Uncertainty Quantification for Kernel Methods

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, kernelized LASSO and ε-SVR.

Index Terms—kernel methods, confidence regions, nonparametric regression, classification, support vector machines

I. INTRODUCTION

KERNAL methods build on the fundamental concept of Reproducing Kernel Hilbert Spaces [1], [2] and are widely used in machine learning [3], [4]. One of the reasons of their popularity is the representer theorem [5], [6] 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 [7], [8], rooted in statistical learning theory [9], are highly successful 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 [10] or exponential families [4] 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 [2], but such approaches are usually lack of 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 [11], [12], more specifically, we build on the Sign-Perturbed Sums (SPS) algorithm [13] and its generalizations, the Data Perturbation (DP) methods [14].

B. Cs. Csáji was supported by the National Research, Development and Innovation Office (NKFIH), Hungary, KH_17 125698, and the János Bolyai Research Fellowship, Hungarian Academy of Sciences, BO/00217/16/6.

B. Cs. Csáji and K. B. Kis are with MTA SZTAKI: Institute for Computer Science and Control, Hungarian Academy of Sciences, Kende utca 13-17, H-1111 Budapest, Hungary. balazs.csaji@sztaki.mta.hu, krisztian.kis@sztaki.mta.hu

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, e.g., we do not want to assume that this function belongs to the Hilbert space in which we search our estimate, we take a “honest” approach [15] 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 parametrized distributions), it has strong finite-sample guarantees. We argue that, for all ideal representations, the constructed region contains the representation exactly with a user-chosen probability. In case the noises are independent and symmetric about zero, and the objective function of the used kernel method is 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 SVMs and kernelized LASSO.

Our approach has some similarities to bootstrap [16] and conformal prediction [17]. 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.

II. REPRODUCING KERNEL HILBERT SPACES

A Hilbert space, \( \mathcal{H} \), of functions \( f : \mathcal{X} \rightarrow \mathbb{R} \), with inner product \( \langle \cdot, \cdot \rangle_{\mathcal{H}} \), is called a Reproducing Kernel Hilbert Space (RKHS) [2], if the point evaluation functional \( \delta_z : f \mapsto f(z) \) is continuous (or equivalently bounded) for all \( z \in \mathcal{X} \), at any \( f \in \mathcal{H} \). Then, by using the Riesz representation theorem, one can construct a kernel, \( k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R} \), having the reproducing property, that is for all \( z \in \mathcal{X} \) and \( f \in \mathcal{H} \), we have

\[
\langle k(\cdot, z), f \rangle_{\mathcal{H}} = f(z),
\]

In particular, we have \( k(z, s) = \langle k(\cdot, z), k(\cdot, s) \rangle_{\mathcal{H}} \). A kernel is thus both symmetric and positive-definite; we know from the Moore-Aronszajn theorem that the converse is also true: for every positive-definite function there is a unique RKHS [1].

Typical kernels include, for example, 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 Euclidean inner product.
By a data sample, \( D_n \), we mean a finite set of input-output measurements, \((x_1, y_1), \ldots, (x_n, y_n)\) \(\in \mathcal{X} \times \mathbb{R} \), with \( \mathcal{X} \neq \emptyset \). We also introduce the notations \( x = (x_1, \ldots, x_n)^T \in \mathcal{X} \) and \( y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n \). The Gram matrix of \( k(\cdot, \cdot) \), w.r.t. input \( x \), is denoted by \( K_x \in \mathbb{R}^{n \times n} \), where \( [K_x]_{ij} = k(x_i, x_j) \).

A kernel is called strictly positive definite if its Gram matrix, \( K_x \), is (strictly) positive definite for distinct \( \{x_i\} \) [4].

One of the fundamental reasons for the successes of kernel methods is the so-called representer theorem, originally given by [5], but the generalization presented here is due to [6].

