Uncertainty Quantification and Kernels

Distribution-Free Inference for Regression and Classification

Balázs Csanád Csáji

SZTAKI: Institute for Computer Science and Control, Budapest, Hungary

Joint work with M.C. Campi, E. Weyer, K.B. Kis, and A. Tamás

Deep Learning Seminar, AI National Laboratory, June 23, 2021

I. Linear Regression

Distribution-Free Finite Sample Exact Confidence Regions

Joint work with: Marco Campi and Erik Weyer

Linear Regression

Consider a standard linear regression problem:

Linear Regression

$$y_t \doteq \varphi_t^{\mathrm{T}}\theta^* + w_t = \langle \phi(x_t), \theta^* \rangle + w_t$$

for $t = 1, \ldots, n$, where

- y_t — output (scalar response variable, measured)
- φ_t — regressor (deterministic, d dimensional, measured)
- w_t — noise terms (zero mean, uncorrelated) with variance σ^2
- θ^* — true parameter (deterministic, d dimensional, unknown)
- n — sample size (the number of measured input-output pairs)
- $\Phi_n \doteq [\varphi_1, \ldots, \varphi_n]^{\mathrm{T}}$ — regression matrix (skinny and full rank)

Ordinary Least Squares

- Given: a sample, $\mathcal{Z}$, of size n of outputs $\{y_t\}$ and regressors $\{\varphi_t\}$.
- A classical approach is the least squares criterion (loss), that is

$$\mathcal{V}(\theta \mid \mathcal{Z}) \doteq \frac{1}{2} \sum_{t=1}^{n} (y_t - \varphi_t^{\mathrm{T}} \theta)^2.$$

- The least squares estimate (LSE) can be found by solving

Normal Equation

$$\nabla_\theta \mathcal{V}(\hat{\theta}_n \mid \mathcal{Z}) = \sum_{t=1}^{n} \varphi_t (y_t - \varphi_t^{\mathrm{T}} \hat{\theta}_n) = 0$$

Asymptotic Normality

– The LSE is (under the "skinny and full rank" assumption)

$$\hat{\theta}_n = \left(\sum_{t=1}^{n} \varphi_t \varphi_t^{\mathrm{T}} \right)^{-1} \left(\sum_{t=1}^{n} \varphi_t y_t \right).$$

– The (scaled) error of LSE is asymptotically Gaussian:

Limiting Distribution

$$\sqrt{n}\,(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, \sigma^2\, R^{-1}) \quad \text{as} \quad n \to \infty$$

where R is the limit of $R_n = \frac{1}{n} \sum_{t=1}^{n} \varphi_t \varphi_t^{\mathrm{T}}$ as $n \to \infty$ (if it exists).

Asymptotic Confidence Ellipsoids

– The standard confidence region construction is

Asymptotic Confidence Ellipsoid

$$\widetilde{\Theta}_{n,\mu} \doteq \left\{ \theta \in \mathbb{R}^d \,:\, (\theta - \hat{\theta}_n)^{\mathrm{T}} R_n (\theta - \hat{\theta}_n) \leq \frac{\mu\, \hat{\sigma}_n^2}{n} \right\}$$

then $\mathbb{P}(\theta^* \in \widetilde{\Theta}_{n,\mu}) \approx F_{\chi^2(d)}(\mu)$, where $F_{\chi^2(d)}$ is the CDF of $\chi^2(d)$,

$$\hat{\sigma}_n^2 \doteq \frac{1}{n-d} \sum_{t=1}^{n} (y_t - \varphi_t^{\mathrm{T}} \hat{\theta}_n)^2,$$

is an estimate of variance σ^2 based on the sample.

– This construction is however only a heuristic for finite samples.

Reference and Sign-Perturbed Sums

Let us introduce a reference sum and $m-1$ sign-perturbed sums.

Reference Sum

$$S_0(\theta) \doteq R_n^{-\frac{1}{2}} \sum_{t=1}^{n} \varphi_t (y_t - \varphi_t^{\mathrm{T}} \theta)$$

Sign-Perturbed Sums

$$S_i(\theta) \doteq R_n^{-\frac{1}{2}} \sum_{t=1}^{n} \varphi_t \, \alpha_{i,t} (y_t - \varphi_t^{\mathrm{T}} \theta)$$

for $i = 1, \ldots, m-1$, where $\alpha_{i,t}$ $(t = 1, \ldots, n)$ are i.i.d. random signs, that is $\alpha_{i,t} = \pm 1$ with probability $1/2$ each (Rademacher).

Intuitive Idea: Distributional Invariance

- Assume $\{w_t\}$ are independent, each w_t is symmetric about zero.
- Observe that, if $\theta = \theta^*$, we have (for $i = 1, \ldots, m-1$)

Distributional Invariance

$$S_0(\theta^*) = R_n^{-\frac{1}{2}} \sum_{t=1}^{n} \varphi_t w_t$$

$$S_i(\theta^*) = R_n^{-\frac{1}{2}} \sum_{t=1}^{n} \varphi_t \alpha_{i,t} w_t$$

- Consider the ordering $\|S_{(0)}(\theta^*)\|^2 \prec \cdots \prec \|S_{(m-1)}(\theta^*)\|^2$

  (note: "$\prec$" is the canonical "$<$" with random tie-breaking)

  All orderings are equally probable! (they are conditionally i.i.d.)

Intuitive Idea: Reference Dominance

- What if $\theta \neq \theta^*$? How well can we distinguish "false" parameters?
- In fact, the reference paraboloid $\|S_0(\theta)\|^2$ increases faster than $\{\|S_i(\theta)\|^2\}$, therefore will eventually dominate the ordering.
- Intuitively, for "large enough" $\|\tilde{\theta}\|$, where $\tilde{\theta} \doteq \theta^* - \theta$

Eventual Dominance of the Reference Paraboloid

$$\left\| \sum_{t=1}^{n} \varphi_t \varphi_t^{\mathrm{T}} \tilde{\theta} + \sum_{t=1}^{n} \varphi_t w_t \right\|_{R_n^{-1}}^2 > \left\| \sum_{t=1}^{n} \pm \varphi_t \varphi_t^{\mathrm{T}} \tilde{\theta} + \sum_{t=1}^{n} \pm \varphi_t w_t \right\|_{R_n^{-1}}^2$$

with "high probability" (for simplicity $\pm$ is used instead of $\{\alpha_{i,t}\}$).

Non-Asymptotic Confidence Regions

The rank of $\|S_0(\theta)\|^2$ in the ordering of $\{\|S_i(\theta)\|^2\}$ w.r.t. $\prec$ is

$$\mathcal{R}(\theta) \doteq 1 + \sum_{i=1}^{m-1} \mathbb{I}\big(\|S_i(\theta)\|^2 \prec \|S_0(\theta)\|^2\big),$$

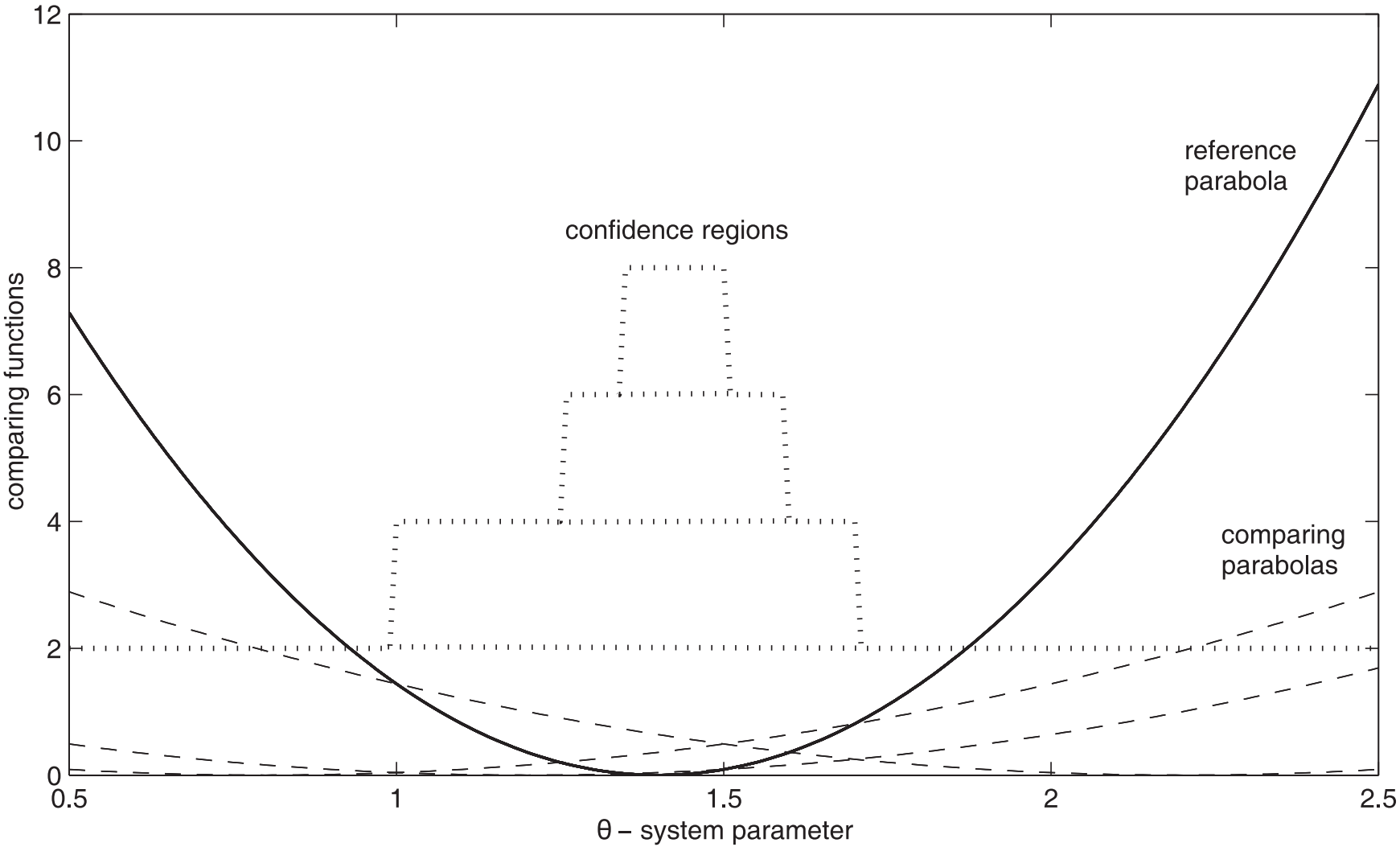
where $\mathbb{I}(\cdot)$ is an indicator function.

