Distribution-Free Uncertainty Quantification for Kernel Methods by Gradient Perturbations

Balázs Csanád Csáji & Krisztián Balázs Kis

Institute for Computer Science and Control (SZTAKI), Budapest, Hungary

Overview

- Data-driven *uncertainty quantification* (UQ) for models built by kernel methods.
- UQ takes the form of *confidence regions* for ideal representations of the true function.
- The core idea is to *perturb the residuals* in the *gradient* of the objective function.
- *Distribution-free* (unlike GP regression), only some mild regularities are assumed.
- *Non-asymptotic* (finite sample) guarantees.
- *Exact* (user-chosen) coverage probabilities.
- Convex quadratic problems and symmetric noises ⇒ the regions are *star convex* and have *ellipsoidal outer approximations*.
- Examples: LS-SVM, KRR, SVR & kLASSO.

Preliminaries

We are given a *data sample*, $\mathcal{D}_n$, of observations,

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

with $\mathcal{X} \neq \emptyset$. Let $x \doteq (x_1, \dots, x_n)^{\mathrm{T}} \in \mathcal{X}^n$ and $y \doteq (y_1, \dots, y_n)^{\mathrm{T}} \in \mathbb{R}^n$. The Gram matrix of a *kernel* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, w.r.t. input vector x, is

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

Let $\mathcal{H}$ be an RKHS induced by kernel k. Then, for any *objective* function g having the form

$$g(f, \mathcal{D}_n) \doteq \mathrm{L}\big((x_1, y_1, f(x_1)), \dots, (x_n, y_n, f(x_n))\big) + \Lambda(\,\|f\|_{\mathcal{H}}\,),$$

where L is an arbitrary loss function, Λ is a non-decreasing regularizer, there is a solution with

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

which is ensured by the *representer theorem*.

Ideal Representations

Let the data be generated by noisy observations of an underlying *true function*, f_*, for $i = 1, \dots, n$,

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

where $\{\varepsilon_i\}$ are the noises; let $\varepsilon \doteq (\varepsilon_1, \dots, \varepsilon_n)^{\mathrm{T}}$.

Let $\mathcal{H}_\alpha \subseteq \mathcal{H}$ be the subspace of f_α functions. An $f_0 \in \mathcal{H}_\alpha$, having coefficients $\alpha^* \in \mathbb{R}^n$, is called an *ideal representation* of f_* w.r.t. $\mathcal{D}_n$, if for all i,

$$f_0(x_i) \;=\; f_*(x_i).$$

Note that α^* is unique if $\mathrm{rank}(\mathrm{K}_x) = n$, since ideal coefficients satisfy $\mathrm{K}_x \alpha^* = (f_*(x_1), \dots, f_*(x_n))^{\mathrm{T}}$.

Distributional Invariance

An $\mathbb{R}^n$-valued random vector ε is *distributionally invariant* w.r.t. a compact group of transformations, $(\mathcal{G}, \circ)$, where "$\circ$" is 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.

E.g.: $\{\varepsilon_i\}$ are *exchangeable* ($\mathcal{G}$: permutations); or independent and *symmetric* ($\mathcal{G}$: sign-changes).

Main Assumptions

A1 *The kernel, k, is strictly positive definite and all inputs, $\{x_i\}$, are almost surely distinct.*

A2 *The input vector x and the noise vector ε are independent (from each other, not internally).*

A3 *The noises, $\{\varepsilon_i\}$, are distrib. invariant w.r.t. 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 $\bar{g}$,*

$$\nabla_\alpha\, g(f_\alpha, \mathcal{D}_n) \;=\; \bar{g}(x, \alpha, \hat{\varepsilon}(x, y, \alpha)),$$

where the residuals are defined as

$$\hat{\varepsilon}(x, y, \alpha) \;\doteq\; y - \mathrm{K}_x \alpha.$$

A1 ⇒ α^* is a.s. unique; A2 ⇒ no autoregression; A3 ⇒ ε can be perturbed; A4 holds in most cases.

Perturbed Gradients

Let us define a *reference* function, $Z_0 : \mathbb{R}^n \to \mathbb{R}$, and $m - 1$ *perturbed* functions, $\{Z_i\}$, with $Z_i : \mathbb{R}^n \to \mathbb{R}$, where m is a hyper-parameter, as

$$Z_0(\alpha) \;\doteq\; \|\, \Psi(x)\, \bar{g}(x, \alpha, G_0(\hat{\varepsilon}(x, y, \alpha)))\, \|^2,$$

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

for $i = 1, \dots, m-1$, where $\Psi(x)$ is a weighting matrix, G_0 is the identity element of $\mathcal{G}$ and $\{G_i\}$ are uniformly sampled i.i.d. elements from $\mathcal{G}$.