**Theorem 1.** Suppose we are given a sample, \( D_n \), a positive-definite kernel \( k(\cdot, \cdot) \), an associated RKHS with a norm \( \| \cdot \|_\mathcal{H} \) induced by its inner product \( \langle \cdot, \cdot \rangle_\mathcal{H} \), and a class of functions

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

then, for any monotonically increasing regularization function, \( \Omega : [0, \infty) \to [0, \infty) \), and an arbitrary loss function \( L : (\mathcal{X} \times \mathbb{R}^2)^n \to \mathbb{R} \cup \{\infty\} \), the objective (risk functional) \( g(f, D_n) = \Omega\left(\left|\left| f - f_{0}\right|\right|_\mathcal{H}\right) \), has a minimizer admitting the following representation

\[
f_{0}(z) = \sum_{i=1}^{n} \alpha_i k(z_i, z),
\]

where \( \alpha = (\alpha_1, \ldots, \alpha_n)^T \in \mathbb{R}^n \) is the vector of coefficients. If the regularizer, \( \Omega \), is strictly monotonically increasing, then each minimizer admits a representation having form (3).

The theorem can be extended with a bias term [7], in which case if the solution exists, it also contains a multiple of the bias term. For further generalizations, see [18], [19].

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 optimization problem whose dimension does not exceed the size of the data sample we have, i.e., \( n \).

If the data is noisy, the obtained estimate is a random function and it is of natural interest to study its distribution, e.g., to evaluate its uncertainty or to test hypotheses about the system.

### III. **CONFIDENCE REGIONS FOR KERNEL METHODS**

Now, we turn our attention to a stochastic variant of the problem. There are several advantages for taking a statistical point of view on kernel methods, including conditional modeling, dealing with structured responses, handling missing measurements and building prediction regions [4].

Following a standard statistical viewpoint [20], we assume that the outputs \( \{y_i\} \) are generated by some noisy observations of an underlying "true" function, \( f_\star \), that is \( y_i = f_\star(x_i) + \varepsilon_i \), for \( i = 1, \ldots, n \), where \( \{\varepsilon_i\} \) are the noise terms. The entire noise vector is denoted by \( \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \).

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, particularly, we do not want to assume that \( f_\star \) 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 [15] and search for functions which correctly describe the hidden function, \( f_\star \), 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 \( \mathcal{H}_0 \subseteq \mathcal{H} \) denote the subspace of functions that can be represented as (3). A function \( f_\star \in \mathcal{H}_0 \), having coefficients \( \alpha_0 \in \mathbb{R}^n \), is called an ideal or noise-free representation of function \( f_\star \) w.r.t. data sample \( D_n \), if \( f_\star(x_i) = f_{\star}(x_i) \), for all \( x_1, \ldots, x_n \). The set of all ideal representations is denoted by \( \mathcal{H}_0 \subseteq \mathcal{H}_0 \) and the set of their (ideal) coefficients by \( A_0 \subseteq \mathbb{R}^n \).

Note that an ideal representation does not simply interpolate the observed (noisy) outputs \( \{y_i\} \), but it interpolates the unobserved (noise-free) outputs, \( \{f_\star(x_i)\} \). Also note that if \( K_x \) is positive definite, which is arch-typically the case in practice, then the existence of an ideal representation is guaranteed.

A confidence region, \( C(\cdot, \cdot) \), is a (possibly randomized) function that maps a set to sample \( D_n \) and probability \( p \) [20].

**Definition 2.** A confidence region \( C(\cdot, \cdot) \) is called exact, w.r.t. class \( \mathcal{F} \), if \( \mathbb{P}\left( f \in C(D_n, p) \right) = p \), for all \( f \in \mathcal{F} \). It is called honest if, for all \( f \in \mathcal{F} \), we have \( \mathbb{P}\left( f \in C(D_n, p) \right) \geq p \).

Therefore, we will aim at building confidence regions for ideal representations of the underlying true function, and not directly for \( f_\star \), which might not even be in our RKHS.

Standard ways to construct confidence regions for nonparametric estimators typically either make strong distributional assumptions, like assuming Gaussian processes [10], or resort to asymptotic results, e.g., Donsker-type theorems for Kolmogorov-Smirnov confidence bands [2]. An alternative approach is to build on Rademacher complexities, which can provide non-asymptotic, distribution-free confidence bands [2]. Nevertheless, these regions are conservative (not exact) and are constructed independently of the applied kernel method. In contrast, our approach provides exact, distribution-free, non-asymptotic confidence sets for a user-chosen kernel method.

**A. Non-Asymptotic, Distribution-Free Framework**

This section presents the proposed general framework to quantify the uncertainty of kernel-based estimates. It is inspired by and builds on results from finite-sample system identification, such as the SPS and DP methods [11], [12], [13], [14].

