Distribution-Free Uncertainty Quantification for Kernel Methods by Gradient Perturbations

Balázs Csanád Csáji · Krisztián Balázs Kis

Abstract We propose a data-driven approach to quantify the uncertainty of models constructed by kernel methods. Our approach minimizes the needed distributional assumptions, hence, instead of working with, for example, Gaussian processes or exponential families, it only requires knowledge about some mild regularity of the measurement noise, such as it is being symmetric or exchangeable. We show, by building on recent results from finite-sample system identification, that by perturbing the residuals in the gradient of the objective function, information can be extracted about the amount of uncertainty our model has. Particularly, we provide an algorithm to build exact, non-asymptotically guaranteed, distribution-free confidence regions for ideal, noise-free representations of the function we try to estimate. For the typical convex quadratic problems and symmetric noises, the regions are star convex centered around a given nominal estimate, and have efficient ellipsoidal outer approximations. Finally, we illustrate the ideas on typical kernel methods, such as LS-SVC, KRR, ε-SVR and kernelized LASSO.

Keywords kernel methods · confidence regions · nonparametric regression · classification · support vector machines · distribution-free methods

Balázs Csanád Csáj
EPIC Centre of Excellence
MTA SZTAKI: Institute for Computer Science and Control
Hungarian Academy of Sciences, Budapest, Hungary
Tel.: +(36)-1-279-6231 Fax: +(36)-1-279-7503
E-mail: balazs.csaji@sztaki.mta.hu

Krisztián Balázs Kis
EPIC Centre of Excellence
MTA SZTAKI: Institute for Computer Science and Control
Hungarian Academy of Sciences, Budapest, Hungary
Tel.: +(36)-1-279-6111 Fax: +(36)-1-279-7503
E-mail: krisztian.kis@sztaki.mta.hu

1 Introduction

Kernel methods build on the fundamental concept of Reproducing Kernel Hilbert Spaces (Aronszajn, 1950; Giné and Nickl, 2015) and are widely used in machine learning (Shawe-Taylor and Cristianini, 2004; Hofmann et al., 2008) and related fields, such as system identification (Pillonetto et al., 2014). One of the reasons of their popularity is the representer theorem (Kimeldorf and Wahba, 1971; Schölkopf et al., 2001) which shows that finding an estimate in an infinite dimensional space of functions can be traced back to a finite dimensional problem. Support vector machines (Schölkopf and Smola, 2001; Steinwart and Christmann, 2008), rooted in statistical learning theory (Vapnik, 1998), are typical examples of kernel methods.

Besides how to construct efficient models from data, it is also a fundamental question how to quantify the *uncertainty* of the obtained models. While standard approaches like Gaussian processes (Rasmussen and Williams, 2006) or exponential families (Hofmann et al., 2008) offer a nice theoretical framework, making strong statistical assumptions on the system is sometimes unrealistic, since in practice we typically have very limited knowledge about the noise affecting the measurements. Building on asymptotic results, such as limiting distributions, is also widespread (Giné and Nickl, 2015), but they usually lack finite sample guarantees.

Here, we propose a *non-asymptotic*, *distribution-free* approach to quantify the uncertainty of kernel-based models, which can be used for *hypothesis testing* and *confidence region* constructions. We build on recent developments in finite-sample system identification (Campi and Weyer, 2005; Carè et al., 2018), more specifically, we build on the Sign-Perturbed Sums (SPS) algorithm (Csáji et al., 2015) and its generalizations, the Data Peturbation (DP) methods (Kolumbán, 2016).

We consider the case where there is an underlying "true" function that generates the measurements, but we only have noisy observations of its outputs. Since we want to minimize the needed assumptions, for example, we do not want to assume that the true underlying function belongs to the Hilbert space in which we search our estimate, we take a "honest" approach (Li, 1989) and consider "ideal" representations of the target function from our function space. A representation is ideal w.r.t. the data sample, if its outputs coincide with the corresponding (hidden) noise-free outputs of the true underlying function for all available inputs.

Despite our method is *distribution-free*, i.e., it does not depend on any parameterized distributions, it has strong *finite-sample guarantees*. We argue that, the constructed confidence region contains the ideal representation *exactly* with a user-chosen probability. In case the noises are independent and symmetric about zero, and the objective function is convex quadratic, the resulting regions are *star convex* and have efficient *ellipsoidal outer approximations*, which can be computed by solving semi-definite optimization problems. Finally, we demonstrate our approach on typical kernel methods, such as KRR, SVMs and kernelized LASSO.

Our approach has some similarities to bootstrap (Efron and Tibshirani, 1994) and conformal prediction (Vovk et al., 2005). One of the fundamental differences w.r.t bootstrap is, e.g., that we avoid building alternative samples and fitting bootstrap estimates to them (since it is computationally challenging), but perturb directly the gradient of the objective function. Key differences w.r.t. conformal prediction are, e.g., that we want to quantify the uncertainty of the model and not necessarily that of the next observation (though the two problems are related), and more importantly, exchangeability is not fundamental for our approach.

2 Preliminaries

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 the point evaluation functional

$$\delta_z : f \to f(z), \tag{1}$$

is continuous (or equivalently bounded) for all $z \in \mathcal{X}$, at any $f \in \mathcal{H}$ (Giné and Nickl, 2015). Then, by using the Riesz representation theorem, one can construct a (unique) kernel, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, having the *reproducing* property, that is

$$\langle k(\cdot, z), f \rangle_{\mathcal{H}} \,=\, f(z), \tag{2}$$

for all $z \in \mathcal{X}$ and $f \in \mathcal{H}$. In particular, the kernel satisfies for all $z, s \in \mathcal{X}$ that

$$k(z, s) \,=\, \langle k(\cdot, z), k(\cdot, s) \rangle_{\mathcal{H}} \,. \tag{3}$$

Hence, the kernel of an RKHS is a symmetric and positive-definite function; moreover, the Moore-Aronszajn theorem states that the converse is also true: for every symmetric, positive-definite function there is a unique RKHS (Aronszajn, 1950).

Typical kernels include, e.g., the Gaussian kernel $k(z, s) = \exp(-\|z-s\|^2/2\sigma^2)$, with $\sigma > 0$, the polynomial kernel, $k(z, s) = (\langle z, s \rangle + c)^p$, with $c \geq 0$ and $p \in \mathbb{N}$, and the sigmoidal kernel, $k(z, s) = \tanh(a \langle z, s \rangle + b)$ for some $a, b \geq 0$, where $\langle \cdot, \cdot \rangle$ denotes the standard Euclidean inner product (Hofmann et al., 2008).

By a *data sample*, $\mathcal{D}_n$, we mean a finite set of input-output measurements,

$$(x_1, y_1), \, \dots, \, (x_n, y_n) \,\in\, \mathcal{X} \times \mathbb{R}, \tag{4}$$

with $\mathcal{X} \neq \emptyset$. We also introduce $x \doteq (x_1, \dots, x_n)^{\mathrm{T}} \in \mathcal{X}^n$ and $y \doteq (y_1, \dots, y_n)^{\mathrm{T}} \in \mathbb{R}^n$. The *Gram matrix* of $k(\cdot, \cdot)$, w.r.t. input x, is denoted by $\mathrm{K}_x \in \mathbb{R}^{n \times n}$, where

$$[\, \mathrm{K}_x \,]_{i,j} \,\doteq\, k(x_i, x_j). \tag{5}$$

A kernel is called *strictly* positive definite if its Gram matrix, K_x, is (strictly) positive definite for *distinct* inputs $\{x_i\}$ (Hofmann et al., 2008).

One of the fundamental reasons for the successes of kernel methods is the so-called *representer theorem*, originally given by Kimeldorf and Wahba (1971), but the generalization presented here is due to Schölkopf et al. (2001).

Theorem 1 *Suppose we are given a sample, $\mathcal{D}_n$, 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 of functions*

$$\mathcal{F} \,\doteq\, \Big\{ f : \mathcal{X} \to \mathbb{R} \;\big|\; f(z) = \sum_{i=1}^{\infty} \beta_i k(z, z_i), \; \beta_i \in \mathbb{R}, \; z_i \in \mathcal{X}, \; \|f\|_{\mathcal{H}} < \infty \Big\}, \tag{6}$$

then, for any monotonically increasing regularization function, $\Lambda : [0, \infty) \to [0, \infty)$, and an arbitrary loss function $\mathrm{L} : (\mathcal{X} \times \mathbb{R}^2)^n \to \mathbb{R} \cup \{\infty\}$, the objective

$$g(f, \mathcal{D}_n) \,\doteq\, \mathrm{L}\big((x_1, y_1, f(x_1)), \dots, (x_n, y_n, f(x_n)) \big) \,+\, \Lambda(\, \|f\|_{\mathcal{H}} \,), \tag{7}$$

has a minimizer admitting the following representation

$$f_\alpha(z) \,=\, \sum_{i=1}^{n} \alpha_i \, k(z, x_i), \tag{8}$$

where $\alpha \doteq (\alpha_1, \dots, \alpha_n)^{\mathrm{T}} \in \mathbb{R}^n$ is the vector of coefficients. If Λ is strictly monotonically increasing, then each minimizer admits a representation having form (8).

The theorem can be extended with a bias term (Schölkopf and Smola, 2001), in which case if the solution exists, it also contains a multiple of the bias term. For further generalizations, see (Yu et al., 2013; Argyriou and Dinuzzo, 2014).

The power of the representer theorem comes from the fact that it shows that computing the point estimate in a high, typically infinite, dimensional space of models can be reduced to a much simpler (finite dimensional) optimization problem whose dimension does not exceed the size of the data sample we have, that is n.

If the data is noisy, then of course, the obtained estimate is a *random* function and it is of natural interest to study the *distribution* of the resulting function, for example, to evaluate its *uncertainty* or to test *hypotheses* about the system.

3 Confidence Regions for Kernel Methods

Now, we turn our attention to a stochastic variant of the problem discussed above. There are several advantages of taking a statistical point of view on kernel methods, including conditional modeling, dealing with structured responses, handling missing measurements and building prediction regions (Hofmann et al., 2008).

Following a standard statistical viewpoint (Davies et al., 2009), we assume that the outputs $\{y_i\}$ are generated by some noisy observations of an underlying "true" function, denoted by f_*, that is for all $i = 1, \dots, n$, the outputs can be written as

$$y_i \doteq f_*(x_i) + \varepsilon_i, \tag{9}$$

where $\{\varepsilon_i\}$ are the noise terms. The entire noise vector is $\varepsilon \doteq (\varepsilon_1, \dots, \varepsilon_n)^{\mathrm{T}}$. The noiseless outputs of function f_* will be denote by $y_i^* \doteq f_*(x_i)$, for $i = 1, \dots, n$.