Sign-Perturbed Sums (SPS) Confidence Regions

$$\widehat{\Theta}_n \doteq \Big\{ \theta \in \mathbb{R}^d \,:\, \mathcal{R}(\,\theta\,) \leq m - q \Big\}$$

where $m > q > 0$ are user-chosen integers (design parameters).

Simple Illustration ($y_t = \theta^* + w_t,\ n = 3,\ m = 4$)



Exact Coverage Probability

(A1) $\{w_t\}$ is a sequence of independent random variables.
Each w_t has a symmetric probability distribution about zero.

(A2) The outer product of regressors is invertible, $\det(R_n) \neq 0$.

Exact Confidence of SPS Confidence Regions

$$\mathbb{P}\left(\theta^* \in \widehat{\Theta}_n\right) = 1 - \frac{q}{m}$$

for any finite sample. Parameters m and q are under our control.

Note that $\|S_0(\hat{\theta}_n)\|^2 = 0$, thus $\hat{\theta}_n \in \widehat{\Theta}_n$, assuming it is non-empty (that is, the confidence region is built around the LS estimate).

Star Convexity

Set $\mathcal{X} \subseteq \mathbb{R}^d$ is star convex if there is a star center $c \in \mathbb{R}^d$ with

$$\forall x \in \mathcal{X}, \forall \beta \in [0,1] \, : \, \beta x + (1-\beta) c \in \mathcal{X}.$$

Star Convexity of SPS Confidence Regions

$\widehat{\Theta}_n$ is star convex with the LSE, $\hat{\theta}_n$, as a star center

Hint: $\widehat{\Theta}_n$ is the union and intersection of ellipsoids containing LSE.

Strong Consistency

(A1) independence, symmetricity: $\{w_t\}$ are independent, symmetric

(A2) invertibility: $R_n \doteq \frac{1}{n}\sum_{t=1}^{n} \varphi_t \varphi_t^{\mathrm{T}}$ is invertible

(A3) regressor growth rate: $\sum_{t=1}^{\infty} \|\varphi_t\|^4 / t^2 < \infty$

(A4) noise moment growth rate: $\sum_{t=1}^{\infty} \left(\mathbb{E}[w_t^2]\right)^2 / t^2 < \infty$

(A5) Cesàro summability: $\lim_{n\to\infty} R_n = R$, which is positive definite

Strong Consistency of SPS Confidence Regions

$$\mathbb{P}\left(\bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} \left\{ \widehat{\Theta}_n \subseteq B_\varepsilon(\theta^*) \right\} \right) = 1,$$

where $B_\varepsilon(\theta^*) \doteq \{ \theta \in \mathbb{R}^d : \|\theta - \theta^*\| \leq \varepsilon \}$ is a norm ball.

Ellipsoidal Outer Approximation

– The reference paraboloid can be rewritten as

$$\|S_0(\theta)\|^2 = (\theta - \hat{\theta}_n)^{\mathrm{T}} R_n (\theta - \hat{\theta}_n).$$

– From which an alternative description of the confidence region is

$$\widehat{\Theta}_n \subseteq \left\{ \theta \in \mathbb{R}^d \,:\, (\theta - \hat{\theta}_n)^{\mathrm{T}} R_n (\theta - \hat{\theta}_n) \leq r(\theta) \right\},$$

where $r(\theta)$ is the qth largest value of $\{\|S_i(\theta)\|^2\}_{i \neq 0}$.

Ellipsoidal Outer Approximation

$$\widehat{\Theta}_n \subseteq \left\{ \theta \in \mathbb{R}^d \,:\, (\theta - \hat{\theta}_n)^{\mathrm{T}} R_n (\theta - \hat{\theta}_n) \leq r^* \right\}$$

– The question is of course how to find such an r^* efficiently.

Quadratically Constrained Quadratic Program

$\max\{\|S_i(\theta)\|^2 : \|S_0(\theta)\|^2 \leq \|S_i(\theta)\|^2\}$ can be obtained by

$$\begin{aligned} &\text{maximize} && \|z\|^2 \\ &\text{subject to} && z^{\mathrm{T}} A_i z + 2 z^{\mathrm{T}} b_i + c_i \leq 0 \end{aligned}$$

$$\begin{aligned} A_i &\doteq I - R_n^{-\frac{1}{2}} Q_i R_n^{-1} Q_i R_n^{-\frac{1}{2}\mathrm{T}} \\ b_i &\doteq R_n^{-\frac{1}{2}} Q_i R_n^{-1} (\psi_i - Q_i \hat{\theta}_n) \\ c_i &\doteq -\psi_i^{\mathrm{T}} R_n^{-1} \psi_i + 2\hat{\theta}_n^{\mathrm{T}} Q_i R_n^{-1} \psi_i - \hat{\theta}_n^{\mathrm{T}} Q_i R_n^{-1} Q_i \hat{\theta}_n \end{aligned}$$

$$Q_i \doteq \sum_{t=1}^{n} \alpha_{i,t} \varphi_t \varphi_t^{\mathrm{T}}, \qquad \psi_i \doteq \sum_{t=1}^{n} \alpha_{i,t} \varphi_t y_t$$

Semi-Definite Program

– Problem: the previous QCQP is not convex.

– Fortunately, strong duality holds and its dual can be written as:

