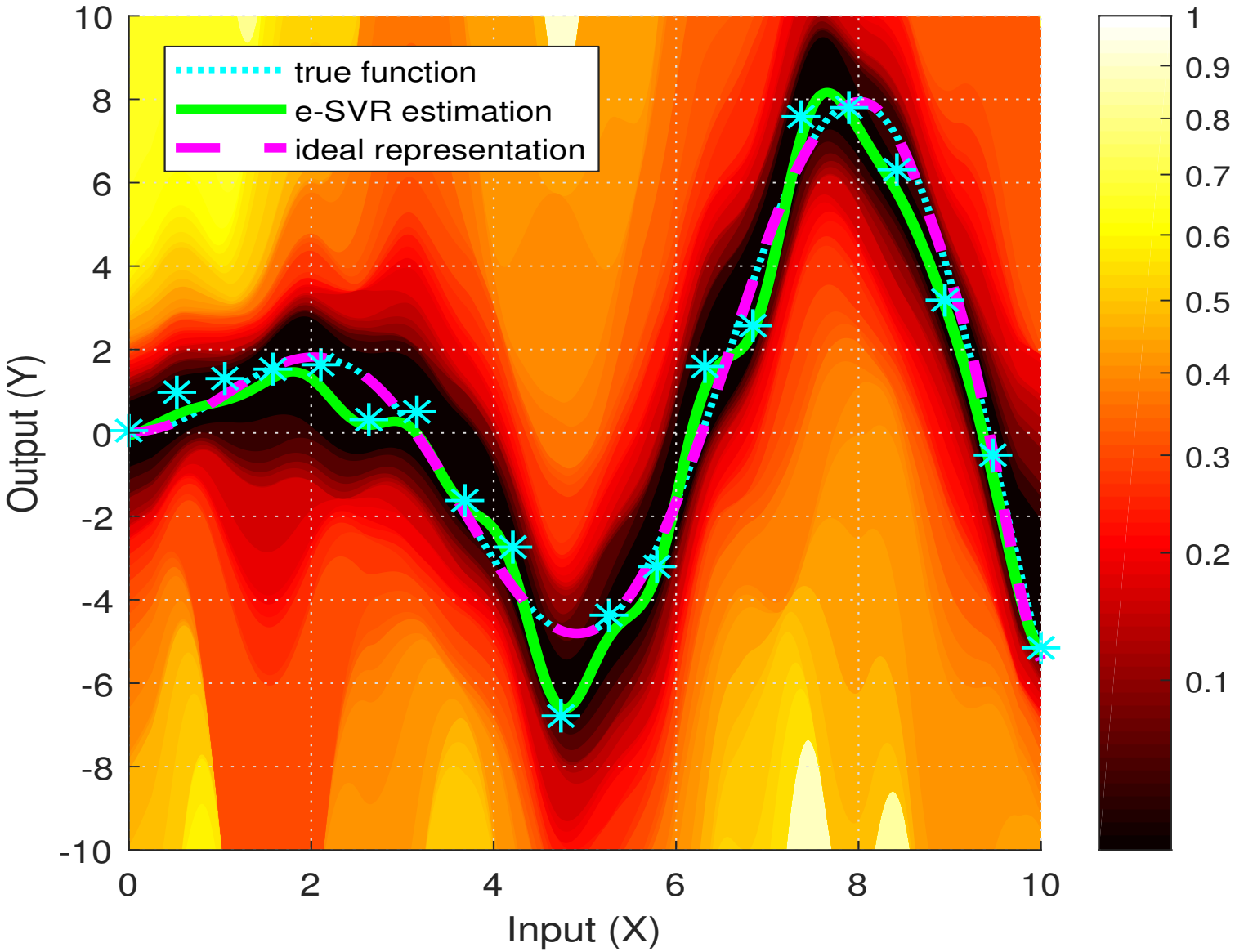
$$g^*(\alpha, \beta) = y^{\mathrm{T}}(\alpha - \beta) - \frac{1}{2}(\alpha - \beta)^{\mathrm{T}} K (\alpha - \beta) - \bar{\varepsilon}(\alpha + \beta)^{\mathrm{T}} \mathbb{1}$$

subject to $\alpha, \beta \in [0, c/n]^n$ and $(\alpha - \beta)^{\mathrm{T}} \mathbb{1} = 0$.