3.1 Ideal Representations

We aim at *quantifying the uncertainty* of our estimated model. A standard way to measure the quality of a point-estimate is to build *confidence regions* around it. However, it is not obvious what we should aim for with our confidence regions. For example, since all of our models live in our RKHS, $\mathcal{H}$, we would like to treat the confidence region as a subset of $\mathcal{H}$. On the other hand, we want to minimize the assumptions, for example, we may not want to assume that f_* is an element of $\mathcal{H}$. Furthermore, since unless we make strong smoothness assumptions on the underlying unobserved function, we only have information about it at the actual inputs, $\{x_i\}$. Hence, we aim for a "honest" nonparametric approach (Li, 1989) and search for functions which correctly describe the hidden function, f_*, on the given inputs. Then, by the representer theorem, we may restrict ourselves to a finite dimensional subspace of $\mathcal{H}$. This leads us to the definition of *ideal representations*:

Definition 1 Let $\mathcal{H}_\alpha \subseteq \mathcal{H}$ denote the subspace of functions that can be represented as (8). A function $f_0 \in \mathcal{H}_\alpha$, having coefficients $\alpha_0 \in \mathbb{R}^n$, is called an *ideal* or noise-free representation of the "true" unobserved function f_*, if we have

$$f_0(x_i) = y_i^* \doteq f_*(x_i), \qquad \text{for all} \qquad i \in \{1, \dots, n\}. \tag{10}$$

The set of all ideal representations, w.r.t. data sample $\mathcal{D}_n$, is denoted by $\mathcal{H}_0 \subseteq \mathcal{H}_\alpha$, and the set of their coefficients, called *ideal coefficients*, is denoted by $A_0 \subseteq \mathbb{R}^n$.

An ideal representation does not simply interpolate the observed (noisy) outputs $\{y_i\}$, but it interpolates the *unobserved* (noise-free) outputs, that is $\{y_i^*\}$.

A natural question which arises is: when does such an ideal representation exist? To answer this question, first note that since ideal representations have the form (8), equation system (10) can be rewritten in a matrix form by using the Gram matrix. That is, vector α is an ideal coefficient vector, if and only if

$$\mathrm{K}_x \alpha \,=\, y^*, \tag{11}$$

where $y^* \doteq (y_1^*, \dots, y_n^*)^\mathrm{T}$. If K_x is (strictly) positive definite, which is the case if for example the kernel is Gaussian and all inputs are distinct, then $\mathrm{rank}(\mathrm{K}_x) = n$ and *every* $f_* : \mathcal{X} \to \mathbb{R}$ has a *unique* ideal representation w.r.t. data sample $\mathcal{D}_n$.

On the other hand, if $\mathrm{rank}(\mathrm{K}_x) < n$, then (11) places a restriction on the functions which have ideal representations. For example, if $\mathcal{X} = \mathbb{R}$ and $\ker(z, s) = \langle z, s \rangle = z^\mathrm{T} s$, then $\mathrm{rank}(\mathrm{K}_x) = 1$ and in general only functions which are *linear* on the data sample have ideal representations. This is of course not surprising, as it is well-known that the choice of the *kernel* encodes our *inductive bias* on the underlying true function we aim at estimating (Schölkopf and Smola, 2001).

If $\mathrm{rank}(\mathrm{K}_x) < n$ and there is an α which satisfies (11), then there are infinitely many ideal representations, as for all $\nu \in \mathrm{null}(\mathrm{K}_x)$, the null space of K_x, we have $\mathrm{K}_x(\alpha + \nu) = \mathrm{K}_x\alpha + \mathrm{K}_x\nu = \mathrm{K}_x\alpha = y^*$. The opposite is also true, if α and β both satisfy (11), then $\mathrm{K}_x(\alpha - \beta) = \mathrm{K}_x\alpha - \mathrm{K}_x\beta = 0$, thus, $\alpha - \beta \in \mathrm{null}(\mathrm{K}_x)$. Hence, to avoid allowing infinitely many ideal representations, we may form *equivalence classes* by treating coefficient vectors α and β equivalent if $\mathrm{K}_x\alpha = \mathrm{K}_x\beta$. Then, we can work with the resulting *quotient space* of coefficients to ensure that there is only one ideal representation (i.e., one equivalence class of such representations).

All of our theory goes trough if we work with the quotient space of representations, but to simplify the presentation we make the assumption (cf. Section 4.2) that K_x is full rank, therefore, there always *uniquely exists* an ideal representation (for any "true" function), whose unique coefficient vector will be denoted by α^*.

3.2 Exact and Honest Confidence Regions

Let $(\Omega, \mathcal{A}, \{\mathbb{P}_\theta\}_{\theta \in \Theta})$ be a *statistical space*, where Θ denotes an arbitrary *index set*. In other words, for all $\theta \in \Theta$, $(\Omega, \mathcal{A}, \mathbb{P}_\theta)$ is a probability space, where Ω is the sample space, $\mathcal{A}$ is the σ-algebra of events, and $\mathbb{P}_\theta$ is a probability measure. Note that it is *not assumed* that $\Theta \subseteq \mathbb{R}^d$, for some d; therefore, this formulation covers *nonparametric* inference, as well (and that is why we do not call θ a "parameter").

In our case, index θ is identified with the underlying true function, therefore, each possible f_* induces a different probability distribution according to which the observations are generated. Confidence regions constitute a classical form of statistical inference, when we aim at constructing sets which cover with high probability some target function of θ (DeGroot and Schervish, 2012). These sets are usually random as they are typically built using observations. In our case, we will build confidence regions for the ideal coefficient vector (equivalently, the ideal representation), which itself is a random element, as it depends on the sample.

Let γ be a random element (it corresponds to the available observations), let $g(\theta, \gamma)$ be some target function of θ (which can possibly also depend on the observations) and let $p \in [0, 1]$ be a target probability, also called significance level. A confidence region for $g(\theta, \gamma)$ is a random set, $C(p, \gamma) \subseteq \mathrm{range}(g)$, i.e., the codomain of function g. The following definition formalizes two important types of stochastic guarantees for confidence regions (Davies et al., 2009).

Definition 2 A confidence region $C(p, \gamma)$ for $g(\theta, \gamma)$ is called *exact*, if

$$\forall\, \theta \in \Theta : \mathbb{P}_\theta(\, g(\theta, \gamma) \in C(p, \gamma)\,) \;=\; p, \tag{12}$$

and it is called *honest*, if it satisfies $\forall\, \theta \in \Theta : \mathbb{P}_\theta(\, g(\theta, \gamma) \in C(p, \gamma)\,) \;\geq\; p$.

In our case, γ is basically[1] the sample of input-output pairs, $\mathcal{D}_n$; and the target object we aim at covering is $g(\theta, \gamma) = \alpha^*_\theta$, i.e., the (unique) ideal coefficient vector corresponding to the underlying true function (identified by θ) and the sample. Since the ideal coefficient vector uniquely determines the ideal representation (together with the inputs, which however we observe), it is enough to estimate the former. The main question of this paper is how can we construct exact or honest confidence regions for the ideal coefficient vector based on a finite sample without strong distributional assumptions on the statistical space.

Henceforth, we will treat θ (the underlying true function) fixed, and omit the θ indexes from the notations, to simplify the formulas. Therefore, instead of writing $\mathbb{P}_\theta$ or α^*_θ, we will simply use $\mathbb{P}$ or α^*. The results are of course valid for all θ.

Standard ways to construct confidence regions for kernel-based estimates typically either make *strong distributional assumptions*, like assuming Gaussian processes (Rasmussen and Williams, 2006), or resort to *asymptotic* results, such as Donsker-type theorems for Kolmogorov-Smirnov confidence bands. An alternative approach is to build on *Rademacher complexities*, which can provide non-asymptotic, distribution-free confidence bands (Giné and Nickl, 2015). Nevertheless, these regions are conservative (not exact) and are constructed independently of the applied kernel method. In contrast, our approach provides *exact*, *non-asymptotic*, *distribution-free* confidence sets for a *user-chosen* kernel estimate.

4 Non-Asymptotic, Distribution-Free Framework

This section presents the proposed framework to quantify the uncertainty of kernel-based estimates. It is inspired by and builds on recent results from finite-sample system identification, such as the SPS and DP methods (Campi and Weyer, 2005; Csáji et al., 2015; Csáji, 2016; Kolumbán, 2016; Carè et al., 2018). Novelties with respect to these approaches are, e.g., that our framework considers *nonparametric* regression and does not require the “true” function to be in the model class.

4.1 Distributional Invariance

The proposed method is distribution-free in the sense that it does not presuppose any parametric distribution about the noise vector ε. We only assume some mild regularity about the measurement noises, more precisely that their (joint) distribution is invariant with respect to a known group of transformations.

[1] We used the word “basically”, since there will also be some other random elements in the construction, e.g., for tie-breaking, and those should also constitute part of observation γ.

Definition 3 An $\mathbb{R}^n$-valued random vector v is *distributionally invariant* with respect to 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 transformation $G \in \mathcal{G}$, random vectors v and $G(v)$ have the same distribution.

The two most important examples of the above definition are as follows.

- If $\{\varepsilon_i\}$ are *exchangeable* random variables, then the (joint) distribution of the noise vector ε is invariant w.r.t. multiplications by permutation matrices (which are orthogonal and form a finite, thus compact, group).
- On the other hand, if $\{\varepsilon_i\}$ are independent, each having a (possibly different!) *symmetric* distribution about zero, then the (joint) distribution of ε is invariant w.r.t. multiplications by diagonal matrices having $+1$ or -1 as diagonal elements (which are also orthogonal, and form a finite group).

Both of these examples assume only mild regularities about the measurement noises: for example, it is a standard assumption in statistical learning theory that the sample is independent and identically distributed (i.i.d.) which immediately implies exchangeability (which is a more general concept than i.i.d.). But even this assumption can be omitted if we work with symmetric noises, which are widespread as most standard distributions in statistics are symmetric, such as Gauss, Laplace, Cauchy, Student's t, uniform, plus a large class of multimodal ones.

Note that for these examples no assumptions about other properties of the (noise) distributions are needed, e.g., they can be *heavy-tailed*, with even *infinite variance*, skewed, their expectations need not exist, hence, *no moment assumptions* are necessary. For the case of symmetric distributions, it is even allowed that the observations are affected by a noise where each ε_i has a *different* distribution.