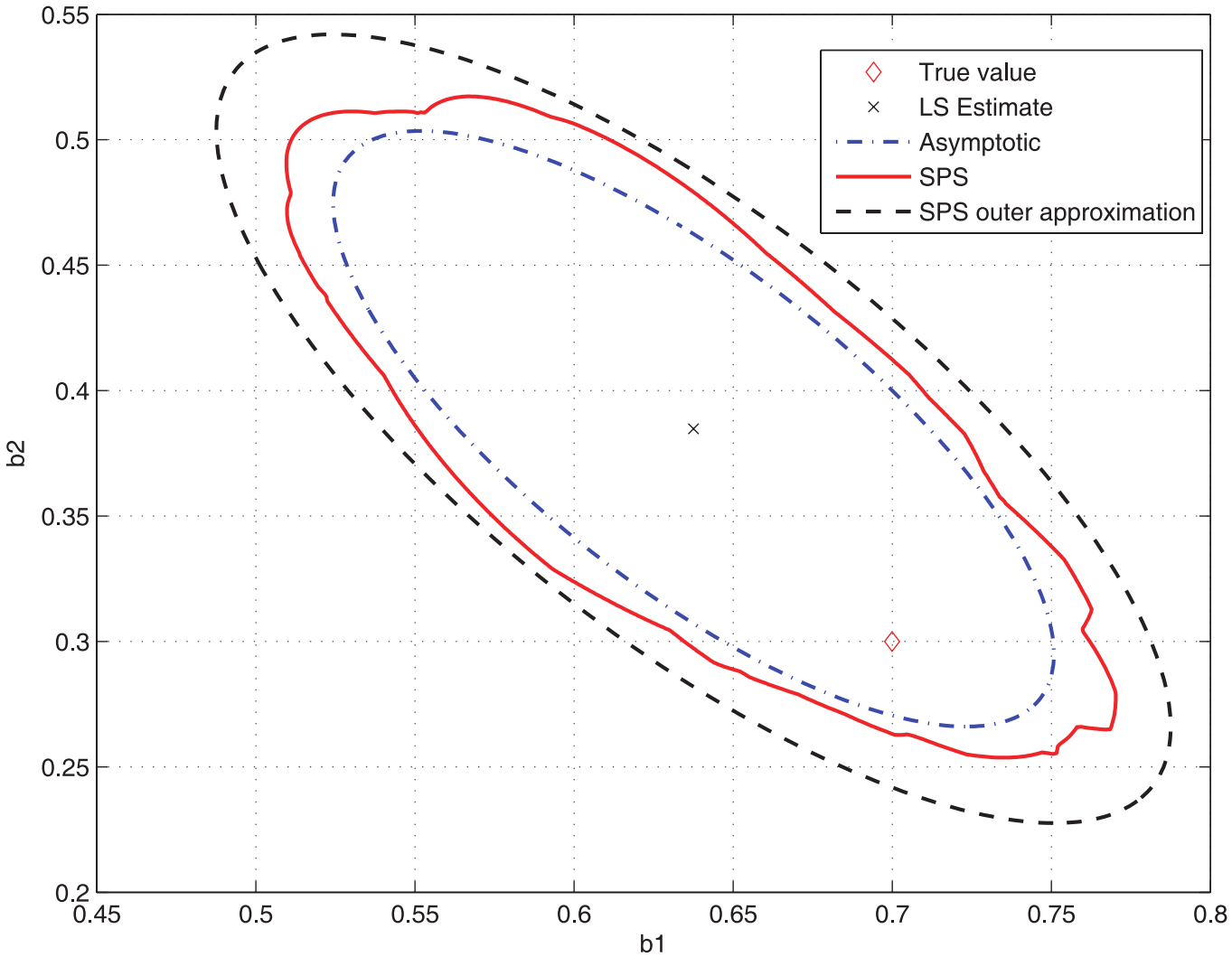
Dual Problem

$$\begin{aligned} \text{minimize} \quad & \gamma \\ \text{subject to} \quad & \lambda \geq 0 \\ & \begin{bmatrix} -I + \lambda A_i & \lambda b_i \\ \lambda b_i^{\mathrm{T}} & \lambda c_i + \gamma \end{bmatrix} \succeq 0 \end{aligned}$$

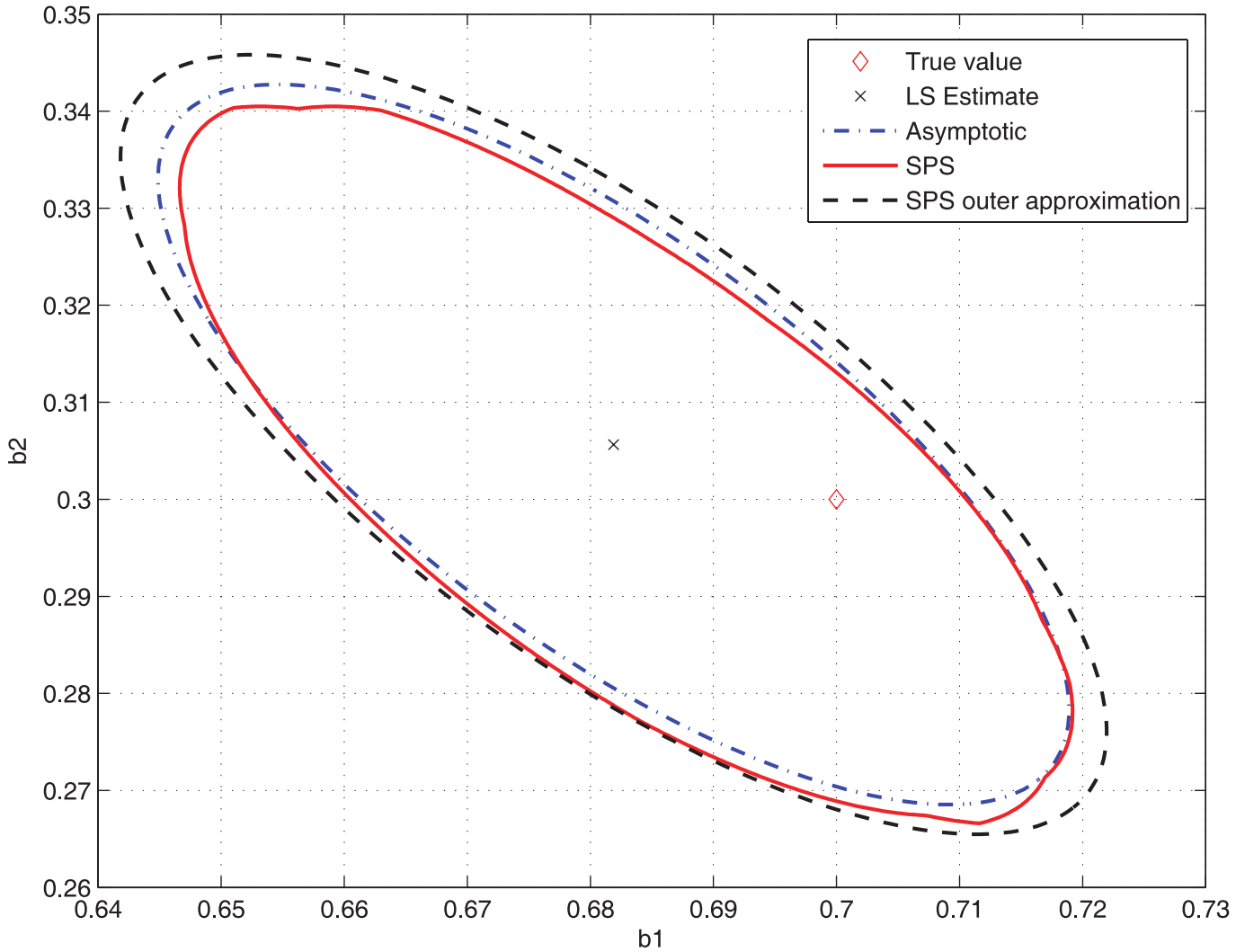
where "$\succeq 0$" denotes that a matrix is positive semidefinite.

– Radius r^* can then be found by solving $m-1$ such convex problems, obtaining $\{\gamma_i^*\}$, and defining r^* the qth largest one.

Numerical Example ($n = 25,\ m = 100,\ q = 5,\ 95\%$)



Numerical Example ($n = 400$, $m = 100$, $q = 5$, 95%)



Summary for Linear Regression

- A finite sample estimation method, called Sign-Perturbed Sums (SPS), was presented for standard linear regression problems.
- It builds confidence regions around the least squares estimate.
- Only mild statistical assumptions are needed, e.g., symmetry.
- Not needed: stationarity, moments, particular distributions, etc.
- For (rational) probabilities, exact confidence sets can be built.
- SPS is strongly consistent: the confidence regions almost surely shrink around the true parameter, as the sample size increases.
- SPS is star convex with the least squares estimate as a star center.
- It also has efficiently computable ellipsoidal outer approximations.
- It has many extensions: it can handle closed-loop LTI (dynamical) systems, GARCH processes, it can detect undermodelling, etc.

II. Kernel Regression

Distribution-Free Confidence Sets for Ideal Kernel Models

Joint work with: Krisztián Balázs Kis

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., $\forall z : \exists\, \kappa > 0$ 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 Regression

– 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 set $x \doteq (x_1, \ldots, x_n)^{\mathrm{T}} \in \mathcal{X}^n$ and $y \doteq (y_1, \ldots, y_n)^{\mathrm{T}} \in \mathbb{R}^n$.

– 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 $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

– 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 $\{x_i\}$.

Regularizated 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, Ω, 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.

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 ε 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 ε 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 ε 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. α 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 $\Rightarrow$ ε 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 α^* 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 Φ 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