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

**Definition 3.** An \( \mathbb{R}^n \)-valued random vector \( v \) is distributionally invariant with respect to a compact group of transformations,
(G, o), where “o” is the function composition and each G ∈ G maps \( \mathbb{R}^n \) to itself, if for all transformation G ∈ 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 \( \varepsilon \) 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 \( \varepsilon \) 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 are mild regularities, since for example i.i.d. implies exchangeability and most standard distributions are symmetric (e.g., Gauss, Laplace, Cauchy, Student’s t, uniform, plus several multimodal ones).

Note that no assumptions about other properties of the (noise) distributions are needed, e.g., they can be heavy-tailed, with even infinite variance, skewed, and no moment assumptions are necessary. For the case of symmetric distributions, each noise term may even have a different marginal distribution.

1) Main Assumptions: Before the general construction of our method is explained, we highlight its main assumptions.

**Assumption 1.** The input, \( x \), and the noise, \( \varepsilon \), are independent.

**Assumption 2.** Noise \( \varepsilon \) is distributionally invariant w.r.t. a known group of transformations, \( (G, o) \).

**Assumption 3.** The (sub-)gradient of the objective function w.r.t. \( \alpha \) exists and it only depends on the output vector, \( y \), through the residuals \( \hat{\varepsilon}(x, y, \alpha) \). i.e., there is \( \hat{g} \),

\[
\nabla_{\alpha} g(f_{\alpha}, D_{n}) = \hat{g}(x, \alpha, \hat{\varepsilon}(x, y, \alpha)).
\]  

(4)

Assumption 1 implies that the noises, \( \{ \varepsilon_i \} \), do not affect the inputs, \( \{ x_i \} \); e.g., the system is not autoregressive. On the other hand, it allows deterministic inputs, as a special case.

Assumption 2 states that we known transformations that do not change the (joint) distribution of the noise. As it was discussed earlier, symmetry and exchangeability are two standard examples for which we know such group of transformations, but the applied methodology is not limited to those.

For Assumption 3, it is enough if a sub-gradient is defined for each \( \alpha \), hence, e.g., the case of \( \varepsilon \)-insensitive loss functions is also covered. Even in such cases, we still treat \( \hat{g} \) as a vector-valued function and choose arbitrarily from the sub-gradients.

This requirement is also mild as it is typically the case that the objective function has subgradients and only depends on \( y \) through the residuals, which immediately imply Assumption 3.

To see this assume, for simplicity, that \( g \) is differentiable; then clearly, if \( g(f_{\alpha}, D_{n}) = g_0(x, \alpha, \hat{\varepsilon}(x, y, \alpha)) \), then

\[
\nabla_{\alpha} g(f_{\alpha}, D_{n}) = \nabla_{\alpha} (g_0(x, \alpha, y - K_x \alpha))
\]  

\[= -K_x (\nabla_{\alpha} g_0(x, \alpha, y - K_x \alpha))
\]  

\[= \hat{g}(x, \alpha, \hat{\varepsilon}(x, y, \alpha)),
\]  

(5)

where we applied the chain rule, used the fact that matrix \( K_x \) is symmetric and the definition \( \hat{\varepsilon}(x, y, \alpha) = y - K_x \alpha \).

2) Perturbed Gradients and Distributional Invariance: 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 that \( H_0 : \alpha \in A_0 \), i.e., whether it defines an ideal representation; against the alternative hypothesis \( H_1 : \alpha \notin A_0 \). Under \( H_0 \), the residuals of \( f_\alpha \) are the “true” (unobserved) errors, since by definition (for ideal representations)

\[
\hat{\varepsilon}(x, y, \alpha) = y - K_x \alpha = [f_\alpha(x_1) + \varepsilon_1, \ldots, f_\alpha(x_n) + \varepsilon_n]^T
\]  

\[= [f_\alpha(x_1), \ldots, f_\alpha(x_n)]^T = \varepsilon.
\]  

(6)

Consequently, based on the group of invariant transformations, \( G \), we know that the (joint) distribution of the residuals does not change if we transform them by any \( G \in G \) (under \( H_0 \)). Then, we can generate alternative realizations of the residuals, \( \hat{\varepsilon}(x, y, \alpha) \), by applying random transformations from \( G \), and these alternative realizations will behave “similarly” (in the statistical sense) to the original residuals (i.e., the true noise).

However, under \( H_1 \), if coefficient vector \( \alpha \) does not define an ideal representation, \( \hat{\varepsilon}(x, y, \alpha) \), 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. \( \alpha \)) with the transformed residuals and apply a rank test [21].