4.2 Main Assumptions

Before the general construction of our method is explained, first, we highlight the core assumptions we apply. We also discuss their relevance and implications.

Assumption 1 *The kernel, $k(\cdot, \cdot)$, is strictly positive definite and all inputs, $\{x_i\}$, are distinct with probability one (in other words, $\forall\, i \neq j : \mathbb{P}(x_i = x_j) = 0$).*

As we discussed in Section 3.1, this assumption ensures that $\text{rank}(\text{K}_x) = n$ (a.s.), hence there *uniquely exists* an ideal representation (a.s.), whose unique ideal coefficient vector is denoted by α^*. The primary choices are *universal* kernels for which $\mathcal{H}$ is dense in the space of continuous functions on compact domains of $\mathcal{X}$.

Assumption 2 *The input vector x and the noise vector ε are independent.*

Assumption 2 implies that the measurement noises, $\{\varepsilon_i\}$, do not affect the inputs, $\{x_i\}$; for example, the system is not autoregressive. It is possible to extend our approach to dynamical systems, e.g., using similar ideas as in (Csáji et al., 2012; Csáji and Weyer, 2015; Csáji, 2016), but we leave the extension for future research. Note that Assumption 2 allows deterministic inputs, as a special case.

Assumption 3 *Noise ε is distributionally invariant w.r.t. a known group of transformations, $(\mathcal{G}, \circ)$, where each $G \in \mathcal{G}$ acts on $\mathbb{R}^n$ and $\circ$ is the function composition.*

Assumption 3 states that we known transformations that do not change the (joint) distribution of the measurement noises. As it was discussed in Section 4.1, symmetry and exchangeablity are two standard examples for which we know such group of transformations. Thus, if the noise vector is either exchangeable (e.g., it is i.i.d.), or symmetric, or both properties hold, then the theory applies. We also note that the suggested methodology is not limited to exchangeabe or symmetric noises, e.g., power defined noises constitute another example (Kolumbán, 2016).

Assumption 4 *The gradient, or a subgradient, of the objective w.r.t. α exists and it only depends on the output vector, y, through the residuals, i.e., there is $\bar{g}$,*

$$\nabla_{\alpha}\, g(f_{\alpha}, \mathcal{D}_n) \;=\; \bar{g}(x, \alpha, \widehat{\varepsilon}(x, y, \alpha)), \tag{13}$$

where the residuals w.r.t. the sample and the coefficients are defined as

$$\widehat{\varepsilon}(x, y, \alpha) \;\doteq\; y \,-\, \mathrm{K}_x \alpha. \tag{14}$$

For Assumption 4, it is enough if a subgradient is defined for each coefficient vector α, hence, e.g., the cases of ε-*insensitive* and *Huber* loss functions are also covered. Even in such cases (when we work with subderivaties), we still treat $\bar{g}$ as a vector-valued function and choose arbitrarily from the set of possible subgradients.

This requirement is also very mild as it is typically the case that the objective function is differentiable or convex and has subgradients (we will present several demonstrative examples in Section 5); furthermore, the objective typically only depends on y through the residuals, which immediately imply Assumption 4.

To see this assume that g is differentiable; then clearly, if the objective function can be written as $g(f_{\alpha}, \mathcal{D}_n) \,=\, g_0(x, \alpha, \widehat{\varepsilon}(x, y, \alpha))$ for some function g_0, then

$$\begin{aligned} \nabla_{\alpha}\, g(f_{\alpha}, \mathcal{D}_n) \;&=\; \nabla_{\alpha}(g_0(x, \alpha, y - \mathrm{K}_x \alpha))) \\ &=\; -\,\mathrm{K}_x\, (\nabla_{\alpha}\, g_0)\, (x, \alpha, y - \mathrm{K}_x \alpha)) \\ &=\; \bar{g}(x, \alpha, \widehat{\varepsilon}(x, y, \alpha)), \end{aligned} \tag{15}$$

where during the derivation we applied the chain rule, used the fact that matrix K_x is symmetric and the definition of the residuals, $\widehat{\varepsilon}(x, y, \alpha) \,=\, y - \mathrm{K}_x \alpha$.

4.3 Perturbed Gradients

At first, the proposed method can be understood as a *hypothesis testing* approach. Given coefficient vector $\alpha \in \mathbb{R}^n$ we test the null hypothesis $H_0 : \alpha \,=\, \alpha^*$, i.e., it is the ideal coefficient vector; against the alternative hypothesis $H_1 : \alpha \,\neq\, \alpha^*$. Under H_0, the residuals of f_{α} coincide with the "true" (unobserved) noise terms, since by definition (for ideal representations), we have

$$\begin{aligned} \widehat{\varepsilon}(x, y, \alpha^*) \;&=\; y \,-\, \mathrm{K}_x \alpha^* \\ &=\; [\, f_*(x_1) + \varepsilon_1, \dots, f_*(x_n) + \varepsilon_n \,]^{\mathrm{T}} \\ &-\; [\, f_*(x_1), \dots, f_*(x_n) \,]^{\mathrm{T}} \;=\; \varepsilon. \end{aligned} \tag{16}$$

Consequently, based on the group of invariant transformations, $\mathcal{G}$, we know that the (joint) distribution of the residuals does not change if we transform them by any $G \in \mathcal{G}$ (under H_0). Then, we can generate alternative realizations of the residuals, $\widehat{\varepsilon}(x, y, \alpha^*)$, by applying a random transformation $G \in \mathcal{G}$, and the resulting alternative realization, $G(\widehat{\varepsilon}(x, y, \alpha^*))$, will behave "similarly" (in the statistical sense) to the original residual vector (i.e., the true noise vector).

However, under H_1, if coefficient vector α does not define an ideal representation, $\widehat{\varepsilon}(x, y, \alpha)$, in general, will not coincide with the true noises. Therefore, the distributions of their randomly transformed variants will be distorted and will statistically *not* behave "similarly" to the original residuals.

Of course, we need a way to measure "similar behavior". Since we want to measure the uncertainty of a model constructed by using a certain objective function, we will measure similarity by recalculating (the magnitude of) its gradient (w.r.t. α) with the transformed residuals and apply a rank test (Good, 2005).

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 user-chosen hyper-parameter, as follows

$$Z_0(\alpha) \doteq \| \, \Psi(x) \, \bar{g}(x, \alpha, G_0(\widehat{\varepsilon}(x, y, \alpha))) \, \|^2, \tag{17}$$

$$Z_i(\alpha) \doteq \| \, \Psi(x) \, \bar{g}(x, \alpha, G_i(\widehat{\varepsilon}(x, y, \alpha))) \, \|^2, \tag{18}$$

for $i = 1, \dots, m-1$, where $\Psi(x)$ is some (possibly input dependent) positive definite weighting matrix, G_0 is the *identity* element of $\mathcal{G}$ (w.l.o.g. the identity transformation), and $\{G_i\}$ are i.i.d. random transformations from $\mathcal{G}$, sampled using the *uniform* distribution on $\mathcal{G}$. They are generated independently of the other random elements of the system, such as the input vector x and the noise vector ε.

For symmetric noises, transformation $G_i \in \mathcal{G}$ is basically a random $n \times n$ diagonal matrix whose diagonal elements are $+1$ or -1, each having $1/2$ probability to be selected, independently of the other elements of the diagonal.

On the other hand, for the case of exchangeable noise terms, each transformation $G_i \in \mathcal{G}$ is a randomly (uniformly) chosen $n \times n$ permutation matrix.

Weighting matrix $\Psi(x)$ is included in the construction to allow some additional flexibility, e.g., if we have some a priori information on the measurement noises. We will see an example for the special case of quadratic objectives in Section 4.6. In case no such information is available, $\Psi(x)$ can be chosen as identity.

We can observe that for the ideal coefficient vector α^*, we have

$$\begin{aligned} Z_0(\alpha^*) \;&=\; \| \, \Psi(x) \, \bar{g}(x, \alpha^*, \varepsilon) \, \|^2 \\ &\overset{d}{=}\; \| \, \Psi(x) \, \bar{g}(x, \alpha^*, G_i(\varepsilon)) \, \|^2 \\ &=\; Z_i(\alpha^*), \end{aligned} \tag{19}$$

for $i = 1, \dots, m-1$, where „$\overset{d}{=}$" denotes equality in distribution. Therefore, the $\{Z_i(\alpha^*)\}_{i=0}^{m-1}$ variables have the same (marginal) distribution, though, they are of course not independent. It can be shown, however, that they are *conditionally* independent, and therefore all of their possible orderings are equally likely, with possible tie-breakings, which can be used to measure *similar* behavior.

On the other hand, for $\alpha \neq \alpha^*$, this distributional equivalence does not hold, and we expect that if $\| \, \alpha - \alpha^* \, \|$ is large enough, the reference element $Z_0(\alpha)$ will dominate the perturbed elements, $\{Z_i(\alpha)\}_{i=1}^{m-1}$, with high probability, from which we can detect (statistically) that coefficient vector α is not the ideal one, $\alpha \neq \alpha^*$.

4.4 Normalized Ranks

Now, we make our argument, including possible tie-breakings, more precise by introducing the concept of normalized ranks. Formally, the *normalized rank* of the reference element, $Z_0(\alpha)$, among all $\{Z_i(\alpha)\}_{i=0}^{m-1}$ elements is defined as follows

$$\mathcal{R}(\alpha) \doteq \mathcal{R}_m(\alpha) \doteq \frac{1}{m}\bigg[1 + \sum_{i=1}^{m-1} \mathbb{I}\left(Z_0(\alpha) \prec_\pi Z_i(\alpha)\right)\bigg], \tag{20}$$

where $\mathbb{I}(\cdot)$ is an indicator function, namely, its value is 1 if its argument is true and 0 otherwise; $m \in \mathbb{N}$ is a user-chosen hyper-parameter; and binary relation "$\prec_\pi$" is the standard "$<$" with random tie-breaking (according to a fixed, pre-generated random order). More precisely, let π be a random (uniformly chosen) permutation of the set $\{0, \dots, m-1\}$. Then, given m arbitrary real numbers, $Z_0, \dots, Z_{m-1}$, we can construct a strict total order, denoted by "$\prec_\pi$", by defining $Z_k \prec_\pi Z_j$ if and only if $Z_k < Z_j$ or it both holds that $Z_k = Z_j$ and $\pi(k) < \pi(j)$.