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

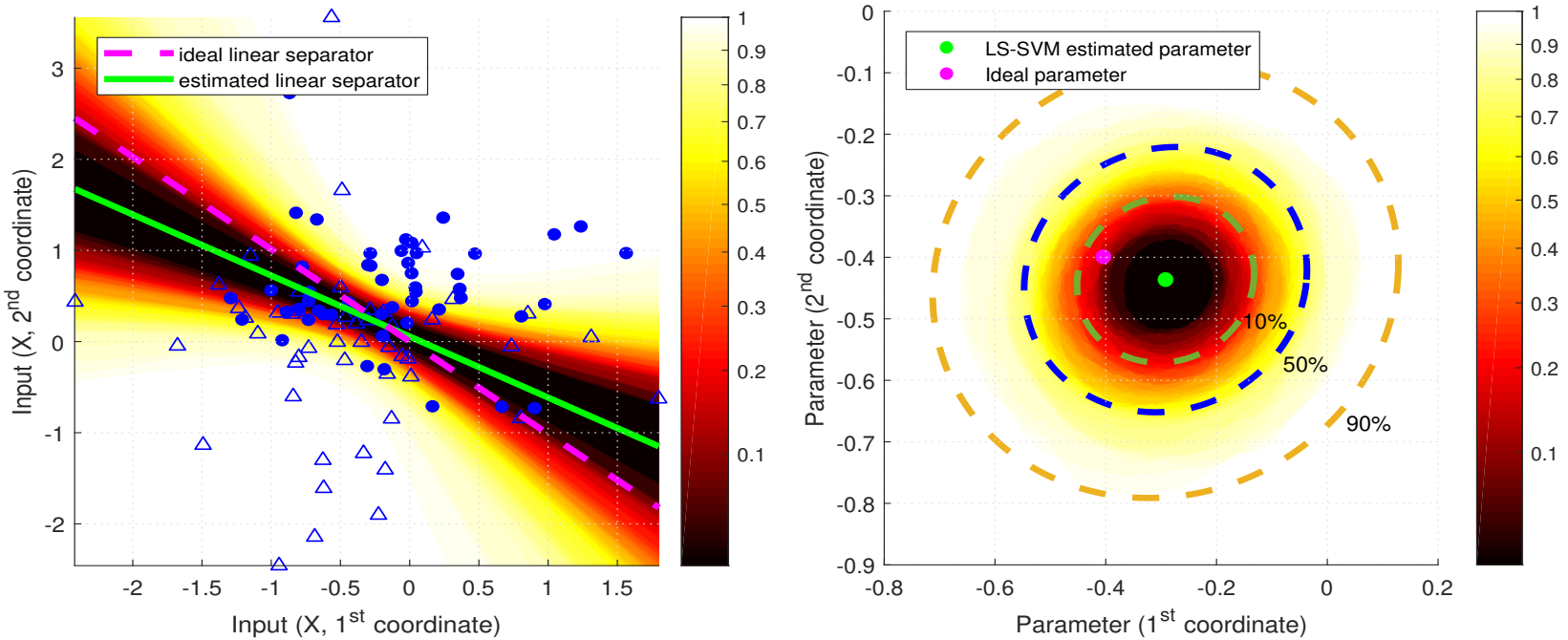
where $\mathbb{1}_n \in \mathbb{R}^n$ is the all-one vector, "$\odot$" denotes the Hadamard (entrywise) product, $X \doteq [\tilde{x}_1, \ldots, \tilde{x}_n]^{\mathrm{T}}$ with $\tilde{x}_k \doteq [1, x_k^{\mathrm{T}}]^{\mathrm{T}}$ and $B \doteq \mathrm{diag}(0, 1, \ldots, 1)$, the role of matrix B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

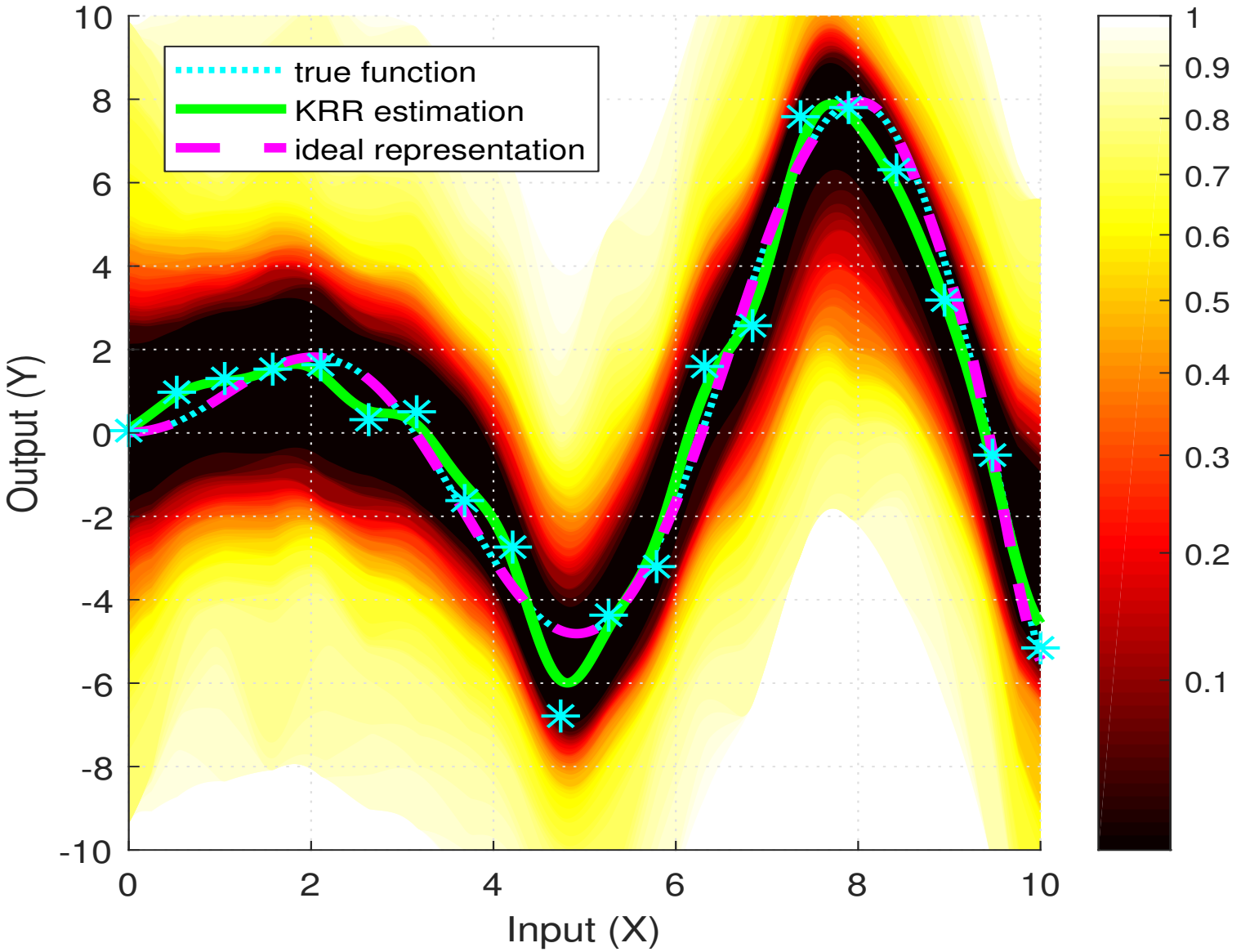
– Using the representer theorem and the reproducing property,

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

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

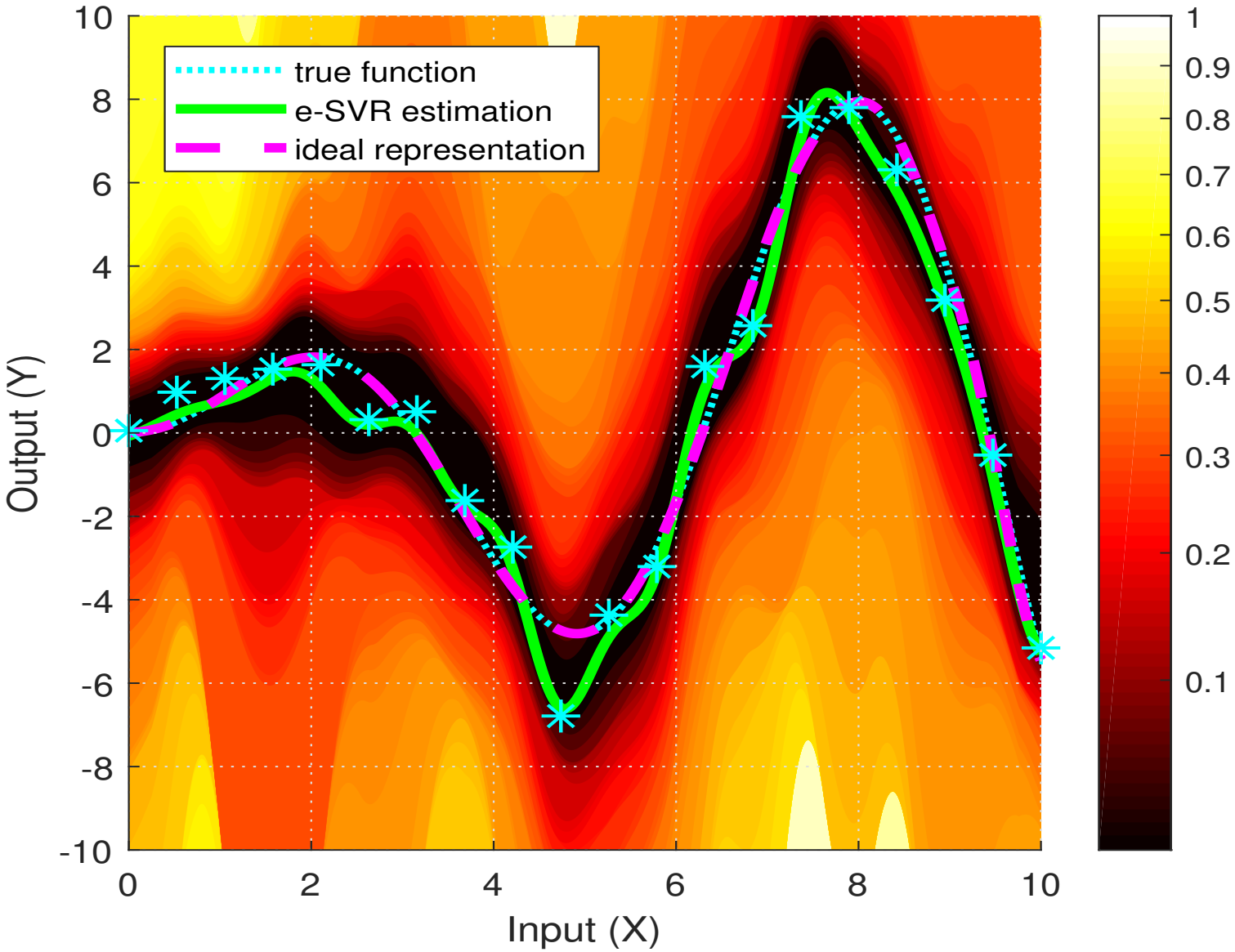
$$g^*(\alpha, \beta) = y^{\mathrm{T}}(\alpha - \beta) - \frac{1}{2}(\alpha - \beta)^{\mathrm{T}} K (\alpha - \beta) - \bar{\varepsilon}(\alpha + \beta)^{\mathrm{T}} \mathbb{1}$$