## Evaluation Function for Support Vector Regression

$$Z_i(\alpha) \doteq \left\| G_i (y - K\alpha) - \bar{\varepsilon}\, \mathrm{sign}(\alpha) \right\|^2$$

# Experiment: Confidence Regions for SVR



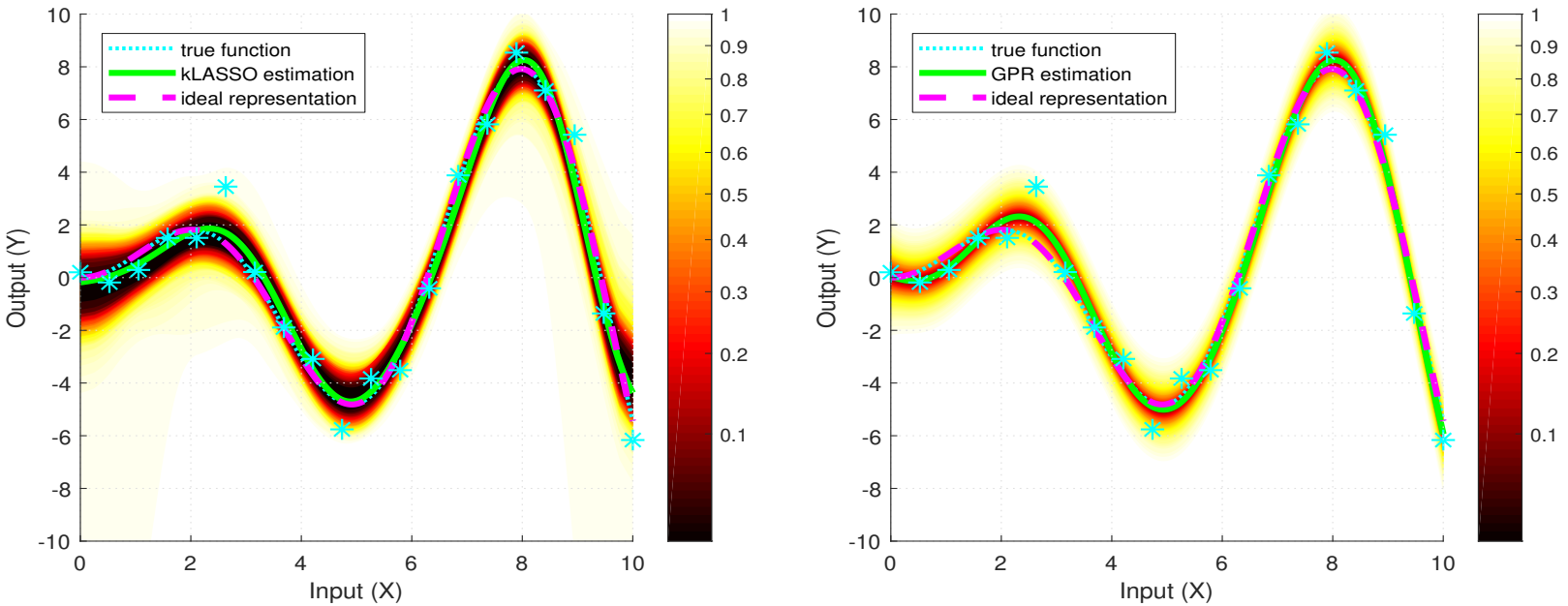
# Confidence Sets for Kernelized LASSO

– The kernelized version of LASSO leads to the objective,

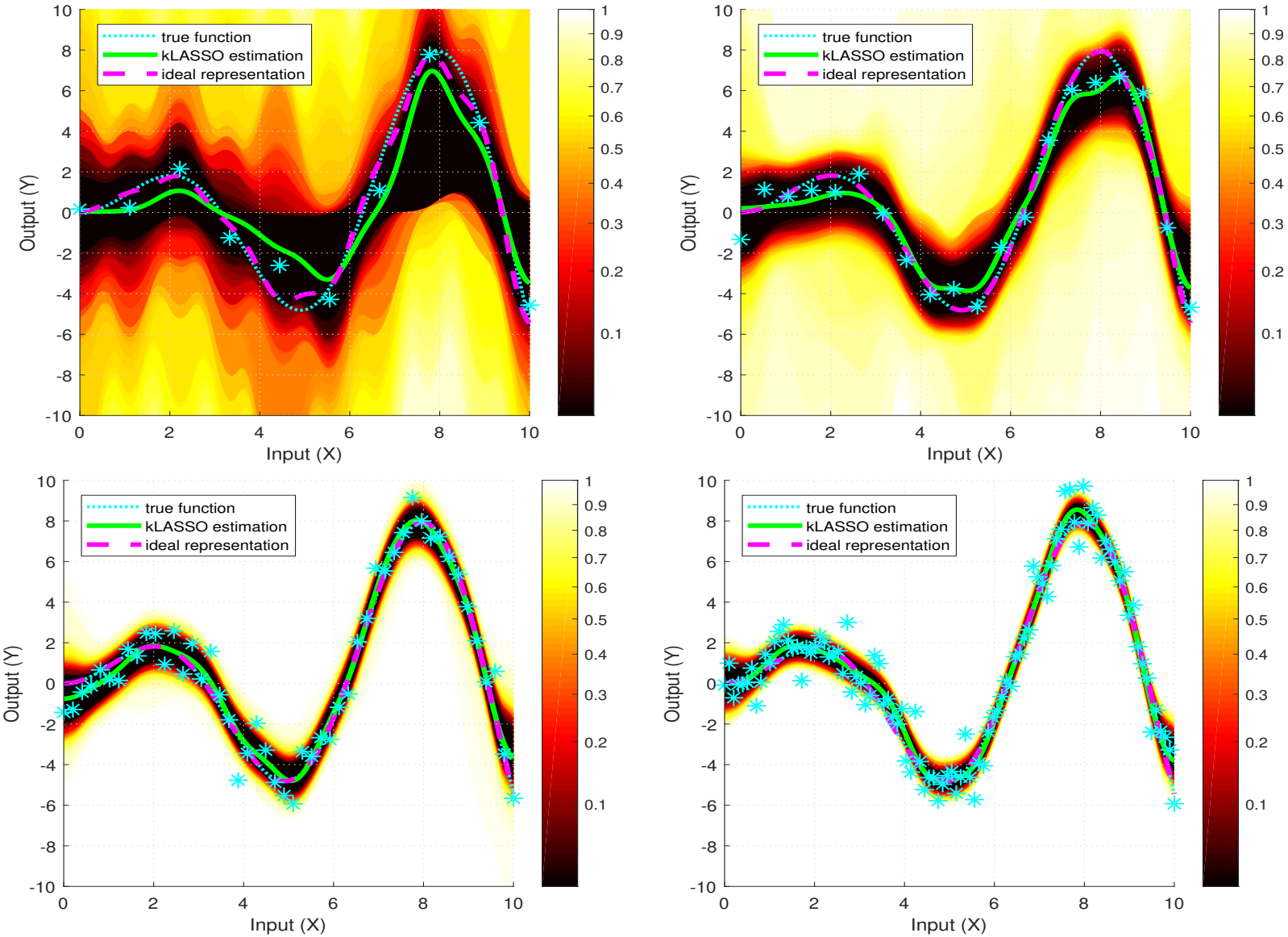
$$g(f) \doteq 1/2 \, \| \, y - K\alpha \, \|^2 + \lambda \, \| \, \alpha \, \|_1.$$

## Evaluation Function for Kernelized LASSO

$$Z_i(\alpha) \doteq \| \, K \, G_i \, (K\alpha - y) + \lambda \, \mathrm{sign}(\alpha) \, \|^2$$



# Experiment: Consistency (n = 10, 20, 50, and 100)



# Summary: Guarantees for Ideal Representations

- A data-driven uncertainty quantification (UQ) approach was preseted for (regression) models constructed by kernel methods.
- UQ takes the form of confidence regions for ideal representations of the true function which we only observe via measurement noise.
- The core idea is to perturb the residuals in the gradient of the objective function with some distributionally invariant operations.
- The resulting sets have exact (user-chosen) coverage probabilities.
- The framework is distribution-free (unlike GP regression), only mild regularities are assumed about the noise (like symmetry).
- The method has non-asymptotic (finite sample) guarantees.
- Convex quadratic problems and symmetric noises $\Rightarrow$ the regions are star convex and have ellipsoidal outer approximations.
- The ideas were demonstrated on LS-SVM, KRR, SVR & kLASSO.