4.5 Exact Confidence

Parameter m influences the resolution of the confidence probability we can achieve. Namely, a probability $p \in (0,1)$ is *admissible* if it can be written in the form of $p = 1 - q/m$, where q is an integer satisfying $0 < q < m$. On the other hand, since both m and q are (hyper) parameters, their values are user-chosen. Hence, every rational probability $p \in (0,1)$ is admissible, by choosing m and q appropriately. Then, a confidence set for an admissible probability $p = p(m,q)$ is

$$A_p \doteq \{\, \alpha : \mathcal{R}(\alpha) \leq p \,\} = \{\, \alpha : \mathcal{R}_m(\alpha) \leq 1 - q/m \,\}. \tag{21}$$

One of the main questions is: what kind of stochastic guarantees do such confidence regions have? The following theorem states that they are *exact*.

Theorem 2 *Under Assumptions 1, 2, 3 and 4, the coverage probability of the constructed confidence region with respect to the ideal coefficient vector α^* is*

$$\mathbb{P}\left(\, \alpha^* \in A_p \,\right) = p = 1 - \frac{q}{m}, \tag{22}$$

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

Proof Following (Csáji et al., 2015), the core idea is to show that variables

$$Z_0(\alpha^*),\; Z_1(\alpha^*),\; \dots,\; Z_{m-1}(\alpha^*) \tag{23}$$

are *uniformly ordered*, which means that each ordering of them, with respect to the strict total order $\prec_\pi$, has the same probability, that is $1/m!$, formally,

$$\mathbb{P}\left(\, Z_{i_0}(\alpha^*) \prec_\pi Z_{i_2}(\alpha^*) \prec_\pi \cdots \prec_\pi Z_{i_{m-1}}(\alpha^*) \,\right) = \frac{1}{m!}, \tag{24}$$

where $(i_0, i_1, \dots, i_{m-1})$ is an arbitrary permutation of $(0, 1, \dots, m-1)$. This ordering property is not obvious, since they are not independent, even though we already observed that they are identically distributed (for ideal coefficients).

By definition, $\alpha^* \in A_p$ if and only if $\mathcal{R}(\alpha^*) \leq 1 - {}^{q}/_{m}$, i.e., if the reference element, $Z_0(\alpha^*)$ takes one of the positions $1, \dots, m-q$ in the ordering of $\{Z_i(\alpha^*)\}_{i=0}^{m-1}$ variables, w.r.t. the strict total order $\prec_\pi$. Then, assuming they are uniformly ordered (yet to be shown), we know that $Z_0(\alpha^*)$ takes each position in the ordering with probability exactly $1/m$. Therefore, for $i \in \{1, \dots, m\}$, we have

$$\mathbb{P}\Big(\, \mathcal{R}(\alpha^*) \,=\, \frac{i}{m} \,\Big) \,=\, \frac{1}{m}, \tag{25}$$

from which it follows that $\mathbb{P}\big(\alpha^* \in A_p\big) \,=\, 1 - {}^{q}/_{m}$ by taking into account that events $\{\, \mathcal{R}(\alpha^*) = {}^{i}/_{m} \,\}$ and $\{\, \mathcal{R}(\alpha^*) = {}^{j}/_{m} \,\}$ are disjoint, if $i \neq j$.

In order to show that $\{Z_i(\alpha^*)\}_{i=0}^{m-1}$ are indeed uniformly ordered, we can apply Theorem 2.17 of (Kolumbán, 2016). Our proposed approach can be interpreted as a variant of a DP method, even though formally the DP "performance measures" can depend on the parameters, α, the inputs, x, and the perturbed outputs, $y^{(i)}$, but not directly on the perturbed residuals. Nevertheless, in our case, $y^{(i)}$ is

$$y^{(i)} \,\doteq\, f_\alpha(x) \,+\, G_i(\widehat{\varepsilon}(x, y, \alpha)), \tag{26}$$

where $f_\alpha(x) \doteq [\, f_\alpha(x_1), \dots, f_\alpha(x_n)]^\mathrm{T}$. Then, obviously we can compute the transformed residuals, $G_i(\widehat{\varepsilon}(x, y, \alpha))$, from α, x, and $y^{(i)}$ by using that $G_i(\widehat{\varepsilon}(x, y, \alpha)) = y^{(i)} - f_\alpha(x)$. Hence, the DP performance measure in our case is defined as

$$Z(\alpha, x, y^{(i)}) \,\doteq\, \|\, \Psi(x)\, \bar{g}(x, \alpha, y^{(i)} - f_\alpha(x))\, \|^2, \tag{27}$$

which now fits the DP framework. Our Assumption 4 ensures that this function is well-defined and, together with Assumption 2, it also guarantees that we do not need to compute $\{y^{(i)}\}$ to evaluate the perturbed functions. Our Assumption 3 directly states that the noise, ε, is invariant under a compact group of transformations, which is a requirement of Theorem 2.17, and we already observed that true errors coincide with the residuals of ideal representations, $\widehat{\varepsilon}(x, y, \alpha^*) \,=\, \varepsilon$. □

Theorem 2 shows that the confidence region contains the ideal coefficient vector *exactly* with probability p that statement is *non-asymptotically* guaranteed, despite the method is *distribution-free.* Since m and q are user-chosen (hyper-parameters), the confidence probability is *under our control.* The confidence level does not depend on the weighting matrix, but it influences the *shape* of the region. Ideally, it should be proportional to the square root of the covariance of the estimate.

4.6 Quadratic Objectives and Symmetric Noises

If we work with convex *quadratic* objectives, which have special importance for kernel methods (Hofmann et al., 2008), and assume independent and *symmetric* noises, we get the Sign-Perturbed Sums (SPS) method (Csáji et al., 2015) as a special case (using the inverse square root of the Hessian as a weighting matrix).

The SPS method uses the classical least-squares (LS) objective function,

$$g(f_\alpha, \mathcal{D}_n) \,=\, \|\, z \,-\, \Phi\alpha\, \|^2, \tag{28}$$

where z denotes the vector of outputs and Φ is the regressor matrix. Objective (28) can be seen the canonical form of many quadratic functions (cf. Section 5).

When using the SPS method, we make the following assumptions: the noise terms, $\{\varepsilon_i\}$, are independent and have symmetric distributions about zero; and the regressor matrix, Φ, has independent rows, it is skinny and full rank.

For SPS, the reference and the perturbed functions are defined as

$$Z_i(\alpha) \doteq \|\,(\Phi^{\mathrm{T}}\Phi)^{-1/2}\Phi^{\mathrm{T}}G_i(z - \Phi\alpha)\,\|^2, \tag{29}$$

for $i = 0, \dots, m-1$, where $G_i = \mathrm{diag}(\sigma_{i,1}, \dots, \sigma_{i,n})$, for $i \neq 0$, where random variables $\{\sigma_{i,j}\}$ are i.i.d. having *Rademacher* distribution, i.e., they take values $+1$ and -1 with probability $1/2$ each; and $G_0 = I_n$ is the identity matrix.

It is easy to see that (29) is a special case of construction (17)-(18), where z are the outputs and Φ is computed from the inputs. Besides being *exact*, the confidence regions of SPS have additional important properties, such as they are *star convex* with the LS estimate, $\widehat{\alpha}$, as a star center (Csáji et al., 2015). Moreover, they have *ellipsoidal outer approximations*, that is there are regions of the form

$$A_p^{\circ} \doteq \Big\{\alpha \in \mathbb{R}^n \,:\, (\alpha - \widehat{\alpha})^{\mathrm{T}}\frac{1}{n}\Phi^{\mathrm{T}}\Phi(\alpha - \widehat{\alpha}) \leq r\Big\}, \tag{30}$$

where $A_p \subseteq A_p^{\circ}$ and radius of the ellipsoid, r, can be computed (in polynomial time) by solving semi-definite programming problems (Csáji et al., 2015).

Hence, for quadratic problems, the obtained regions are star convex, thus connected, have ellipsoidal outer approximation, thus bounded. These properties ensure that it is easy to work with them. For example, using star convexity and boundedness, we can efficiently explore the region by knowing that every point of it can be reached from the given star center by a line segment inside the region. Moreover, the ellipsoidal outer approximation provides a compact representation.

5 Applications and Experiments

In this section, we show specific applications of the proposed uncertainty quantification (UQ) approach for typical kernel methods, such as LS-SVC, KRR, ε-SVR and KLASSO, in order to demonstrate the usage and the power of the framework.

We also present several numerical experiments to illustrate the family of confidence regions we get for various confidence levels. We always set hyper-parameter m to 100 in the experiments. The figures were constructed by Monte Carlo simulations, i.e., evaluating 1 000 000 random coefficients and drawing the graphs of their induced models with colors indicating their confidence levels.

5.1 Uncertainty Quantification for Least-Squares Support Vector Classification

We start with a classification problem and consider the Least-Squares Support Vector Classification (LS-SVC) method (Suykens and Vandewalle, 1999). LS-SVC under the Euclidean distance is known to be equivalent to hard-margin SVC using the Mahalanobis distance (Ye and Xiong, 2007). It has the advantage that it can be solved by a system of linear equations, in contrast to a quadratic problem.

We assume that $x_k \in \mathbb{R}^d$ and $y_k \in \{+1, -1\}$, for all $k \in \{1, \dots n\}$, as well as that the slack variables, i.e., the algebraic (signed) distances of the objects from the corresponding margins, are *independent* and distributed *symmetrically*, for the ideal representation; which we will identify with the best possible classifier.

For simplicity, we consider *linear* classification, that is models of the form

$$h_\alpha(x_k) \;\doteq\; \mathrm{sign}(\, w^{\mathrm{T}} x_k + b\,) \;=\; \mathrm{sign}(\, \alpha^{\mathrm{T}} \tilde{x}_k\,), \tag{31}$$

where x_k is an input vector, $\alpha \,\doteq\, [\, b,\, w^{\mathrm{T}}\,]^{\mathrm{T}}$ and $\tilde{x}_k \,\doteq\, [\,1,\, x_k^{\mathrm{T}}\,]^{\mathrm{T}}$.