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



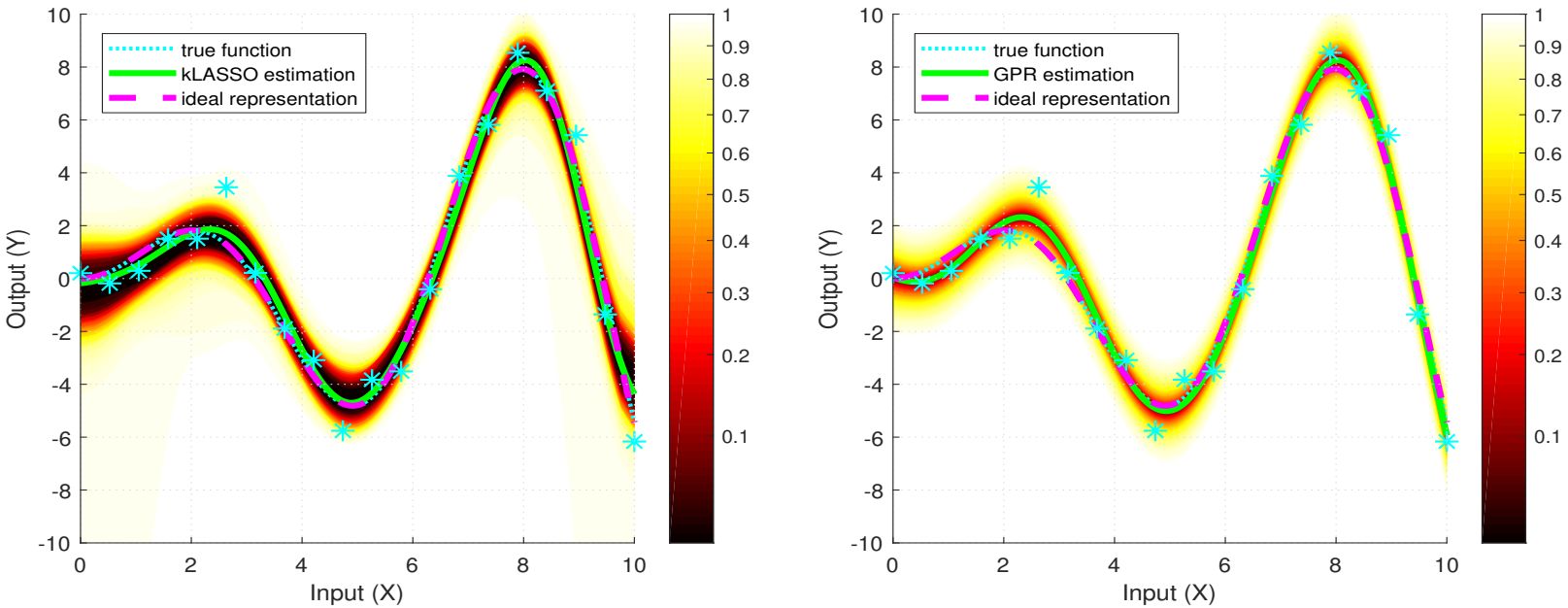
Confidence Sets for Kernelized LASSO

– The kernelized version of LASSO leads to the objective,

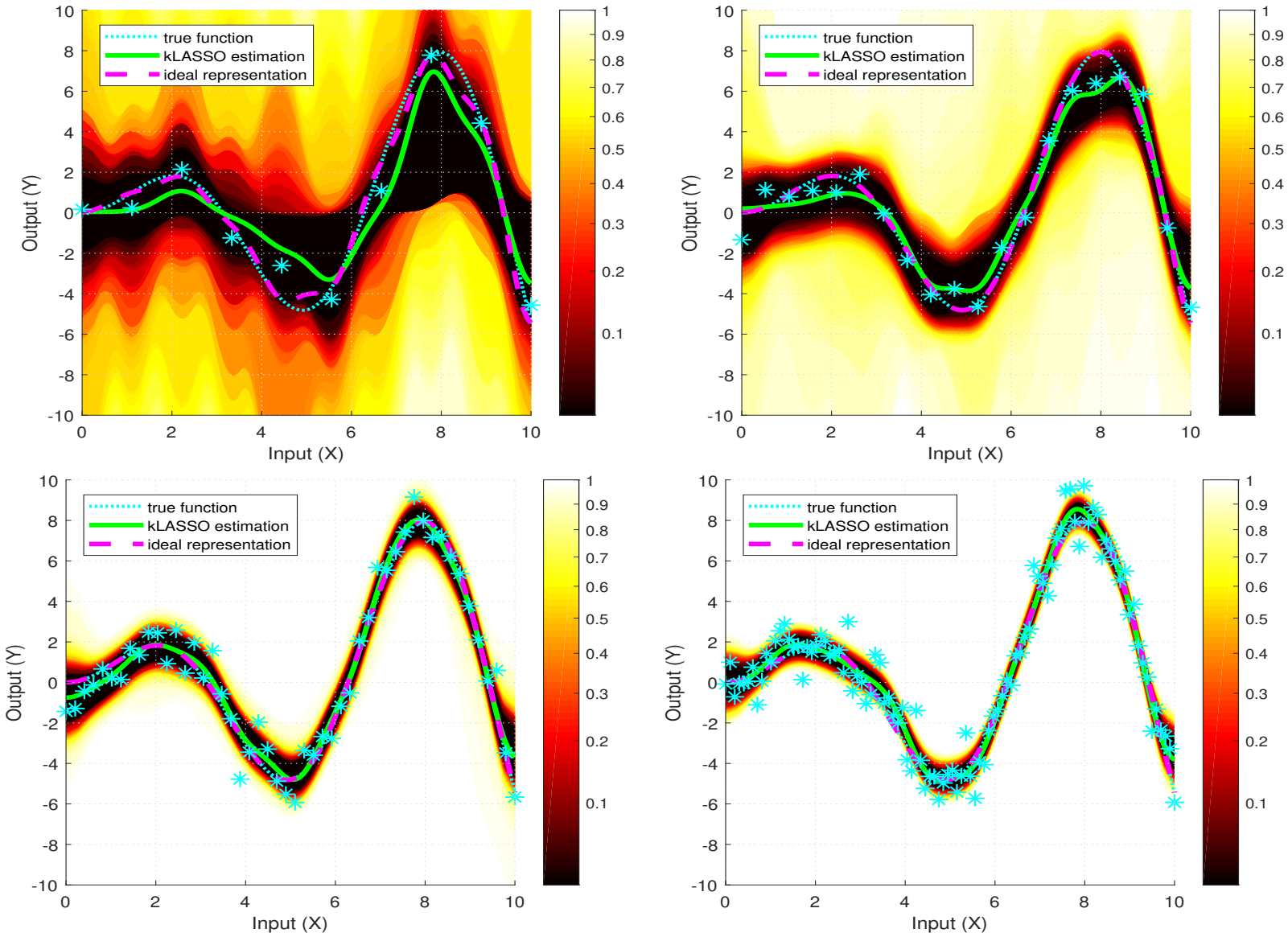
$$g(f) \doteq 1/2 \, \| \, y - K\alpha \, \|^2 + \lambda \, \| \, \alpha \, \|_1.$$

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 for Kernel Regression

- 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. Binary Classification

Distribution-Free Confidence Sets for the Regression Function

Joint work with: Ambrus Tamás

Binary Classification

- In binary classification the sample $\{(x_j, y_j)\}_{j=1}^n$ consists of inputs, $x_j \in \mathbb{X}$, from a measurable space, and labels, $y_i \in \mathbb{Y} \doteq \{-1, +1\}$.
- The sample is i.i.d. and have (unknown) distribution $\mathbb{P}$ on $\mathbb{X} \times \mathbb{Y}$.
- We call any (measurable) $g : \mathbb{X} \to \{-1, +1\}$ function a classifier.
- A loss function penalizes label mismatch, $\ell : \mathbb{Y} \times \mathbb{Y} \to [0, \infty)$.
- Typical choice: zero-one loss, $\ell(\hat{y}, y) \doteq \mathbb{I}(\hat{y} \neq y) = (1 - \hat{y}y)/2$.
- The overall (expected) risk of classifier g is (cf. "test error")

$$R(f) \doteq \mathbb{E}\big[\ell(g(X), Y)\big] = \int_{\mathbb{X}\times\mathbb{Y}} \ell(g(x), y)\, \mathbb{P}(\mathrm{d}x, \mathrm{d}y),$$

  where X and Y are general random elements with $(X, Y) \sim \mathbb{P}$.
- For the zero-one loss, the risk is simply $R(f) = \mathbb{P}(g(X) \neq Y)$.
- In general, we aim at finding a classifier with minimal risk.