Let us define a reference function, \( Z_0 \), and \( m - 1 \) perturbed functions, \( \{ Z_i \} \), with \( Z_i : \mathbb{R}^n \rightarrow \mathbb{R} \), as follows

\[
Z_0(\alpha) = \| \Lambda(x) \hat{g}(x, \alpha, G_0(\hat{\varepsilon}(x, y, \alpha))) \|^2,
\]  

(7)

\[
Z_i(\alpha) = \| \Lambda(x) \hat{g}(x, \alpha, G_i(\hat{\varepsilon}(x, y, \alpha))) \|^2,
\]  

(8)

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

For symmetric noises, transformation \( G_i \in G \) 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 \( G_i \in G \) is a randomly (uniformly) chosen \( n \times n \) permutation matrix.

Weighting matrix \( \Lambda(x) \) is included in the construction to allow some additional flexibility, for example, if we have some a priori information on the noises affecting the observations.

We can observe that for an ideal coefficient, \( \alpha_0 \in A_0 \),

\[
Z_0(\alpha_0) = \| \Lambda(x) \hat{g}(x, \alpha_0, \varepsilon) \|^2 \overset{d}{=} \| \Lambda(x) \hat{g}(x, \alpha_0, G_i(\varepsilon)) \|^2 = Z_i(\alpha_0),
\]  

(9)

for \( i = 1, \ldots, m-1 \), where \( \overset{d}{=} \) denotes equality in distribution. Therefore, \( \{ Z_i(\alpha_0) \}_{i=0}^{m-1} \) have the same 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 \( \alpha \not\in A_0 \), this distributional equivalence will not hold, and we expect that if \( d(\alpha, A_0) \) is large enough, for some metric \( d(\cdot, \cdot) \), the reference element \( Z_0(\alpha) \) will dominate the perturbed elements, \( \{ Z_i(\alpha) \}_{i=1}^m \) from which we can detect (statistically) that coefficient vector \( \alpha \) is not ideal, \( \alpha \not\in A_0 \).

3) Normalized Ranks and Exact Confidence: Now, we make our argument precise, by introducing the concept of normalized ranks. Formally, the normalized rank of the reference element, \( Z_0(\alpha) \), among all \( \{ Z_i(\alpha) \}_{i=1}^m \) is defined as follows

\[
R(\alpha) \equiv R_m(\alpha) \equiv \frac{1}{m} \left[ 1 + \sum_{i=1}^{m-1} I( Z_0(\alpha) \prec Z_i(\alpha) ) \right],
\]

where \( I(\cdot) \) 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 design-parameter (the number of perturbed functions plus the reference one); and relation “\( \prec \)” is the standard “\( < \)” with random tie-breaking (according to a fixed, pre-generated random order). More precisely, let \( \pi \) be a random (uniformly chosen) permutation of the set \( \{ 0, \ldots, m - 1 \} \). Then, given \( m \) arbitrary real numbers, \( Z_0, \ldots, Z_{m-1} \), we can construct a strict total order \( \prec \), by defining \( Z_k \prec Z_j \) if and only if \( Z_k < Z_j \) or it both holds that \( Z_k = Z_j \) and \( \pi(k) < \pi(j) \).

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 value is 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

\[
A_p = \{ \alpha : R(\alpha) \leq p \} = \{ \alpha : R_m(\alpha) \leq 1 - q/m \}.
\]

Theorem 2. Under Assumptions 1, 2 and 3: for all ideal representation, \( \alpha_0 \in A_0 \), the coverage probability is exactly

\[
P \left( \alpha_0 \in A_p \right) = p = 1 - \frac{q}{m},
\]

for any user-chosen positive integers \( q < m \).

The proof builds on the results of [13] and [14]; it can be found in the appendix. Theorem 2 shows that the confidence region contains each ideal coefficient, \( \alpha_0 \in A_0 \), exactly with probability \( p \), which statement is non-asymptotically guaranteed, despite the method is distribution-free. Since \( m \) and \( q \) are user-chosen, the confidence probability is under our control.

B. Quadratic Objectives and Symmetric Noises

If we work with quadratic objectives, which have special importance for kernel methods [4], and assume independent and symmetric noise terms, then we get the Sign-Perturbed Sums (SPS) method [13] as a special case (typically using the inverse square root of the Hessian as a weighting matrix).