The standard (primal) formulation of (soft-margin) LS-SVM classifcation is

$$\text{minimize} \quad \frac{1}{2}\, w^{\mathrm{T}} w \,+\, \lambda \sum_{k=1}^{n} \xi_k^2 \tag{32}$$

$$\text{subject to} \quad y_k(w^{\mathrm{T}} x_k + b) \,=\, 1 - \xi_k \tag{33}$$

for $k = 1, \dots, n$, where $\lambda > 0$ is fixed. Variables $\{\xi_i\}$ are called the *slack variables*. The convex quadratic problem above can be rewritten as minimizing

$$g(f_\alpha, \mathcal{D}_n) \;\doteq\; \frac{1}{2}\, \|\, B\alpha \,\|^2 \,+\, \lambda \,\|\, \mathbb{1}_n - y \odot (X\alpha) \,\|^2, \tag{34}$$

where $\mathbb{1}_n \in \mathbb{R}^n$ is the all-one vector, $\odot$ denotes the Hadamard (entrywise) product, $X \doteq [\, \tilde{x}_1, \dots, \tilde{x}_n\,]^{\mathrm{T}}$ and the role of matrix B is to remove the bias, b, from α, i.e., $B \,\doteq\, \mathrm{diag}(0, 1, \dots, 1)$. Note that the reformulated problem (34) is unconstrained.

Observe that the objective function, $g(f_\alpha, \mathcal{D}_n)$, can be further reformulated to take the canonical form of $\|\, z - \Phi\alpha \,\|^2$ by using the following Φ and z,

$$\Phi \;=\; \begin{bmatrix} \sqrt{\lambda}\, (y \mathbb{1}_d^{\mathrm{T}}) \odot X \\ (1/\sqrt{2})\, B \end{bmatrix}, \qquad \text{and} \qquad z \;=\; \begin{bmatrix} \sqrt{\lambda}\, \mathbb{1}_n \\ 0_d \end{bmatrix}, \tag{35}$$

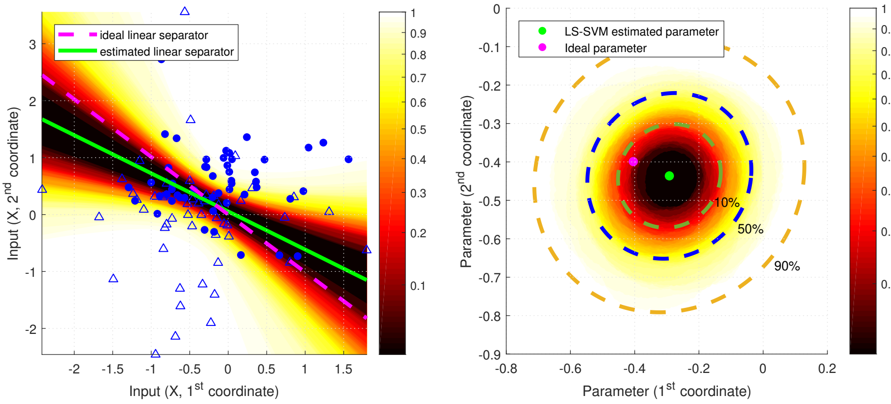
where $0_d \in \mathbb{R}^d$ is the all-zero vector. Then, we can apply SPS to the obtained (ordinary) LS formulation. However, we should be a careful with the transformations, as the new problem has some auxiliary output terms, the zero part of z, for which there are no slack variables. The residuals corresponding to that part are not even stochastic, therefore, the last d terms of the residual vector, $z - \Phi\alpha$, should not be perturbed. Consequently, the transformation matrices $\{G_i\}$ are defined as

$$G_i \;\doteq\; \begin{bmatrix} \bar{G}_i & 0 \\ 0 & I \end{bmatrix}, \tag{36}$$

for $i = 0, \dots, m-1$, where $\bar{G}_0 = I_n$ is the identity, and $\bar{G}_i \doteq \mathrm{diag}(\sigma_{i,1}, \dots, \sigma_{i,n})$, for $i \neq 0$, where $\{\sigma_{i,j}\}$ are i.i.d. Rademacher random variables, as before.

Then, (exact) confidence regions and (honest) ellipsoidal outer approximations can be constructed for the best linear classifier in the domain of coefficients by the SPS method, i.e., (29), with regressor matrix and output vector as defined in (35) and transformations as in (36). The regions will be centered around the LS-SVM classifier, i.e., for all (rational) $p \in (0, 1)$, the coefficients of LS-SVC are contained in A_p, assuming it is non-empty. As each coefficient vector uniquely identifies a classifier, the obtained region can be mapped to the model space, as well.

UQ for LS-SVC is illustrated in Figure 1. The observations were generated by adding Laplace noises to the coordinates of the corresponding class centers. The constructed confidence regions are shown both in the coefficient and model spaces, without the bias term, for simplicity. The possibility of constructing (honest) ellipsoidal outer approximations of the (exact) regions is also illustrated.



(a) UQ for LS-SVC, SPS, model space

(b) UQ for LS-SVC, SPS, coeffcient space

Fig. 1 Exact, non-asymptotic, distribution-free confidence regions for ideal RKHS representations. Parts (a) and (b) present UQ for Least-Squares Support Vector Classification (LS-SVC) with $\lambda = 0.1$ in the model and coefficient spaces, respectively. The ellipsoidal outer approximations of the regions having probabilities 10 %, 50 % and 90 % are also presented in the coefficient space. There were $n = 100$ observations, 50 for each class. The centers of the classes were $(0, 0.5)$ and $(-0.5, 0)$. For each observation i.i.d. Laplace noises were added to the coordinates of the corresponding centers. The parameters of the noises were $\mu = 0$ (location) and $b = 1/2$ (scale). The confidence level of each color can be interpreted by using the scale bars. The regions are increasing, i.e., $A_p \subseteq A_q$ if $p \leq q$, thus, only the smallest levels are shown.

5.2 Uncertainty Quantification for Kernel Ridge Regression

Our next example is Kernel Ridge Regression (KRR) which is a kernelized version of Tikhonov regularized LS (Shawe-Taylor and Cristianini, 2004). The KRR estimate minimizes a quadratic loss function with a Hilbert space norm regularizer,

$$\hat{f}_{\text{KRR}} \in \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} w_i (y_i - f(x_i))^2 + \lambda \, \|f\|_{\mathcal{H}}^2, \tag{37}$$

where $\lambda > 0$, $w_i > 0$, $i = 1, \dots, n$, are some a priori given (constant) weights. After using the representer theorem, the objective function can be rewritten as

$$\begin{aligned} g(f_\alpha, \mathcal{D}_n) &\doteq \frac{1}{n} \sum_{i=1}^{n} w_i (y_i - f_\alpha(x_i))^2 + \lambda \, \|f\|_{\mathcal{H}}^2 \\ &= \frac{1}{n} \, \|\, y - f_\alpha(x) \,\|_W^2 + \lambda \, \|f\|_{\mathcal{H}}^2 \\ &= \frac{1}{n} \, (y - \mathrm{K}_x \alpha)^{\mathrm{T}} W (y - \mathrm{K}_x \alpha) + \lambda \, \alpha^{\mathrm{T}} \mathrm{K}_x \alpha, \end{aligned} \tag{38}$$

where $f_\alpha(x) \doteq [\, f_\alpha(x_1), \dots, f_\alpha(x_n)]^{\mathrm{T}}$, $W \doteq \mathrm{diag}(w_1, \dots, w_n)$, and we used the reproducing property to replace the Hilbert space norm with a quadratic term.

We can reformulate (38) in the canonical form, $\|\, z - \Phi \alpha \,\|^2$, by using

$$\Phi = \begin{bmatrix} (1/\sqrt{n}) \, W^{\frac{1}{2}} \mathrm{K}_x \\ \sqrt{\lambda} \, \mathrm{K}_x^{\frac{1}{2}} \end{bmatrix}, \qquad \text{and} \qquad z = \begin{bmatrix} (1/\sqrt{n}) \, W^{\frac{1}{2}} y \\ 0_n \end{bmatrix}, \tag{39}$$

where $W^{\frac{1}{2}}$ and $\mathrm{K}_x^{\frac{1}{2}}$ denote the square roots of matrices W and K_x, respectively. Note that the square roots exist as these matrices are positive semidefinite.

Then, assuming symmetric and independent measurement noises, formula (29), with regressor matrix and output vector defined by (39), can be applied to build confidence regions. As in the case of LS-SVM classifier, the canonical reformulation also contains some auxiliary terms, the zero part of z, for which there are no real noise terms, therefore, they should not be perturbed. Thus, we should again use the transformations defined by (36) to get guaranteed confidence regions.

Experiments illustrating the family of (exact, non-asymptotic, distribution-free) confidence regions of KRR with Gaussian kernels and Laplacian measurement noises, and comparing the results with that of support vector regression, are shown in Figure 2. The discussion of the comparison is located in Section 5.3.

5.3 Uncertainty Quantification for Support Vector Regression

The previous examples were quadratic and therefore, for symmetric noises, their uncertainty could be quantified with SPS. This may be no more true if we change the applied norms. In this section we study support vector regression, particularly, ε-SVR (Hofmann et al., 2008; Schölkopf and Smola, 2001; Steinwart and Christmann, 2008). A well-known advantage of ε-SVR, for example, over KRR, is that it ensures sparse representations through the ε-insensitive loss function. In order to avoid confusion with the true noise vector, ε, we denote the tolerance parameter of the loss function by $\bar{\varepsilon}$. The primal objective function of ε-SVR is defined as

$$h(f, \mathcal{D}_n) \doteq \frac{1}{2} \| f \|_{\mathcal{H}}^2 + \frac{c}{n} \sum_{k=1}^{n} \max\{ 0, | \langle f, \phi(x_k) \rangle_{\mathcal{H}} - y_k | - \bar{\varepsilon} \}, \tag{40}$$

where $f \in \mathcal{H}$, $c > 0$, and $\phi(z) \doteq k(z, \cdot)$ is the *feature map*. Function (40) can be reformulated by applying slack variables, then using standard arguments based on the Lagrangian and the Karush–Kuhn–Tucker (KKT) conditions, we arrive at the Wolfe dual of ε-SVR (Schölkopf and Smola, 2001), where we have to maximize

$$\begin{aligned} g(f_{\alpha^+,\alpha^-}, \mathcal{D}_n) \;=\; & y^{\mathrm{T}}(\alpha^+ - \alpha^-) - \\ & - \frac{1}{2}(\alpha^+ - \alpha^-)^{\mathrm{T}} \mathrm{K}_x\, (\alpha^+ - \alpha^-) - \bar{\varepsilon}\, (\alpha^+ + \alpha^-)^{\mathrm{T}} \mathbb{1}, \end{aligned} \tag{41}$$