Regression Function

– If distribution $\mathbb{P}$ was known, an ideal choice would be

$$g_* \in \arg\min \Big\{ R(f) \mid g : \mathbb{X} \to \mathbb{Y} \text{ and } g \text{ is measurable} \Big\},$$

called Bayes optimal or target classifier (not unique in general).

– For the zero-one loss, an optimal classifier is (if $\mathbb{P}(\eta(x) \neq 0) = 1$)

$$g_*(x) = \operatorname{sign}(f_*(x)), \qquad \text{where} \qquad f_*(x) \doteq \mathbb{E}\big[Y \mid X = x\big].$$

– Function f_* is a key object, it is called the regression function.

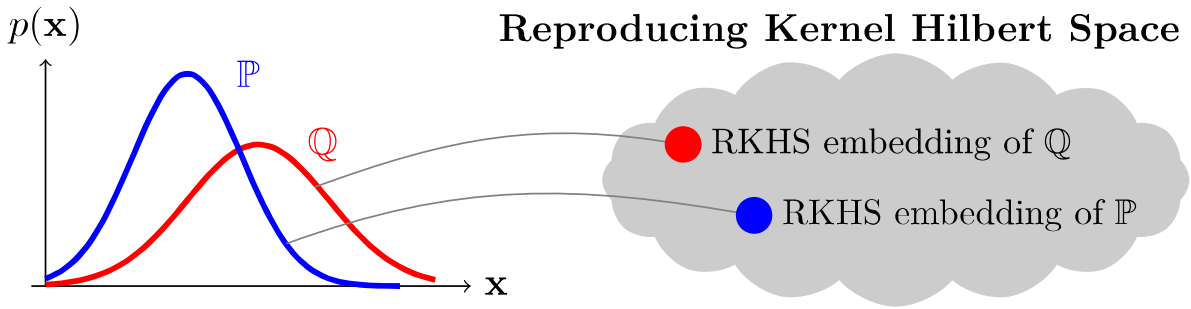
– Note that it contains more information than g_*, as for example the probability of misclassification can also be computed from f_*.

– There are many methods that provide point estimates for f_*, but there are much less that can efficiently build region estimates.

– Here, we aim at building non-asymptotic region estimates for f_*.

Kernel Mean Embedding

– Idea: map distributions to elements of an RKHS with the kernel.



– $\mathcal{D}(\mathbb{X})$ is the set of prob. distributions over meas. space $(\mathbb{X}, \Sigma)$.

– The kernel mean embedding of probability measures into an RKHS $\mathcal{H}$ endowed with a reproducing kernel $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is

$$\mu : \mathcal{D}(\mathbb{X}) \to \mathcal{H},$$

$$\mathbb{P} \to \int_{\mathbb{X}} k(x, \cdot)\, \mathbb{P}(\mathrm{d}x),$$

if this Bochner integral exists, e.g., if $\mathbb{E}_{X \sim \mathbb{P}}\left[\sqrt{k(X, X)}\right] < \infty$.

Universal and Characteristic Kernels

- The kernel embedding has many nice properties, e.g., for $f \in \mathcal{H}$,

$$\mathbb{E}_{X \sim P}\big[\, f(X) \,\big] \;=\; \langle f, \mu_P \rangle_{\mathcal{H}}$$

- If $k(x, y) = \exp(\langle x, y \rangle)$, then we recover the moment generating function (with the Fourier kernel we get the characteristic funct.).
- A kernel is called characteristic if the embedding, μ, is injective.
- A characteristic kernel induces a metric on space $\mathcal{D}(\mathbb{X})$, namely, $d(P, Q) \doteq \|\mu_P - \mu_Q\|_{\mathcal{H}}$, with $d(P, Q) = 0$ if and only if $P = Q$.
- $\mathcal{C}(\mathbb{X})$ is the set of continuous fun. on a compact metric space $\mathbb{X}$.
- A kernel is universal if the corresponding $\mathcal{H}$ is dense in $\mathcal{C}(\mathbb{X})$: for all $f \in \mathcal{C}(\mathbb{X})$ and $\varepsilon > 0$ there is $h \in \mathcal{H}$ such that $\|f - h\|_\infty < \varepsilon$.
- Let $\mathbb{X}$ be a compact metric space and let k be a universal kernel on $\mathbb{X}$, then one can show that k is also characteristic.

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 $p \in \mathbb{N}$).

Resampling Framework

- Let us fix a distribution on $\mathbb{S} \doteq \mathbb{X} \times \mathbb{Y}$, where $\mathbb{X}$ and $\mathbb{Y}$ are the input and output spaces, respectively (in our case $\mathbb{Y} = \{-1, +1\}$).
- The conditional expectation of Y given X can be expressed as

$$f_*(x) \doteq \mathbb{E}\big[\, Y \mid X = x \,\big] = 2 \cdot \mathbb{P}(\, Y = +1 \mid X = x \,) - 1.$$

- We are given an (indexed) family of possible regression functions that also contains f_* (the true system is in the model class):

$$f_* \in \mathcal{F} \doteq \big\{\, f_\theta : \mathbb{X} \to [-1, +1] \mid \theta \in \Theta \,\big\}.$$

- The true "parameter" is denoted by θ^*, namely, $f_{\theta^*} = f_*$.
- Assume that the parametrization is injective (in the $\mathcal{L}^2(\mathbb{X})$ sense).
- Otherwise, Θ can be an arbitrary set! ($\dim(\Theta) = \infty$ is allowed).

Resampling Labels

– The original i.i.d. input-output dataset is denoted by

$$\mathcal{D}_0 \doteq ((x_1, y_1), \dots, (x_n, y_n)).$$

– Given a θ, we can generate $m-1$ alternative samples by

$$\mathcal{D}_i(\theta) \doteq ((x_1, y_{i,1}(\theta)), \dots, (x_n, y_{i,n}(\theta))),$$

for $i = 1, \dots, m-1$, where for all (i, j) label $y_{i,j}(\theta)$ is generated randomly according to the conditional distribution:

$$\mathbb{P}_\theta(Y = y \mid X = x) \doteq 1/2 \left(y(f_\theta(x) + 1)\right).$$

Crucial Observations

– $\mathcal{D}_0$ and $\mathcal{D}_i(\theta^*)$ have the same distribution ("Law"), for i.

– If $\theta \neq \theta^*$, Law$(\mathcal{D}_0)$ is typically different than Law$(\mathcal{D}_i(\theta))$.

Ranking Functions

– Let $\mathbb{A}$ be a measurable space, a function $\psi : \mathbb{A}^m \to [\,m\,]$ where $[\,m\,] \doteq \{1, \ldots, m\}$, is called a ranking function if for all $(a_1, \ldots, a_m) \in \mathbb{A}^m$ it satisfies the two properties:

(P1) For all permutations μ of the set $\{2, \ldots, m\}$, we have

$$\psi\big(\, a_1, a_2, \ldots, a_m \,\big) \;=\; \psi\big(\, a_1, a_{\mu(2)}, \ldots, a_{\mu(m)} \,\big),$$

that is the function is invariant with respect to reordering the last $m-1$ terms of its arguments.

(P2) For all $i, j \in [\,m\,]$, if $a_i \neq a_j$, then we have

$$\psi\big(\, a_i, \{a_k\}_{k\neq i} \,\big) \;\neq\; \psi\big(\, a_j, \{a_k\}_{k\neq j} \,\big).$$

– We can think of ψ as a function which "sorts" the elements and returns the rank of the first element in the order.

Uniform Ordering of Exchangeable Elements

The Main Idea Underlying the Framework

Compare the original dataset with alternative samples randomly generated according to a given hypothesis. Accept the hypothesis if the original dataset behaves "similarly" to the alternative ones and reject otherwise. Measure "similar" behavior with ranking.

– Fundamental quation: how to find a suitable ranking function?

Uniform Ordering Lemma

Let $A_1, \ldots, A_m$ be exchangeable, almost surely pairwise different random elements from $\mathbb{A}$. Then, $\psi\big(A_1, A_2, \ldots, A_m\big)$ has discrete uniform distribution: $\forall\, k \in [m]$, the rank is k with probability $1/m$.

– Pairwise difference is a technical assumptions (cf. tie-breaking).

General Confidence Region Construction

- Given a ranking function ψ (i.e., satisfying P1 and P2).
- User-chosen hyper-parameters $p, q \in [m]$ with $p \leq q$.
- One can build a confidence region based on ψ by

Confidence Region

$$\Theta_{\varrho}^{\psi} \doteq \left\{ \theta \in \Theta : p \leq \psi\left(\mathcal{D}_0, \{\mathcal{D}_k(\theta)\}_{k \neq 0} \right) \leq q \right\}$$

- $\varrho \doteq (m, p, q)$ denotes the hyper-parameters, with $m \geq 1$ being the total number of samples (original & alternative datasets).
- Intuitively: the region contains those models for which the rank of the original dataset compared to the ranks of the alternative ones, generated based on the model, is neither too low nor too high.