# III. Nonparametric Confidence Bands

## Distribution-Free and Non-Asymptotically Guaranteed Regions for the True Function

Joint work with: Bálint Horváth

# Nonparametric Confidence Bands

- 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 possible input $x \in \mathcal{D}$.
- More precisely, we would like to construct a function $I$ with

$$\nu(I) \doteq \mathbb{P}\big(\forall x \in \mathcal{D} : I_1(x) \leq f_*(x) \leq I_2(x)\big) \geq 1 - \alpha$$

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

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

- Based on this, the reliability of $I$ is $\nu(I) = \mathbb{P}(\operatorname{graph}_{\mathcal{D}}(f_*) \subseteq \mathcal{I})$, where we define $\operatorname{graph}_{\mathcal{D}}(f_*) \doteq \{(x, f_*(x)) : x \in \mathcal{D}\}$.

# Paley-Wiener Spaces

- Let $\mathcal{H}$ be the space of $f \in \mathcal{L}^2(\mathbb{R})$ functions, such that the support of the Fourier transform of $f$ is included in $[-\eta, \eta]$, where $\eta > 0$.
- This space of band-limited functions, called the Paley-Wiener space, is an RKHS. Its reproducing kernel is defined as

$$k(z, s) \doteq \frac{\sin(\eta(z-s))}{\pi(z-s)},$$

for $z \neq s$, where $z, s \in \mathbb{R}$; and $k(z, z) \doteq \eta / \pi$.

- It is a (closed) subspace of $\mathcal{L}^2$ and $k$ induces the inner product

$$\langle f, g \rangle_{\mathcal{H}} \doteq \int_{\mathbb{R}} f(x)\, g(x)\, \mathrm{d}x.$$

- Thus, the kernel norm of this space is: $\|f\|_{\mathcal{H}} = \|f\|_2$, for $f \in \mathcal{H}$.
- Henceforth, we work with the above defined Paley-Wiener kernel.

# Main Assumptions

(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 $\mathbb{E}[y_k^2] < \infty$, all for $k \in [n]$.

(A1) Each (measurement) noise, $\varepsilon_k \doteq y_k - f_*(x_k)$, 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 $[0, 1]$ :

$$\int_{\mathbb{R}} f_*^2(x)\, \mathbb{I}(x \notin [0, 1])\, \mathrm{d}x \;\leq\; \delta_0,$$

where $\mathbb{I}(\cdot)$ is an indicator and $\delta_0 > 0$ is a universal constant.

# Bounding the Norm: Noise-Free Case

## Lemma 1: Upper Bound for the Norm (Noiseless Outputs)

Assume that A0, A2, A3 hold and that $y_k = f_*(x_k)$, for all $k \in [n]$. Then, for any risk probability $\alpha \in (0,1)$, we can guarantee that

$$\mathbb{P}\big( \|f_*\|_{\mathcal{H}}^2 \leq \kappa \big) \geq 1 - \alpha,$$

with the following choice of the upper bound $\kappa$:

$$\kappa \doteq \frac{1}{n} \sum_{k=1}^{n} y_k^2 + \sqrt{\frac{\ln(\alpha)}{-2n}} + \delta_0.$$

# Interval Endpoints: Noise-Free Case

- We can compute the minimum norm needed to interpolate the original dataset extended by $(x_0, y_0)$ for any candidate pair.
- First, we extend the Gram matrix with query point $x_0$,

$$K_0(i+1, j+1) \doteq k(x_i, x_j),$$

for $i, j = 0, 1, \ldots, n$ (the extended $K_0$ is a.s. invertible).