The SPS method uses the classical least-squares (LS) objective function \( g(f_\alpha, D_n) = \| z - \Phi \alpha \|^2 \), which can be seen as a canonical form of many quadratic functions. When using SPS, we assume that the noises, \( \{ \varepsilon_i \} \), are independent and have symmetric distributions about zero; and that the regressor matrix, \( \Phi \), has independent rows, it is skinny and full rank.

For SPS, the reference and the perturbed functions are

\[
Z_i(\alpha) = \| (\Phi^T \Phi)^{-1/2} \Phi^T G_i (z - \Phi \alpha) \|^2,
\]

for \( i = 0, \ldots, m - 1 \), where \( G_0 = I_n \), and \( G_i = \text{diag}(\sigma_{i,1}, \ldots, \sigma_{i,n}) \); where (Rademacher) variables \( \{ \sigma_{i,j} \} \) are i.i.d. taking values \( +1 \) and \( -1 \) with probability \( 1/2 \) each.

It is easy to see that (13) is a special case of construction (7)-(8), where \( z \) are the outputs and \( \Phi \) 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, \( \hat{\alpha} \), as a star center [13]. Moreover, they have ellipsoidal outer approximations, that is

\[
A_p^\circ = \{ \alpha \in \mathbb{R}^n : (\alpha - \hat{\alpha})^T \frac{1}{n} \Phi^T \Phi (\alpha - \hat{\alpha}) \leq r \}
\]

where \( A_p \subseteq A_p^\circ \) and radius \( r \) can be computed (in polynomial time) by solving semi-definite programming problems [13].

Therefore, 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 the star convexity and boundedness properties, we can efficiently evaluate the region by knowing that every point of it can be reached from a star center by a line segment inside the region. The ellipsoidal outer approximation provides a compact way to represent the region.

IV. DEMONSTRATIVE EXAMPLES

Now, we present some examples of the proposed uncertainty quantification (UQ) approach for typical kernel methods.

A. 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 [22]. LS-SVC under the Euclidean distance is known to be equivalent to hard margin SVC using the Mahalanobis distance [23]. It has the advantage that it can be solved by a system of linear equations, in contrast to a quadratic optimization problem.

We assume that \( x_k \in \mathbb{R}^d \) and \( y_k \in \{ +1, -1 \} \), for all \( k \in \{ 1, \ldots, n \} \), 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 latter we identify with the best possible classifier (that minimizes the “true” risk).

For simplicity, we consider linear classification, that is

\[
h_{\alpha}(x_k) \triangleq \text{sign}(w^T x_k + b) = \text{sign}(\alpha^T \tilde{x}_k),
\]

where \( x_k \) is an input, \( \alpha \triangleq [b, w^T]^T \) and \( \tilde{x}_k \triangleq [1, x_k^T]^T \).

The (primal) objective function of LS-SVM classification is

\[
g(f_\alpha, D_n) \triangleq \frac{1}{2} \| B \alpha \|^2 + \lambda \| I_n - y \odot (X \alpha) \|^2,
\]

where \( \lambda > 0 \), \( I_n \in \mathbb{R}^n \) is the all-one vector, \( \odot \) denotes the Hadamard (entrywise) product, \( X \triangleq [\tilde{x}_1, \ldots, \tilde{x}_n]^T \) and matrix \( B \) “removes” the bias, \( b \), from \( \alpha \), i.e., \( B \triangleq \text{diag}(0,1, \ldots, 1) \).
Observe that the objective function, \( g(\alpha, D_n) \), can be written in the canonical form of \( \| z - \Phi \alpha \|^2 \) by using

\[
\Phi = \begin{bmatrix}
\sqrt{\lambda}(y^{\top}X^\top) \odot X \\
(1/\sqrt{\lambda}) B
\end{bmatrix}, \quad z = \begin{bmatrix}
\sqrt{\lambda} \mathbb{1}_n \\
0_d
\end{bmatrix},
\]

where \( 0_d \in \mathbb{R}^d \) is the all-zero vector. Then, (exact) confidence regions and (honest) ellipsoidal outer approximations can be constructed in the domain of coefficients by the SPS method, i.e., (13). 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 parts (a) and (d) of Figure 1.

\[\text{B. Uncertainty Quantification for Kernel Ridge Regression}\]

Our next example is Kernel Ridge Regression (KRR) which is a kernelized version of Tikhonov regularized LS [3]. It has a quadratic loss function with a Hilbert space norm regularizer,

\[
g(\alpha, D_n) = \frac{1}{2} \| y - f(\alpha) \|^2_1 + \lambda \| f \|^2_H
\]

where \( f(\alpha) = [f(\alpha)(x_1), \ldots, f(\alpha)(x_n)]^T \), \( \lambda > 0 \), \( W \) is a positive definite weighting matrix and we used the reproducing property to replace the Hilbert space norm with a quadratic term.