Exact Confidence

– The main abstract result of the resampling framework is:

Theorem: Exact Confidence

We have for all ranking function ψ and hyper-parameter $\varrho = (m, p, q)$ with integers $1 \leq p \leq q \leq m$ that

$$\mathbb{P}\left(\theta^* \in \Theta_{\varrho}^{\psi}\right) = \frac{q - p + 1}{m}.$$

– Note: ψ is an arbitrary ranking function (satisfying P1 and P2).

– The coverage probabilty is user-chosen (rational), and exact.

– This probability is independent of the underlying probability distribution generating the data, the result is distribution-free.

– Further, the claim is non-asymptotic (holds for finite samples).

Strong Consistency

- Warning: exact confidence in itself could be misleading as, for example, purely randomized methods can have this property.
- We also study other properties of the methods, e.g., consistency.
- Formally, a method is strongly consistent if

$$\mathbb{P}\left(\bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} \left\{ \theta \in \Theta_{\varrho,n}^{\psi} \right\} \right) = 0,$$

for all parameter $\theta \neq \theta^*$, $\theta \in \Theta$, where $\Theta_{\varrho,n}^{\psi}$ denotes the confidence region constructed based on a sample of size n.

- Informally: eventually, as the sample size tends to infinity, any false parameter will be excluded from the regions with probability one.

Kernel-Based Constructions

– Now, we propose three kernel-based algorithms:

1. A neighborhood based (Algorithm I)
2. An embedding based (Algorithm II)
3. A discrepancy based (Algorithm III)

– Each method builds region-estimates (confidence regions) for the underlying regression function of binary classification.

– They are based on the suggested resampling framework and all of them have exact coverage probabilities and are strongly consistent.

Algorithm I: Neighborhood Based

- If there is a metric on the input space, $\mathbb{X}$, we can estimate f_* based on the original dataset by kNN (k-nearest neighbors).
- Similarly, we can estimate f_* based on the alternative datasets:

$$f^{(i)}_{\theta,n}(x) \doteq \frac{1}{k_n} \sum_{j=1}^{n} y_{i,j}(\theta)\, \mathbb{I}\big(x_j \in N(x, k_n)\big),$$

for $i = 0, \ldots, m-1$, where $\mathbb{I}$ is an indicator function (its value is 1 if its argument is true, and 0 otherwise), $N(x, k_n)$ denotes the k_n closest neighbors of x from $\{x_j\}_{j=1}^{n}$, and $k_n \leq n$ is a constant (window size), which can depend on the sample size n.

- Idea: we can construct a ranking function by comparing the "distances" of these functions from the model generating the data.

Algorithm I: Neighborhood Based

– The $\mathcal{L}^2(\mathbb{X})$ distance of the i th estimate from the model is

$$Z_n^{(i)}(\theta) \doteq \left\| f_{\theta,n}^{(i)} - f_\theta \right\|_2^2,$$

it can be calculated directly or by Monte Carlo approximations.

– Then, we can define the rank of $Z_n^{(0)}$ among $\{Z_n^{(i)}(\theta)\}$ as

$$\mathcal{R}_n(\theta) \doteq 1 + \sum_{i=1}^{m-1} \mathbb{I}\left(Z_n^{(0)} \prec_\pi Z_n^{(i)}(\theta) \right),$$

where "$\prec_\pi$" is the standard "$<$" with random tie-breaking.

– Finally, the confidence region can be constructed as

$$\Theta_{\varrho,n}^{(1)} \doteq \left\{ \theta \in \Theta : \mathcal{R}_n(\theta) \leq q \right\}.$$

Algorithm I: Neighborhood Based

Theorem: Stochastic Guarantees of Algorithm I

Assume that the following properties hold

1. *The input space is* $\mathbb{X} \subseteq \mathbb{R}^d$ *and* $\mathbb{X}$ *is compact.*
2. *The support of the input distribution,* $P_{\mathbb{X}}$*, is the whole* $\mathbb{X}$*.*
3. *The input distribution,* $P_{\mathbb{X}}$*, is absolutely continuous.*

Then, the coverage probability of the constructed region is

$$\mathbb{P}\left(\theta^* \in \Theta^{(1)}_{\varrho,n}\right) = q \,/\, m,$$

i.e., it is exact for any sample size n*. Moreover, if* $\{k_n\}$ *are chosen such that* $k_n \to \infty$ *and* $k_n/n \to 0$*, as* $n \to \infty$*, then the confidence sets are strongly consistent (eventually exclude false parameters).*

Algorithm II: Embedding Based

- Idea: embed the distribution of the original sample and that of the alternative ones in an RKHS using a characteristic kernel.
- The kernel mean embedding of the true distribution generating the data $(*)$ and the one based on a hypothetical model (θ) are

$$h_*(\cdot) \doteq \mathbb{E}\big[k(\cdot, S_*)\big], \qquad \text{and} \qquad h_\theta(\cdot) \doteq \mathbb{E}\big[k(\cdot, S_\theta)\big],$$

where S_* and S_θ are random elements from $\mathbb{S} = \mathbb{R}^d \times \{+1, -1\}$, distributed according to true distribution and the tested one.

- Functions $h_*(\cdot)$ and $h_\theta(\cdot)$ can be estimated from empirical data:

$$h^{(i)}_{\theta,n}(\cdot) \doteq \frac{1}{n}\sum_{j=1}^{n} k(\cdot, s_{i,j}(\theta)),$$

for $i = 0, \ldots, m-1$, where $s_{i,j}(\theta) \doteq (x_j, y_{i,j}(\theta))$.

Algorithm II: Embedding Based

- The kernel is characteristic, therefore, $h_\theta = h_* \iff \theta = \theta^*$.
- The construction of the confidence region is as follows

$$Z_n^{(i)}(\theta) \doteq \sum_{j=0}^{m-1} \| h_{\theta,n}^{(i)} - h_{\theta,n}^{(j)} \|_{\mathcal{H}}^2$$

$$\mathcal{R}_n(\theta) \doteq 1 + \sum_{i=1}^{m-1} \mathbb{I}\left(Z_n^{(0)} \prec_\pi Z_n^{(i)}(\theta) \right)$$

$$\Theta_{\varrho,n}^{(2)} \doteq \left\{ \theta \in \Theta : \mathcal{R}_n(\theta) \leq q \right\}$$

- Note: cumulative distances are used in the definition of $\{Z_n^{(i)}(\theta)\}$.
- The terms $\| h_{\theta,n}^{(i)} - h_{\theta,n}^{(j)} \|_{\mathcal{H}}^2$ can be easily computed in practice using the Gram matrix (based on the reproducing property).

Algorithm II: Embedding Based

Theorem: Stochastic Guarantees of Algorithm II

Assume that the following properties hold

1. $\mathcal{H}$ *is a separable RKHS containing* $\mathbb{S} \to \mathbb{R}$ *functions.*
2. *The kernel is (measurable) bounded and characteristic.*

Then, the confidence regions of Algorithm II have exact coverage

$$\mathbb{P}\left(\theta^* \in \Theta^{(2)}_{\varrho,n}\right) = q \,/\, m,$$

for any sample size n; and they are strongly consistent, if $m \geq 3$.

– One can show that $\mathrm{Var}(k(\cdot, S)) < \infty$, for $S \in \{S_*, S_\theta\}$, therefore a Hilbert space valued strong law of large numbers can be applied.

– Algorithm II is of theoretical interest as it is computationally heavy.

Algorithm III: Discrepancy Based

– In order to formalize the method, let us introduce residuals

$$\varepsilon_{i,j}(\theta) \doteq y_{i,j}(\theta) - f_\theta(x_j)$$

for $i = 0, \dots, m-1$ and $j = 1, \dots, n$. Note that if $i \neq 0$, $\varepsilon_{i,j}(\theta)$ has zero mean for all j, as $f_\theta(x_j) = \mathbb{E}_\theta\big[\, y_{i,j}(\theta) \,|\, x_j \,\big]$.