- The minimum norm interpolation of $(x_0, y_0), \ldots, (x_n, y_n)$ 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 $\tilde{y} \doteq (y_0, y_1, \ldots, y_n)^{\mathrm{T}}$ and $\tilde{\alpha} \doteq (\tilde{\alpha}_0, \ldots, \tilde{\alpha}_n)^{\mathrm{T}}$. The norm square of (interpolant) $\tilde{f}$ is

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

# Guaranteed Coverage: Noise-Free Case

– These lead to the following two (convex) optimization problems:

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

– These are very special problems that can be solved analytically.

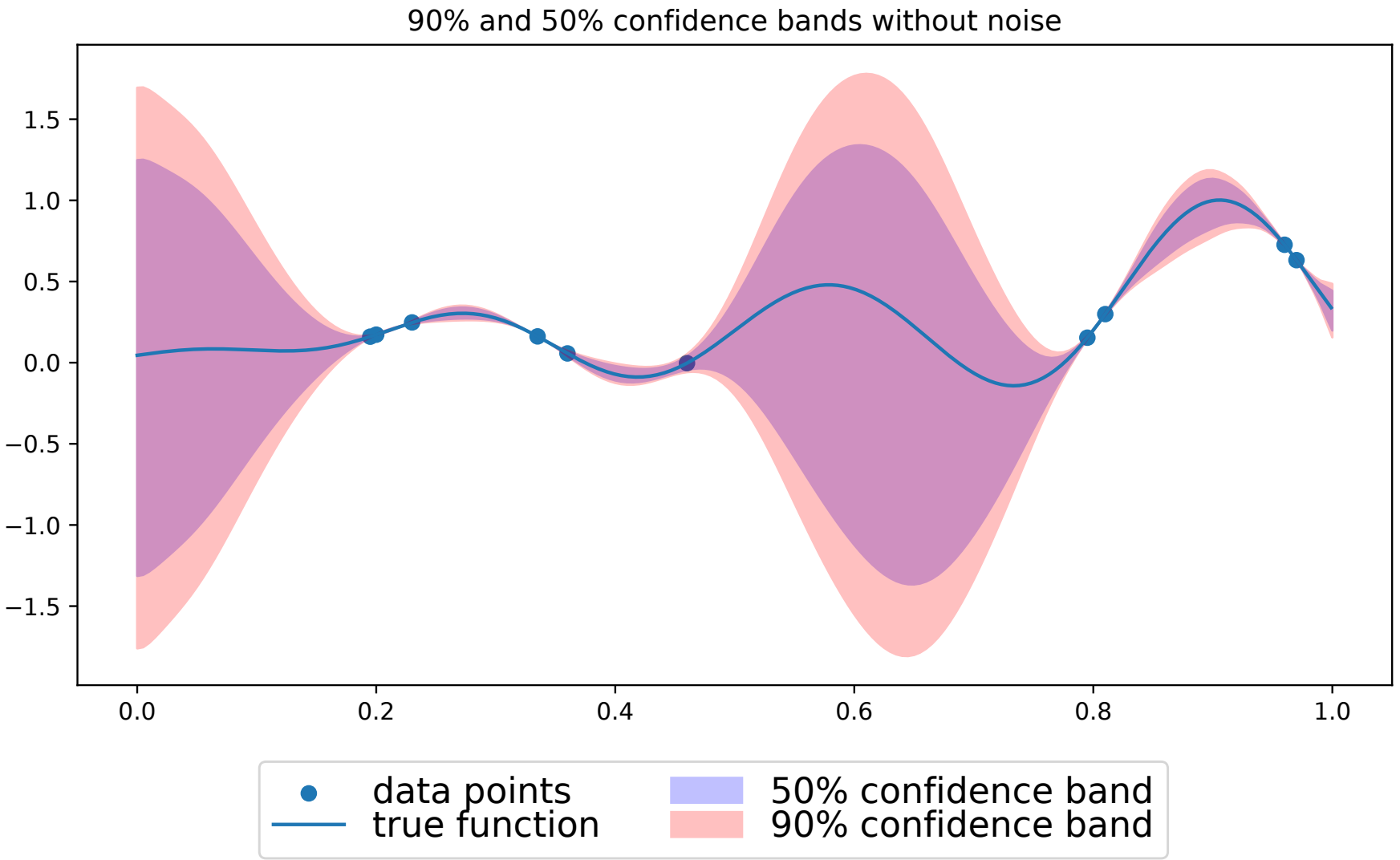
– The optimal values, $y_{\min}$ and $y_{\max}$, determine the endpoints of the confidence interval at $x_0$: $I_1(x_0) \doteq y_{\min}$ and $I_2(x_0) \doteq y_{\max}$.