Note that if $\alpha = \alpha^*$, $Z_0(\alpha^*) \overset{d}{=} Z_i(\alpha^*)$, for all i. On the other hand, for $\alpha \neq \alpha^*$, 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 the reference element, $Z_0(\alpha)$, among all $\{Z_i(\alpha)\}_{i=0}^{m-1}$ elements is

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

where $\mathbb{I}(\cdot)$ is an indicator function and binary relation "$\prec_\pi$" is "$<$" with random tie-breaking.

A *confidence region* for probability $p = 1 - q/m$ is

$$A_p \doteq \big\{\, \alpha : \mathcal{R}(\alpha) \,\leq\, 1 - q/m \,\big\},$$

where $m, q \in \mathbb{N}$ with $0 < q < m$ are user-chosen.

The main *non-asymptotic* and *distribution-free* claim about the stochastic guarantees of A_p is:

Main Theorem

Under Assumptions A1, A2, A3 and A4, the *coverage probability* of the confidence region w.r.t. the ideal coefficient vector α^* is *exactly*

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

for any hyper-parameters with $0 < q < m$.

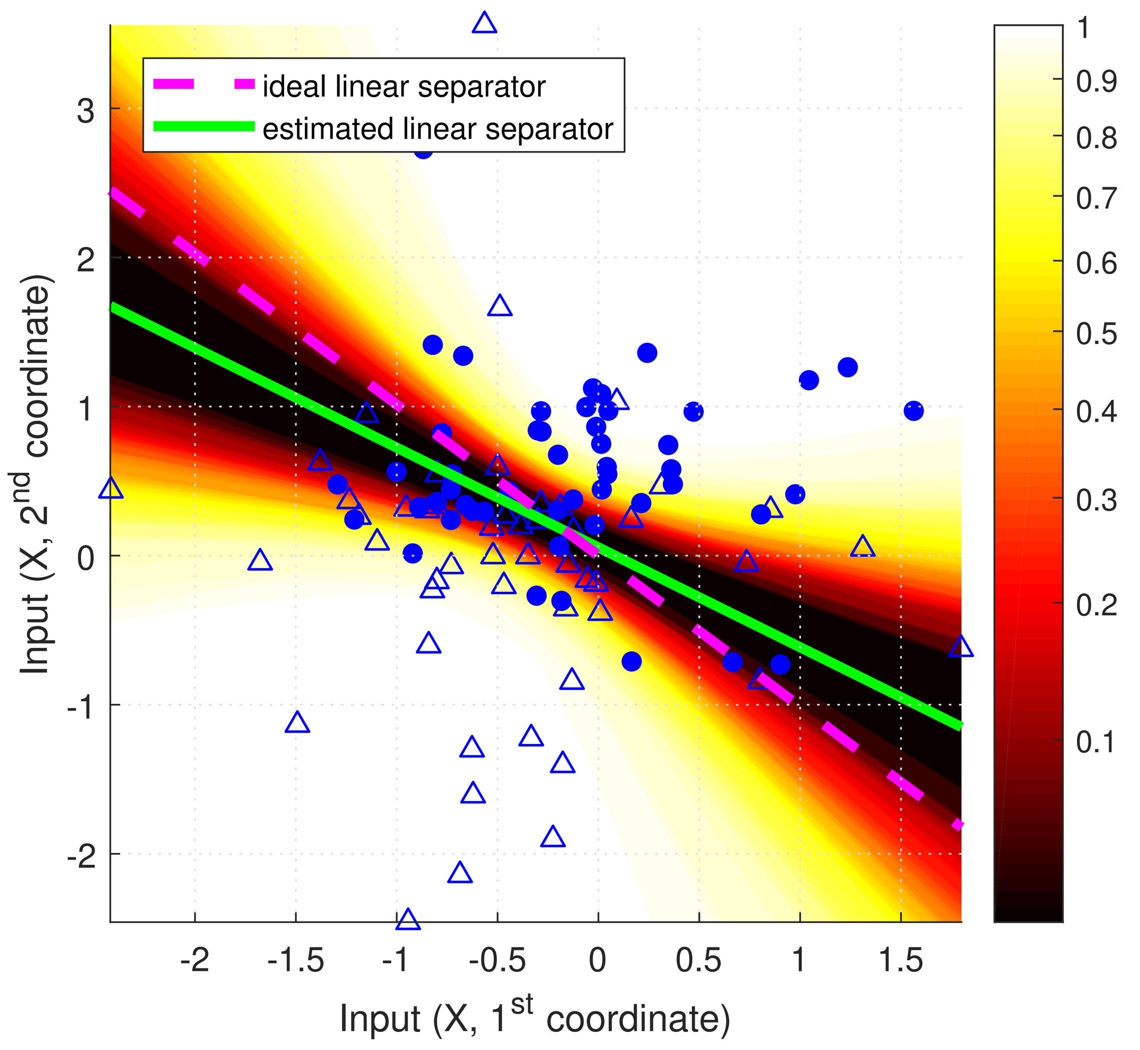

Figure 1: UQ for (linear) LS-SVM classification in the model space based on $n = 100$ observations ($\mathcal{G}$: sign-changes).

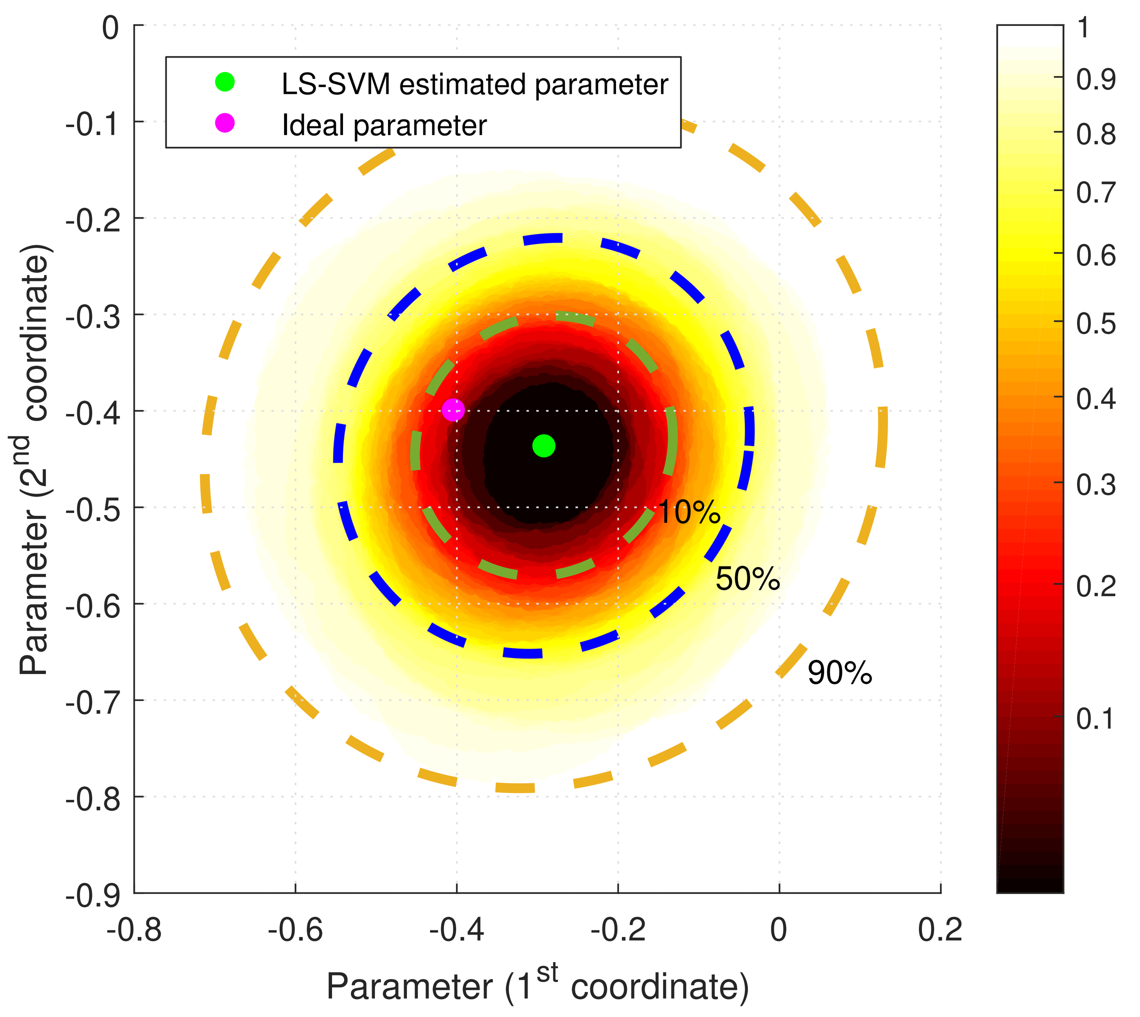

Figure 2: UQ for (linear) LS-SVM classification in the parameter space with various non-asymptotic confidence ellipsoids.