We can reformulate (18) in the canonical form, \( \| z - \Phi \alpha \|^2 \), assuming the Gram matrix, \( K_x \), is positive definite, by using

\[
\Phi = \begin{bmatrix}
(1/\sqrt{\lambda}) W^{1/2} K_x \\
\sqrt{\lambda} K_x^{1/2}
\end{bmatrix}, \quad z = \begin{bmatrix}
(1/\sqrt{\lambda}) W^{1/2} y \\
0_n
\end{bmatrix},
\]

where \( W^{1/2} \) and \( K_x^{1/2} \) are the (unique) positive definite roots of \( W \) and \( K_x \), respectively. Then, assuming symmetric noises, (13) can be applied to build confidence regions, cf. Figure 1.

\[\text{C. Uncertainty Quantification for Kernelized LASSO}\]

The previous examples were quadratic and therefore, for symmetric noises, their uncertainty could be quantified with SPS. This may no more be true if we change the applied norms. For example, consider the kernelized version of LASSO (least absolute shrinkage and selection operator) with objective \[24\]

\[
g(\alpha, D_n) = \frac{1}{2} \| y - K_x \alpha \|^2 + \lambda \| \alpha \|_1,
\]

where \( \| \cdot \|_1 \) is the L1 (or Manhattan) norm. Though, function (20) cannot be written as \( \| z - \Phi \alpha \|^2 \), the proposed construction (7)-(8) can still be applied. A sub-gradient of (20) is

\[
\nabla_\alpha g(\alpha, D_n) = K_x (K_x \alpha - y) + \lambda \text{sign}(\alpha),
\]

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

\[
\mathbf{Z} (\alpha) = \| K_x G_i (K_x \alpha - y) + \lambda \text{sign}(\alpha) \|^2,
\]

were \( G_i \) are from a suitable transformation group, e.g., permutation matrices for exchangeable noises, see Figure 1.

\[\text{D. Uncertainty Quantification for Support Vector Regression}\]

Our last example covers support vector regression, particularly, \( \varepsilon \)-SVR [4], [7], [8]. An advantage of \( \varepsilon \)-SVR, e.g., over KRR, is that it ensures sparse representations through the \( \varepsilon \)-insensitive loss function. To avoid confusion with the true noise vector, \( \varepsilon \), we denote the tolerance parameter of the loss function by \( \bar{\varepsilon} \). The primal objective function of \( \varepsilon \)-SVR is defined as

\[
h(\alpha, D_n) = \frac{1}{2} \| f \|^2_H + \frac{\varepsilon}{n} \sum_{k=1}^n \max \{ 0, |\langle f, \phi(x_k) \rangle_H - y_k | - \bar{\varepsilon} \},
\]

where \( f \in \mathcal{H}, c > 0 \), and \( \phi(z) = k(z, \cdot) \) is the feature map. Function (23) can be reformulated by applying slack variables, then using the usual arguments based on Lagrangian duality and the Karush–Kuhn–Tucker (KKT) conditions, we arrive at the dual form of \( \varepsilon \)-SVR [7], where we have to maximize

\[
g(\alpha^+, \alpha^-, D_n) = y^T (\alpha^+ - \alpha^-) - \frac{1}{2} (\alpha^+ - \alpha^-)^T K_x (\alpha^+ - \alpha^-) - \bar{\varepsilon} (\alpha^+ + \alpha^-)^T 1, 24
\]

subject to the (linear) constraints: \( \alpha^+, \alpha^- \in [0, c/n] n \) and \( (\alpha^+ - \alpha^-)^T 1 = 0 \). One can work directly with the quadratic dual objective, but then the confidence region will be constructed for \( \alpha^+, \alpha^- \). Since, \( \alpha = \alpha^+ - \alpha^- \), the region could be transformed to a confidence region in the space of \( \alpha \) coefficients. Alternatively, one can reformulate (24) directly for \( \alpha \) as

\[
g(\alpha, D_n) = y^T \alpha - \frac{1}{2} \alpha^T K_x \alpha - \bar{\varepsilon} \| \alpha \|_1,
\]

where \( \| \cdot \|_1 \) is the 1-norm. A subgradient of (25) w.r.t. \( \alpha \) is

\[
\nabla_\alpha g(\alpha, D_n) = y - K_x \alpha - \bar{\varepsilon} \text{sign}(\alpha),
\]

where sign(\cdot) is again understood component-wise.