– Algorithm III constructs the confidence region as

$$Z_n^{(i)}(\theta) \doteq \left\| \frac{1}{n} \sum_{j=1}^{n} \varepsilon_{i,j}(\theta)\, k(\cdot, x_j) \right\|_{\mathcal{H}}^2 = \frac{1}{n^2}\, \varepsilon_i^{\mathrm{T}}(\theta)\, K\, \varepsilon_i(\theta)$$

$$\mathcal{R}_n(\theta) \doteq 1 + \sum_{i=1}^{m-1} \mathbb{I}\big(Z_n^{(0)} \prec_\pi Z_n^{(i)}(\theta) \big)$$

$$\Theta_{\varrho,n}^{(3)} \doteq \big\{ \theta \in \Theta : \mathcal{R}_n(\theta) \leq q \big\}$$

Algorithm III: Discrepancy Based

Theorem: Stochastic Guarantees of Algorithm III

Assume that the following properties hold

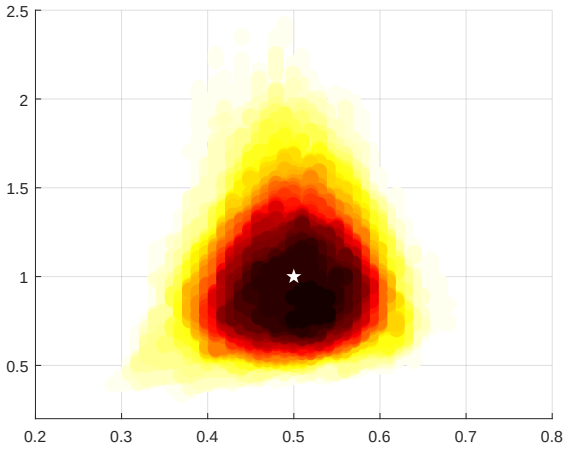
1. $\mathcal{H}$ *is a separable RKHS containing* $\mathbb{X} \to \mathbb{R}$ *functions.*
2. *The kernel is (measurable) bounded and universal.*
3. $\mathbb{X}$ *is a compact Polish metric space (complete and separable).*
4. *Each potential regression function* $f \in \mathcal{F}$ *is continuous.*

Then, the confidence regions of Algorithm III have exact coverage

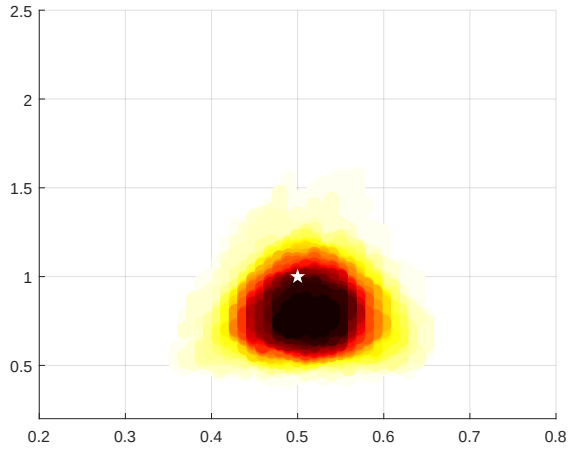
$$\mathbb{P}\left(\theta^* \in \Theta^{(3)}_{\varrho,n} \right) = q / m,$$

for any sample size n; and they are strongly consistent.

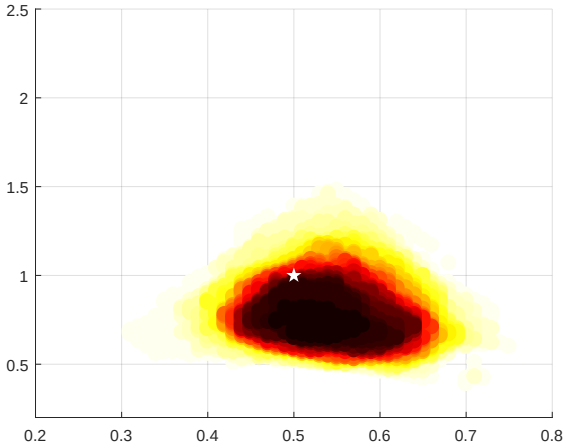
Experiments: Ranks in the Parameter Space



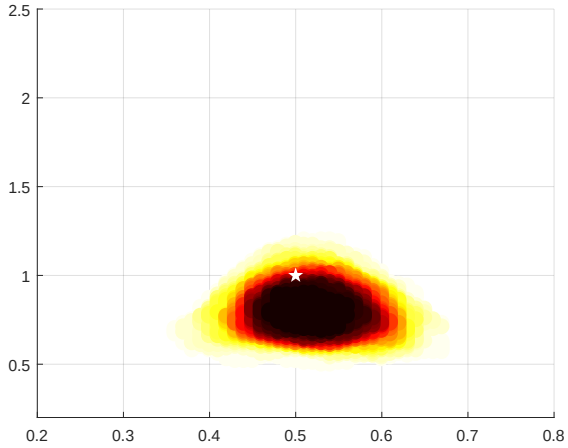
(a) Neighborhood based (kNN)



(b) Neighborhood based (Gauss)

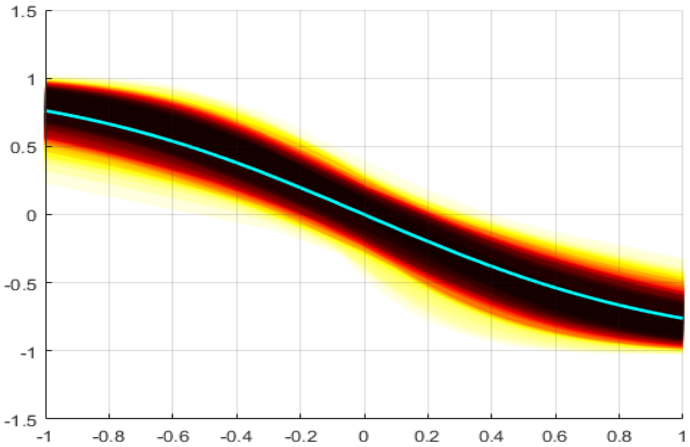


(c) Embedding based (Gauss)

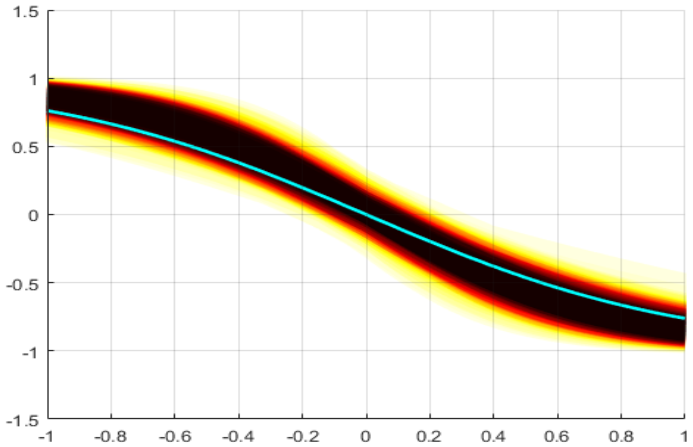


(d) Discrepancy based (Gauss)

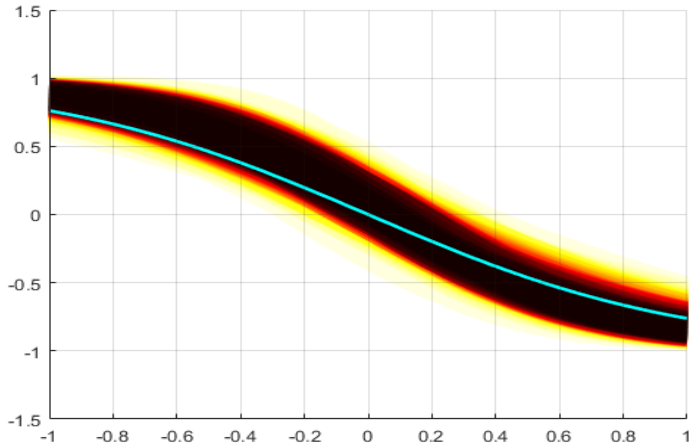
Experiments: Ranks in the Model Space



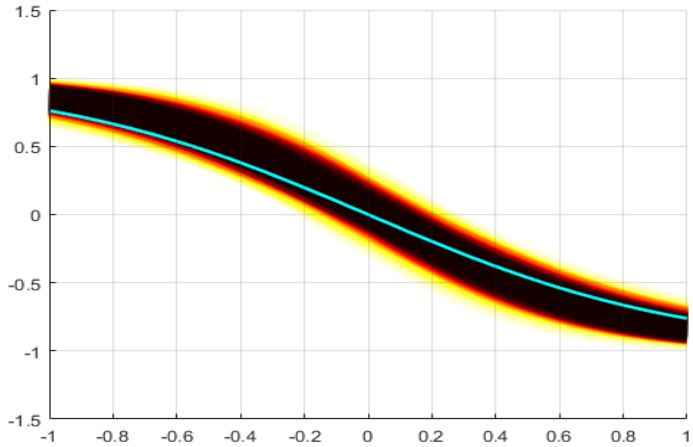
(a) Neighborhood based (kNN)



(b) Neighborhood based (Gauss)



(c) Embedding based (Gauss)



(d) Discrepancy based (Gauss)

Summary for Binary Classification

- The regression function is a key object of binary classification, as it can provide an optimal classifier and can also evaluate the risk.
- We constructed region estimates for the regression function.
- A general framework based on resampling was presented with which confidence regions with exact coverage can be built.
- A general non-asymptotic theorem ensuring this was provided.
- The approach is nonparametric as it can handle arbitrary types of regression functions (e.g., their space can be infinite dimensional).
- Three particular kernel-based (resampling) methods were given based on neighborhoods, (mean) embeddings and discrepancy.
- Besides having exact coverage probabilities, we argued that each method is strongly consistent, as well (under mild assumptions).
- Finally, numerical experiments were shown supporting the ideas.

Thank you for your attention!

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