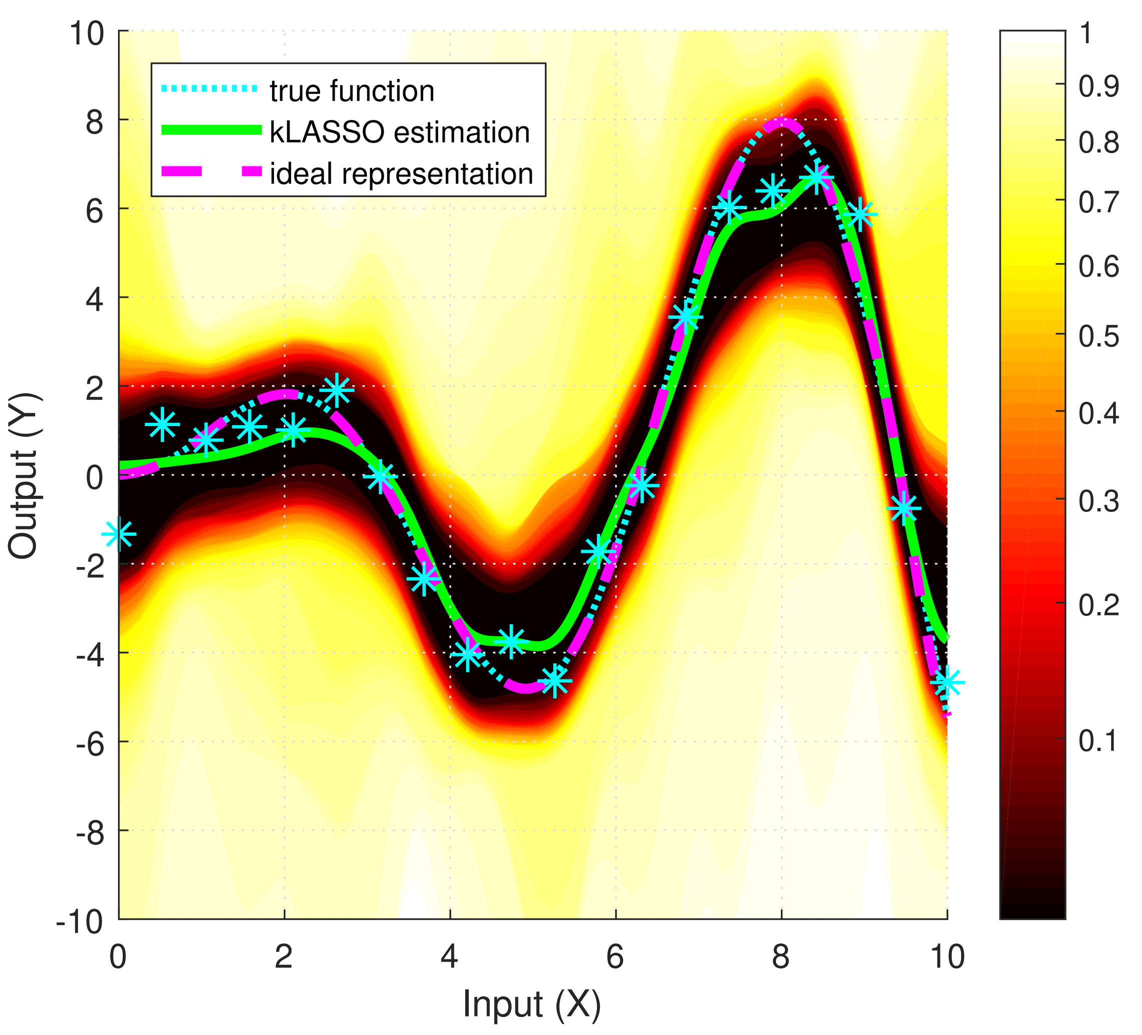

Figure 3: UQ for kernelized LASSO (Gaussian kernel) based on $n = 20$ observations with Laplace noises ($\mathcal{G}$: sign-changes).

Contact Information

Email: csaji@sztaki.hu (Balázs Cs. Csáji)

Website: http://www.sztaki.hu/~csaji/

Poster presented at ECML-PKDD, Würzburg, Germany, 2019