Then, reference and perturbed functions can be defined as

\[
\mathbf{Z}_i (\alpha) = \| G_i (y - K_x \alpha) - \bar{\varepsilon} \text{sign}(\alpha) \|^2,
\]

for \( i = 0, \ldots, m - 1 \), where \( G_0 \) is the identity matrix and \( G_i \) is a (uniformly chosen) element of the transformation group, such as a diagonal matrix with \( \pm 1 \) entries, for symmetric noises. Confidence sets for \( \varepsilon \)-SVR are shown in part (f) of Figure 1.

\[\text{E. Numerical Experiments}\]

Several numerical experiments were carried out, in order to demonstrate the proposed UQ approach. UQ for LS-SVC, KRR, Klasso and \( \varepsilon \)-SVR are illustrated in parts of Figure 1.

For LS-SVC the SPS method was used, and 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 in the coefficient space.

The same data sample was used for all regression models, to allow their comparison. Since in that case 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. As the confidence regions are built
for ideal representations, which belong to the chosen RKHS (unlike the true underlying function), it is meaningful to plot the confidence regions even for unknown input values.

It can be observed that the uncertainty of $\varepsilon$-SVR was the highest, e.g., compared to that of KRR and KLASSO, which can be explained as the price of using $\varepsilon$-insensitive loss functions. Naturally, the uncertainty is also influenced by the specific choice of hyper-parameters. KLASSO models, which ensure sparsity via 1-norm regularization, were evaluated two ways: by exploiting the symmetry of the noises (sign-change matrices) and by exploiting the exchangeability of the noises (permutation matrices). Though these two types of regions were similar for regions having low confidence probabilities, the permutation-based approach produced smaller high-confidence regions.

V. 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 an ideal (noise-free) 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; e.g., the noises are exchangeable or symmetric. The core idea was to evaluate the uncertainty via perturbing the residuals in the gradient of the objective function. The method was also demonstrated on specific examples and by experiments.

The proposed approach to build 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.

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 estimate, and therefore, besides being vital for risk management, it also has the potential to inspire refinements or new constructions.
REFERENCES

[1] N. Aronszajn, “Theory of reproducing kernels,” Transactions of the American Mathematical Society, vol. 68, no. 3, pp. 337–404, 1950.
[2] E. Giné and R. Nickl, Mathematical Foundations of Infinite-Dimensional Statistical Models. Cambridge University Press, 2015, vol. 40.
[3] J. Shawe-Taylor and N. Cristianini, Kernel Methods for Pattern Analysis. Cambridge University Press, 2004.
[4] T. Hofmann, B. Schölkopf, and A. J. Smola, “Kernel methods in machine learning,” The Annals of Statistics, vol. 36, pp. 1171–1220, 2008.
[5] G. Kulin and G. Wahba, “Some results on techebycheffian spline functions,” Journal of Mathematical Analysis and Applications, vol. 33, no. 1, pp. 82–95, 1971.
[6] B. Schölkopf, R. Herbrich, and A. J. Smola, “A generalized representer theorem,” in Annual Conference on Learning Theory (COLT). Springer, 2001, pp. 416–426.
[7] B. Schölkopf and A. J. Smola, Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT Press, 2001.
[8] I. Steinwart and A. Christmann, Support Vector Machines. Springer Science & Business Media, 2008.
[9] V. N. Vapnik, Statistical Learning Theory. Wiley-Interscience, 1998.
[10] C. E. Rasmussen and C. K. I. Williams, Gaussian Processes for Machine Learning, ser. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press, 2006.
[11] M. C. Campi and E. Weyer, “Guaranteed non-asymptotic confidence regions and regularization,” The Annals of Statistics, vol. 36, pp. 1171–1220, 2015.
[12] A. Carè, B. Cs. Csáji, M. Campi, and E. Weyer, “Finite-sample system identification approach for constructing exact non-asymptotic confidence regions in linear regression models,” IEEE Transactions on Signal Processing, vol. 63, no. 1, pp. 169–181, 2015.
[13] S. Kolumbán, “System identification in highly non-informative environment,” Ph.D. dissertation, Budapest University of Technology and Economics and Vrije Universiteit Brussels, 2016.
[14] K.-C. Li, “Honest confidence regions for nonparametric regression,” The Annals of Statistics, pp. 1001–1008, 1989.
[15] B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap. CRC press, 1994.
[16] V. Vovk, A. Gammerman, and G. Shafer, Algorithmic Learning in a Random World. Springer Science & Business Media, 2005.
[17] Yu. Yu, H. Cheng, D. Schuurmans, and C. Szepesvári, “Characterizing the representer theorem,” in International Conference on Machine Learning, 2013, pp. 570–578.
[18] A. Argyriou and F. Dinuzzo, “A unifying view of representer theorems,” in International Conference on Machine Learning (ICML), 2014, pp. 748–756.
[19] P. L. Davies, A. Kovac, and M. Meise, “Nonparametric regression, confidence regions and regularization,” The Annals of Statistics, pp. 2597–2625, 2009.
[20] P. Good, Permutation, Parametric, and Bootstrap Tests of Hypotheses. Springer, 2005.
[21] J. A. K. Suykens and J. Vandewalle, “Least squares support vector machine classifiers,” Neural Processing Letters, vol. 9, no. 3, pp. 293–300, 1999.