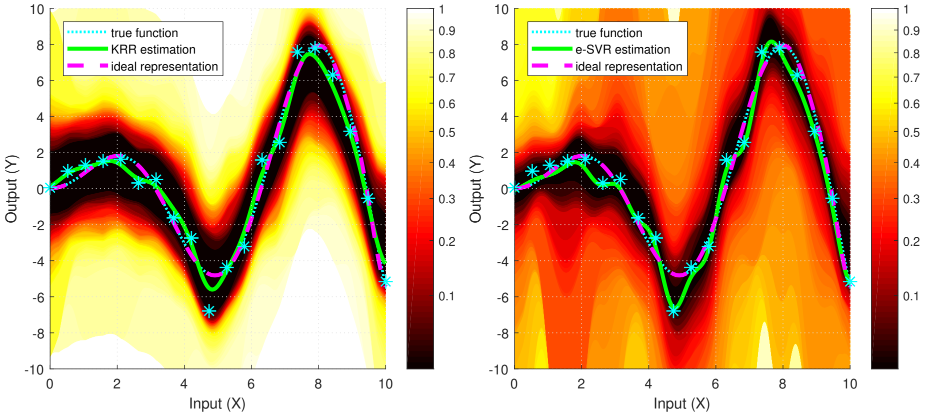
subject to the (linear) constraints: $\alpha^+, \alpha^- \in [\,0, c/n\,]^n$ and $(\alpha^+ - \alpha^-)^{\mathrm{T}} \mathbb{1} = 0$. One can work directly with the quadratic dual objective, but then the confidence region will be constructed for α^+, α^-. Since, $\alpha = \alpha^+ - \alpha^-$, the region could be mapped to a confidence region in the space of coefficient vectors. Alternatively, one can reformulate (41) directly for coefficient vector α as

$$g(f_\alpha, \mathcal{D}_n) \;=\; y^{\mathrm{T}} \alpha - \frac{1}{2}\, \alpha^{\mathrm{T}} \mathrm{K}_x \alpha - \bar{\varepsilon}\, \|\alpha\|_1, \tag{42}$$

where $\| \cdot \|_1$ is the 1-norm. A subgradient of (42) w.r.t. α is given by

$$\nabla_\alpha\, g(f_\alpha, \mathcal{D}_n) \;=\; y \,-\, \mathrm{K}_x \alpha \,-\, \bar{\varepsilon}\, \mathrm{sign}(\alpha), \tag{43}$$

where $\mathrm{sign}(\cdot)$ denotes the signum function and it is understood component-wise.



(a) UQ for KRR ($\lambda = 0.1$), SPS

(b) UQ for ε-SVR ($\bar{\varepsilon} = 0.2$), sign-changes

Fig. 2 Exact, non-asymptotic, distribution-free confidence regions for ideal RKHS representations. Parts (a) and (b) show UQ for Kernel Ridge Regression (KRR) with $\lambda = 0.1$ and ε-Support Vector Regression (ε-SVR) with $c = 250$ and $\bar{\varepsilon} = 0.2$, respectively. The same data was used for both regression problems, namely, the true function was $f_*(x) = x\,\sin(x)$, there were $n = 20$ observations having i.i.d. Laplace noise with parameters $\mu = 0$ (location) and $b = 1/2$ (scale), and Gaussian kernels were applied with $\sigma = 1/2$. Part (a) was built by the Sign-Perturbed Sums (SPS) method, (29), and formula (44) was used with sign-change matrices for part (b). The confidence level of each color can be interpreted by using the scale bars. The regions are increasing, i.e., $A_p \subseteq A_q$ if $p \leq q$, thus, only the smallest levels are shown.

Then, building on the subgradient of the dual objective, i.e., (43), reference and perturbed evaluation functions can be defined, for $i = 0, \dots, m-1$, as

$$Z_i(\alpha) \doteq \|\, G_i\,(y - \mathrm{K}_x\,\alpha) \;-\; \bar{\varepsilon}\,\mathrm{sign}(\alpha)\,\|^2\,, \tag{44}$$

where G_0 is the identity matrix and G_i is a (uniformly chosen) element of the applied compact transformation group, such as a diagonal matrix with ±1 entries, for symmetric noises (or permutation matrices for exchangeable noises, etc.).

A numerical experiment illustrating the obtained family of confidence regions of the ε-SVR estimate for various significance levels is shown in Figure 2.

The same data sample was used for all regression models, to allow their comparison. The noise affecting the observations was Laplacian, thus heavy-tailed. Since the coefficient space is high-dimensional, and there is a one-to-one correspondence between coefficient vectors and kernel models, the confidence regions are mapped and shown in the model space, i.e., in the space of RKHS functions.

Note that it is meaningful to plot the confidence regions even for unknown input values, because the confidence regions are built for the ideal representation, which belongs to the chosen RKHS, unlike the underlying true function.

We can observe that the uncertainty of ε-SVR was higher than that of KRR, which can be explained as the price of using ε-insensitive loss. As the experiments with KLASSO show (cf. Figure 3), the higher uncertainty of ε-SVR is not simply a consequence of sparse representations, as KLASSO also ensures sparsity. Naturally, the confidence regions are also influenced by the specific choice of hyper-parameters which should be taken into account when the confidence regions are compared.

5.4 Uncertainty Quantification for Kernelized LASSO

Our last example covers the LASSO (least absolute shrinkage and selection operator) method, which ensures sparsity via 1-norm regularization. Let us consider the kernelized version of LASSO with objective (Wang et al., 2007):

$$g(f_\alpha, \mathcal{D}_n) \doteq \frac{1}{2} \| y - \mathrm{K}_x \alpha \|^2 + \lambda \| \alpha \|_1, \tag{45}$$

were $\| \cdot \|_1$ is the L1 (or Manhattan) norm. Though, function (45) cannot be written as $\| z - \Phi\alpha \|^2$, the proposed framework, i.e., construction (17)-(18), can still be applied. A sub-gradient of the KLASSO objective (45) is given by

$$\nabla_\alpha g(f_\alpha, \mathcal{D}_n) = \mathrm{K}_x(\mathrm{K}_x \alpha - y) + \lambda \operatorname{sign}(\alpha), \tag{46}$$

where the sign$(\cdot)$ function is applied component-wise. Then, using the construction of (17)-(18), the reference and perturbed functions can be defined as

$$Z_0(\alpha) \doteq \| \mathrm{K}_x (\mathrm{K}_x \alpha - y) + \lambda \operatorname{sign}(\alpha) \|^2, \tag{47}$$

$$Z_i(\alpha) \doteq \| \mathrm{K}_x G_i \, (\mathrm{K}_x \alpha - y) + \lambda \operatorname{sign}(\alpha) \|^2, \tag{48}$$

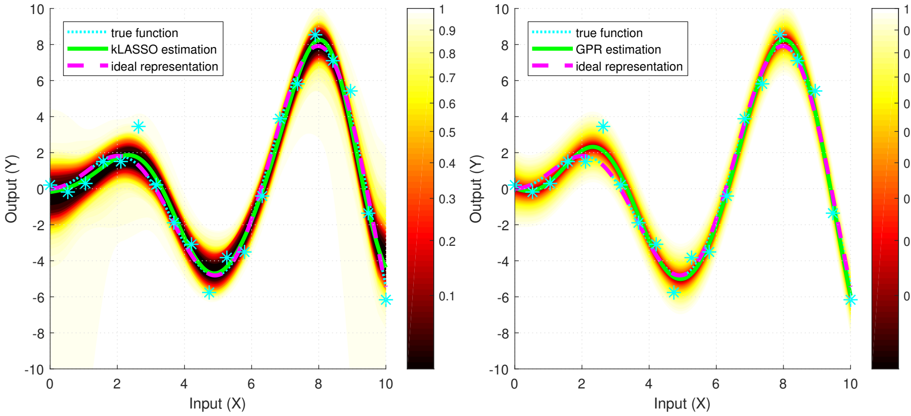
were $\{G_i\}$ are from a suitable transformation group, e.g., diagonal matrices with Rademacher random variables as diagonal elements for symmetric noises.

Numerical experiments illustrating the confidence regions we get for KLASSO are presented in Figure 3. The figure also presents the confidence regions constructed by applying the standard Gaussian Process (GP) regression with estimated parameters. Note that the GP confidence regions are only approximate, namely, they do not come with strict finite-sample guarantees unless the noise is indeed Gaussian. Moreover, during our experiment the noise had a Laplace distribution, which has a heavier tail than Gaussians, therefore even if the true covariance of the noise was known, the confidence regions of GP regression would underestimate the uncertainty of the estimate (would be too optimistic), while the confidence regions of our framework are always non-conservative, independently of the particular distribution of the noise, assuming it has the necessary invariance.

Also note that for our method the noises can even have different (marginal) distributions for each input. Therefore, even though the confidence regions generated by GP are smaller than the ones our framework produces, the GP regions are imprecise and underestimate the uncertainty of the model, while ours come with strict finite-sample guarantees for a broad class of noises (e.g., symmetric ones).

6 Conclusions

In this paper we addressed the problem of quantifying the *uncertainty* of kernel estimates by using minimal distributional assumptions. The main aim was to measure the uncertainty of finding the (noise-free) *ideal representation* of the underlying (hidden) function at the available inputs. By building on recent developments in finite-sample system identification, we proposed a method that delivers *exact*, *distribution-free* confidence regions with strong *finite-sample guarantees*, based on the knowledge of some mild regularity of the measurement noises. The standard



(a) UQ for KLASSO with Gaussian kernel (b) UQ with Gaussian Process Regression

Fig. 3 Exact, non-asymptotic, distribution-free confidence regions for ideal RKHS representations obtained using our framework and approximate confidence regions obtained by Gaussian Process (GP) regression (Rasmussen and Williams, 2006). Part (a) shows UQ for Kernelized LASSO with $\lambda = 1$, and part (b) shows UQ with GP. The applied transformations were sign-change matrices. The same data was used for both regression problems, namely, the true function was $f_*(x) = x \sin(x)$, there were $n = 20$ observations having i.i.d. Laplace noise with parameters $\mu = 0$ (location) and $b = 1/2$ (scale), and Gaussian kernels were applied with $\sigma = 1$. The confidence level of each color can be interpreted by using the scale bars. The confidence regions are increasing, i.e., $A_p \subseteq A_q$ if $p \leq q$, therefore, only the smallest levels are shown.

examples of such regularities are *exchangeable* or *symmetric* noise terms. Note that either of these properties in itself is sufficient for the theory to be applicable.