**Theorem 1: Guaranteed Coverage (Noiseless Outputs)**

Assume that A0, A2, A3 and $y_k = f_*(x_k)$, for $k \in [n]$, are satisfied. Let $\alpha \in (0,1)$ be a risk probability. The construction guarantees

$$\mathbb{P}\big( \operatorname{graph}_{\mathcal{D}}(f_*) \subseteq \mathcal{I} \big) \geq 1 - \alpha.$$

# Nonparametric Conf. Bands: Noise-Free Setting



# Bounding the Norm: Noisy Case

– With gradient-perturbation methods, we can built simultaneous confidence intervals for the first $d \leq n$ observed inputs; that is

$$\mathbb{P}\big(\forall k \in [d] : f_*(x_k) \in [\nu_k, \mu_k]\big) \geq 1 - \beta,$$

for $k \in [d]$, where $\beta \in (0, 1)$ is a (user-chosen) risk probability.

## Lemma 2: Upper Bound for the Norm (Noisy Outputs)

Assume that A0, A1, A2, and A3 hold and that we have built simultaneous confidence intervals, as above. Then,

$$\mathbb{P}\big(\|f_*\|_{\mathcal{H}}^2 \leq \tau\big) \geq 1 - \alpha - \beta,$$

with the following choice of the upper bound $\tau$:

$$\tau \doteq \frac{1}{d}\sum_{k=1}^{d} \max\{\nu_k^2, \mu_k^2\} + \sqrt{\frac{\ln(\alpha)}{-2d}} + \delta_0.$$

# Guaranteed Coverage: Noisy Case

– These lead to the following two (convex) optimization problems:

$$\begin{aligned} \min / \max \quad & z_0 \\ \text{subject to} \quad & (z_0, \ldots, z_d)\widetilde{K}_0^{-1}(z_0, \ldots, z_d)^{\mathrm{T}} \leq \tau \\ & \nu_1 \leq z_1 \leq \mu_1, \ \ldots, \ \nu_d \leq z_d \leq \mu_d \end{aligned}$$

– Given input $x_0$, $\widetilde{K}_0(i+1, j+1) \doteq k(x_i, x_j)$, for $i, j = 0, 1, \ldots, d$

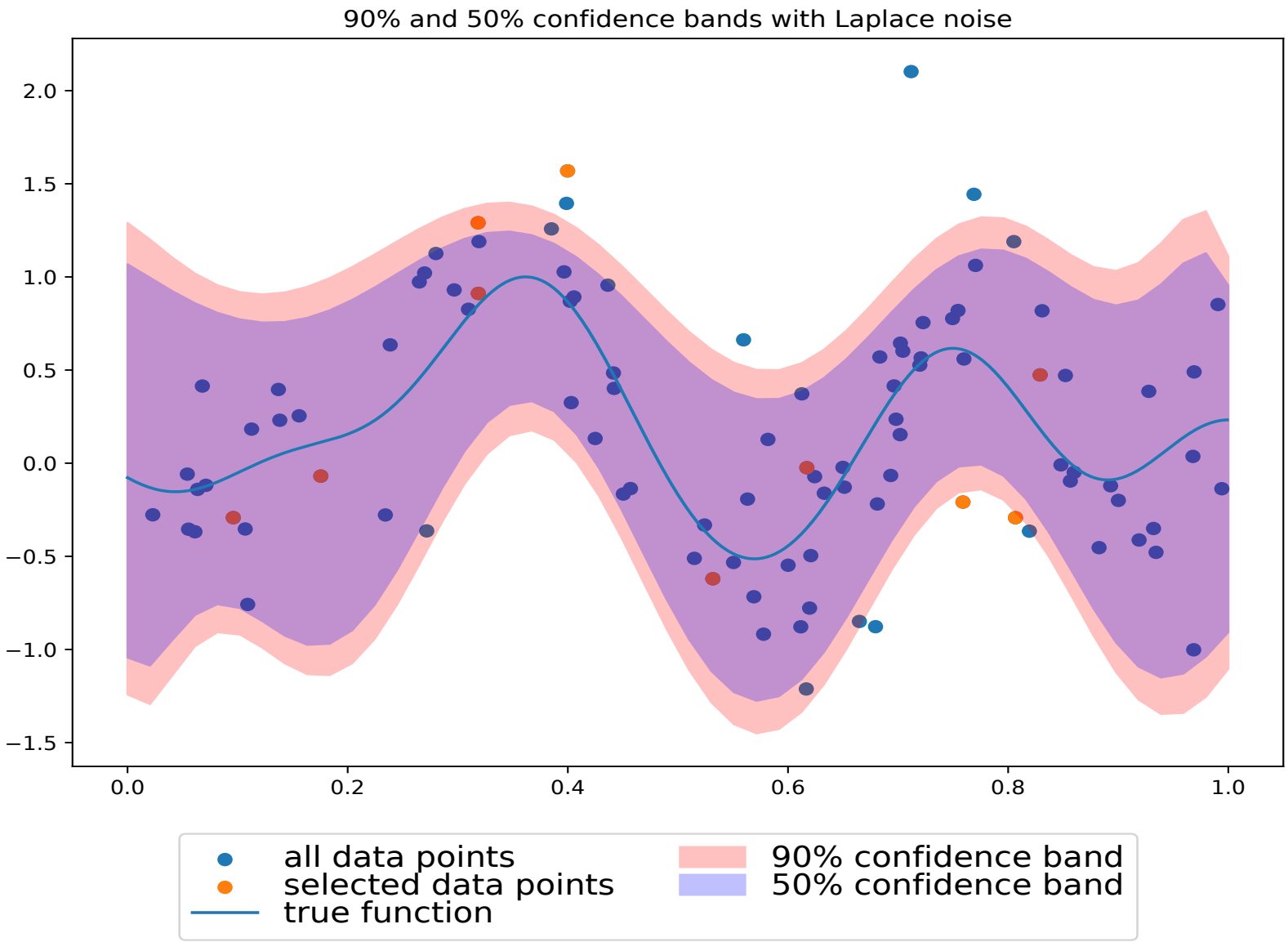
– The optimal values, $y_{\min}$ and $y_{\max}$, determine the endpoints of the confidence interval at $x_0$: $I_1(x_0) \doteq y_{\min}$ and $I_2(x_0) \doteq y_{\max}$.

**Theorem 2: Guaranteed Coverage (Noisy Outputs)**

Assume that A0, A1, A2, and A3 are satisfied. Let $\alpha, \beta \in (0, 1)$ be given risk probabilities. Then, the construction guarantees

$$\mathbb{P}\big(\operatorname{graph}_{\mathcal{D}}(f_*) \subseteq \mathcal{I}\big) \geq 1 - \alpha - \beta.$$

# Nonparam. Conf. Bands with Measurement Noise



# Summary: Nonparametric Confidence Bands

- A nonparametric and distribution-free method was introduced to build confidence bands for bounded, band-limited functions.
- The confidence band is simultaneously guaranteed for all inputs.
- The construction was first presented for the noiseless case.
- The main idea is to first calculate a (stochastically guaranteed) upper bound for the kernel norm (which measures smoothness).
- Then, each candidate $(x_0, y_0)$ can be tested whether there is a function from the Paley-Wiener space that interpolates the original dataset extended with $(x_0, y_0)$ having a norm below the bound.
- Later, the method was extended allowing symmetric noises.
- Besides having non-asymptotic guarantees, the approach was also demonstrated numerically, supporting its feasibility.

# Thank you for your attention!

www.sztaki.hu/~csaji ✉ csaji@sztaki.hu