The needed statistical assumptions are very mild, as for example, no particular (parametric) family of distributions was assumed, *no moment assumptions* were made (the noises can be heavy-tailed, and may even have infinite variances); moreover, for the case of symmetric noises, it is allowed that each noise term affecting the observations has a different distribution, i.e., the noise can be *nonstationary.*

The core idea of the approach is to evaluate the uncertainty of the estimate by *perturbing the residuals* in the *gradient* of the objective function. The norms of the (potentially weighted) perturbed gradients are then compared to that of the unperturbed one, and a *rank test* is applied for the construction of the region.

The proposed method was also demonstrated on *specific examples* of kernel methods. Particularly, we showed how to construct exact, non-asymptotic, distribution-free confidence regions for least-squares support vector classification, kernel ridge regression, support vector regression and kernelized LASSO.

Several *numerical experiments* were presented, as well, demonstrating that the method provides meaningful regions even for *heavy-tailed* (e.g., Laplacian) noises. The figures illustrate whole families of confidence regions for various standard kernel estimates. Ellipsoidal outer approximations are also shown for LS-SVC. Additionally, the method was compared to Gaussian Process (GP) regression, and it was found that although the (approximate) GP confidence regions are smaller in general than our (exact) confidence sets, but the GP regions are typically imprecise and they underestimate the real uncertainty, e.g., if the noises are heavy-tailed.

Our approach to build non-asymptotic, distribution-free, non-conservative confidence regions for kernel methods can be a promising alternative to existing constructions, which arch-typically either build on strong distributional assumptions or on asymptotic theories or only bound the error between the true and empirical risks. As our approach explicitly builds on the constructions of the underlying kernel methods, it can provide *new insights* on how the specific methods influence the uncertainty of the estimates, and therefore, besides being vital for risk management, it also has the potential to inspire refinements or new constructions.

There are several open questions about the framework which can facilitate future research directions. For example, finding efficient *outer-approximations* for cases when the objective function is not convex quadratic should be addressed. Also the *consistency* of the method should be studied to see whether the uncertainty decreases as the sample size tends to infinity. Finally, it would be interesting, as well, to extend the method to (stochastic) *dynamical systems* and to formally analyze the *size and shape* of the constructed regions in a finite-sample setting.

Acknowledgements This research was supported by the National Research, Development and Innovation Office (NKFIH), grant numbers ED_18-2-2018-0006, 2018-1.2.1-NKP-00008 and KH_17 125698. The authors are grateful to Algo Carè for the valuable discussions.

A. Additional Numerical Experiments

In this appendix we provide additional numerical experiments supporting the presented framework. The effects of various *measurement noises*, *kernel functions* and *sample sizes* on the obtained (families of) exact, non-asymptotic, distribution-free confidence regions were studied. The true function was always $f_*(x) = x \sin(x)$ and the inputs were chosen equidistantly from $[\,0, 10\,]$. The regions were evaluated by the same methodology (Monte Carlo simulations) as in Section 5.

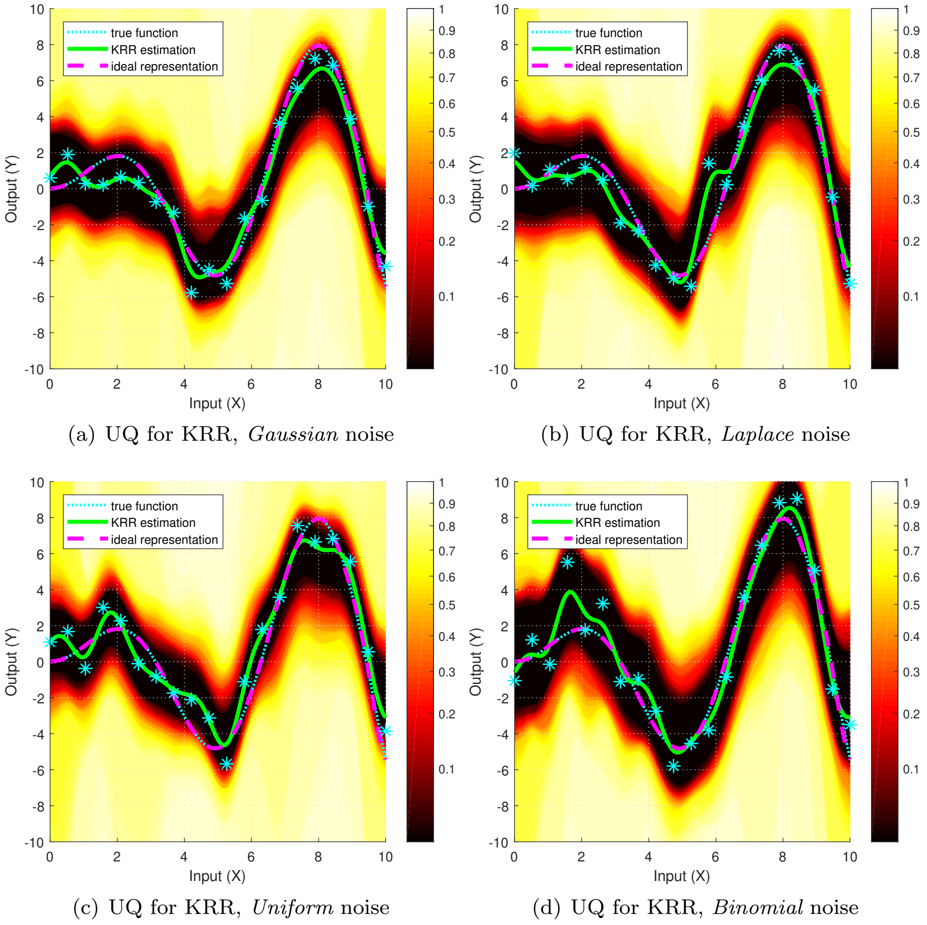
A.1 Various Noise Distributions

First, we investigated how the distribution of the noise affects the regions. Particularly, we applied *Gaussian*, *Laplacian*, *Uniform* and *Binomial* noises on the outputs of the true function and built the regions for Kernel Ridge Regression (KRR). All noises had zero mean (for the Binomial case the theoretical mean was subtracted from the generated noises), and the parameters of the distributions were set in a way to ensure that all of their variances were the same (i.e., one).

Figure 4 illustrates the obtained families of confidence sets. It can be observed that their shapes and sizes show only small fluctuations indicating that the particular choice of the distribution has a limited effect on the confidence regions (assuming it has zero expectation and we keep the variance of the noise fixed).

A.2 Different Kernel Functions

Next, the effect of the applied kernel was studied. Figure 5 illustrates UQ for kernelized LASSO with *Gaussian*, *Laplacian*, *truncated parabolic* ($k(x, y) = \max\{1 -$



(a) UQ for KRR, *Gaussian* noise

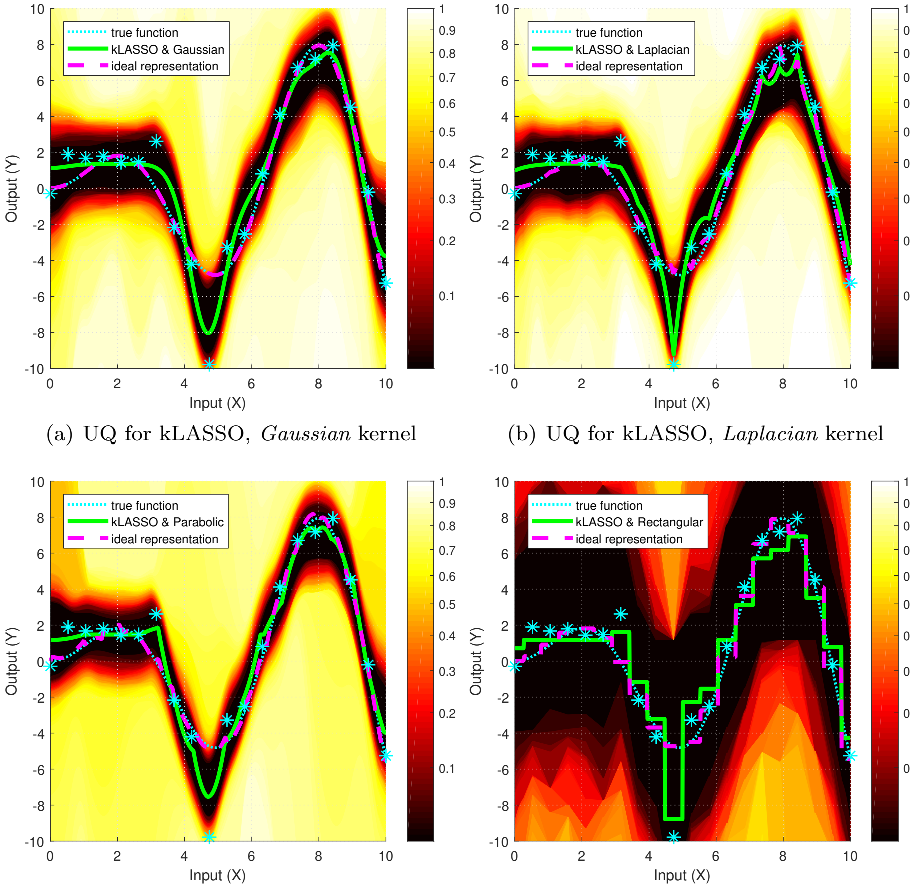
(b) UQ for KRR, *Laplace* noise

(c) UQ for KRR, *Uniform* noise

(d) UQ for KRR, *Binomial* noise

Fig. 4 Exact, non-asymptotic, distribution-free confidence regions for ideal representations w.r.t. *various noise distributions*. The figure shows UQ for Kernel Ridge Regression (KRR) with $\lambda = 0.1$ and Gaussian kernels with $\sigma = 1/2$. Parts (a), (b), (c) and (d) demonstrate the obtained family of confidence regions for i.i.d. Gaussian, Laplace, Uniform and Binomial noises, respectively. The parameters of all distributions were set to ensure that each of them has zero mean and unit variance. For the Binomial case, the "number of trials" parameter was 20, and so the "success probability" p was set to satisfy $20p(1-p) = 1$ (thus, $p \approx 0.052786$). Then, from each Binomial observation $20p$ was subtracted to ensure zero mean. In all cases $n = 20$ outputs were measured at equidistant inputs. The Sign-Perturbed Sums (SPS) method was applied to construct the regions, hence, the applied transformations were sign-changes. The confidence levels can be interpreted by using the scale bars. The regions are increasing, i.e., $A_p \subseteq A_q$ if $p \leq q$, therefore, only the smallest levels are shown.

$c\|x-y\|^2,\, 0\}$) and *rectangular kernels* ($k(x,y) = \mathbb{I}(\|x-y\| \leq c)$, where the noises were Laplacian. The results show that the choice of the kernel has a significant effect on both the obtained point-estimate (regression model) and the corresponding confidence sets, e.g., the Laplacian kernel was more sensitive to outliers and the regions for the rectangular kernel were much larger than the other ones.



(a) UQ for kLASSO, *Gaussian* kernel

(b) UQ for kLASSO, *Laplacian* kernel

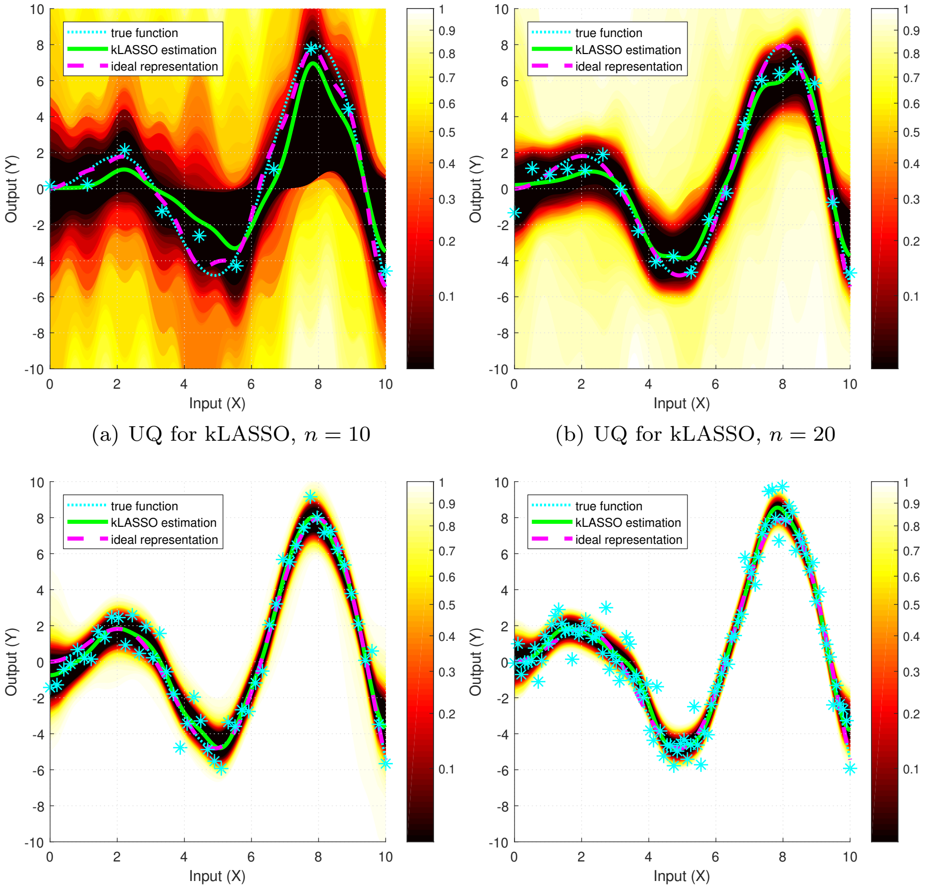
(c) UQ for kLASSO, *Parabolic* kernel

(d) UQ for kLASSO, *Rectangular* kernel

Fig. 5 Exact, non-asymptotic, distribution-free confidence regions for ideal representations w.r.t. *different kernel functions*. The figure shows UQ for kernelized LASSO with $\lambda = 1$. There were $n = 20$ observations having i.i.d. Laplace noises with parameters $\mu = 0$ (location) and $b = 1/2$ (scale). Parts (a), (b), (c) and (d) demonstrate the obtained family of confidence regions when using Gaussian, Laplacian, truncated parabolic and rectangular kernels, respectively. For the Gaussian and Laplacian kernels $\sigma = 1/2$, for the truncated parabolic kernel $c = 1$, and for the rectangular kernel $c = 1/38$. The same data was used for all regression problems, and the applied transformations were sign-changes. Observe that the Laplacian kernel was more sensitive to the outlier between 4 and 5. The obtained regions for the rectangular kernel are much larger than the other regions, indicating a high uncertainty of such an overly localized approach. The confidence levels can be interpreted by using the scale bars. The regions are increasing, i.e., $A_p \subseteq A_q$ if $p \leq q$, therefore, only the smallest levels are shown.

A.3 Increasing the Sample Size

Finally, we have experimented with kernelized LASSO to see how increasing of the sample size affects the obtained confidence regions. The measurement noises were Laplacian (hence heaviy-talied), and the applied *sample sizes* were $n = 10, 20, 50,$



(a) UQ for kLASSO, $n = 10$

(b) UQ for kLASSO, $n = 20$

(c) UQ for kLASSO, $n = 50$

(d) UQ for kLASSO, $n = 100$

Fig. 6 Exact, non-asymptotic, distribution-free confidence regions for ideal representations w.r.t. *increasing sample sizes.* The figure shows UQ for kernelized LASSO with $\lambda = 1$ and using Gaussian kernels with $\sigma = 1/2$. The observations had i.i.d. Laplace noises with parameters $\mu = 0$ (location) and $b = 1/2$ (scale). Parts (a), (b), (c) and (d) demonstrate the obtained family of confidence regions when using samples of size $n = 10, 20, 50$, and 100, respectively. The applied transformations were sign-changes. Observe that the confidence regions shrink around the ideal representations, despite the number of coefficients also increases with the sample size. This is indicative of the phenomenon that the regions have a consistency property. This may be especially true if we apply a universal kernel, such as the Gaussian one, for which the ideal representations can approximate arbitrary well any continuous functions on a compact domain. The confidence levels can be interpreted by using the scale bars. The regions are increasing, i.e., $A_p \subseteq A_q$ if $p \leq q$, therefore, only the smallest levels are shown.

and 100. The results are shown in Figure 6 and are indicative of the phenomenon that the confidence regions, and hence the uncertainties, shrink as the sample size tends to infinity, even though the number of coefficients increases with the sample size. This experiment supports that the approach is “consistent”, nevertheless, we leave the theoretical investigation of this phenomenon for further study.

References

Argyriou A., Dinuzzo F. (2014) A unifying view of representer theorems. In: International Conference on Machine Learning (ICML), pp. 748–756

Aronszajn N. (1950) Theory of reproducing kernels. Transactions of the American Mathematical Society 68(3):337–404

Campi M. C., Weyer E. (2005) Guaranteed non-asymptotic confidence regions in system identification. Automatica 41(10):1751–1764

Carè A., Csáji B. Cs., Campi M., Weyer E. (2018) Finite-sample system identification: An overview and a new correlation method. IEEE Control Systems Letters 2(1):61 – 66

Csáji B. Cs. (2016) Score permutation based finite sample inference for generalized autoregressive conditional heteroskedasticity (GARCH) models. In: 19th International Conference on Artificial Intelligence and Statistics (AISTATS), Cadiz, Spain, pp. 296–304

Csáji B. Cs., Weyer E. (2015) Closed-loop applicability of the Sign-Perturbed Sums method. In: 54th IEEE Conference on Decision and Control (CDC), IEEE, pp. 1441–1446

Csáji B. Cs., Campi M. C., Weyer E. (2012) Sign-Perturbed Sums (SPS): A method for constructing exact finite-sample confidence regions for general linear systems. In: 51st IEEE Conference on Decision and Control, Maui, Hawaii, pp. 7321–7326

Csáji B. Cs., Campi M. C., Weyer E. (2015) Sign-Perturbed Sums: A new system identification approach for constructing exact non-asymptotic confidence regions in linear regression models. IEEE Transactions on Signal Processing 63:169–181

Davies P. L., Kovac A., Meise M. (2009) Nonparametric regression, confidence regions and regularization. The Annals of Statistics pp. 2597–2625

DeGroot M. H., Schervish M. J. (2012) Probability and Statistics, 4th edn. Pearson Education

Efron B., Tibshirani R. J. (1994) An Introduction to the Bootstrap. CRC press

Giné E., Nickl R. (2015) Mathematical Foundations of Infinite-Dimensional Statistical Models, vol 40. Cambridge University Press

Good P. (2005) Permutation, Parametric, and Bootstrap Tests of Hypotheses. Springer

Hofmann T., Schölkopf B., Smola A. J. (2008) Kernel methods in machine learning. The Annals of Statistics 36:1171–1220

Kimeldorf G., Wahba G. (1971) Some results on tchebycheffian spline functions. Journal of Mathematical Analysis and Applications 33(1):82–95

Kolumbán S. (2016) System identification in highly non-informative environment. PhD thesis, Budapest University of Technology and Economics, Hungary, and Vrije Univesiteit Brussels, Belgium

Li K.-C. (1989) Honest confidence regions for nonparametric regression. The Annals of Statistics pp. 1001–1008

Pillonetto G., Dinuzzo F., Chen T., De Nicolao G., Ljung L. (2014) Kernel methods in system identification, machine learning and function estimation: A survey. Automatica 50(3):657–682

Rasmussen C. E., Williams C. K. I. (2006) Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning, MIT Press, Cambridge, MA

Schölkopf B., Smola A. J. (2001) Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT Press

Schölkopf B., Herbrich R., Smola A. J. (2001) A generalized representer theorem. In: Annual Conference on Learning Theory (COLT), Springer, pp. 416–426

Shawe-Taylor J., Cristianini N. (2004) Kernel Methods for Pattern Analysis. Cambridge University Press

Steinwart I., Christmann A. (2008) Support Vector Machines. Springer Science & Business Media

Suykens J. A. K., Vandewalle J. (1999) Least squares support vector machine classifiers. Neural Processing Letters 9(3):293–300

Vapnik V. N. (1998) Statistical Learning Theory. Wiley-Interscience

Vovk V., Gammerman A., Shafer G. (2005) Algorithmic Learning in a Random World. Springer Science & Business Media

Wang G., Yeung D.-Y., Lochovsky F. H. (2007) The kernel path in kernelized LASSO. In: Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 580–587

Ye J., Xiong T. (2007) SVM versus least squares SVM. In: 11th International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 644–651

Yu Y., Cheng H., Schuurmans D., Szepesvári Cs. (2013) Characterizing the representer theorem. In: International Conference on Machine Learning, pp. 570–578