APPENDIX

Proof of Theorem 2

Following [13], the core idea is to show that \( \{Z_i(\alpha_0)\}_{i=0}^{m-1} \) are uniformly ordered, which means that each ordering of them has the same probability, \( 1/m! \). This is not obvious, since they are not independent, even though we already observed that they are identically distributed (for ideal coefficients).

By definition, \( \alpha_0 \in A_p \) if and only if \( R(\alpha_0) \leq 1 - q/m \), i.e., if the reference element, \( Z_0(\alpha_0) \) takes one of the positions \( 1, \ldots, m - q \) in the ordering of \( \{Z_i(\alpha_0)\}_{i=0}^{m-1} \) variables, w.r.t. \( \leq \). Then, assuming they are uniformly ordered (yet to be shown), we know that \( Z_0(\alpha_0) \) takes each position in the ordering with probability \( 1/m \). Thus \( P[R(\alpha_0) = i/m] = 1/m \) for \( i \in \{1, \ldots, m\} \), from which \( P[\alpha_0 \in A_p] = 1 - q/m \), since \( R(\alpha_0) = i/m \) and \( R(\alpha_0) = j/m \) are disjoint if \( i \neq j \).

In order to show that \( \{Z_i(\alpha_0)\}_{i=0}^{m-1} \) are indeed uniformly ordered, we can apply Theorem 2.17 of [14]. Our proposed approach is in fact a DP method, even though formally the “performance measures” of DP methods can depend on the parameters, \( \alpha \), the inputs, \( x \), and the perturbed outputs, \( y^{(i)} \), but not directly on the perturbed residuals. In our case, \( y^{(i)} \) is

\[
y^{(i)} = f_\alpha(x) + G_1(\bar{\varepsilon}(x, y, \alpha)),
\]

where \( f_\alpha(x) = [f_\alpha(x_1), \ldots, f_\alpha(x_n)]^T \). Then, obviously we can get \( G_1(\bar{\varepsilon}(x, y, \alpha)) \) from \( f_\alpha \), and \( y^{(i)} \) by \( G_1(\bar{\varepsilon}(x, y, \alpha)) = y^{(i)} - f_\alpha(x) \), thus, for us, the DP performance measure is

\[
Z(\alpha, x, y^{(i)}) = \|\Lambda(x) \bar{g}(x, \alpha, y^{(i)} - f_\alpha(x))\|^2,
\]

which perfectly fits the DP framework. Our Assumption 3 ensures that this function is well-defined and, together with Assumption 1, it also guarantees that we do not need to compute \( \{y^{(i)}\} \) to evaluate the perturbed functions. Our Assumption 2 directly states that the noise, \( \varepsilon \), is invariant under a compact group of transformations, which is a requirement of Theorem 2.17, and we already observed that \( \bar{\varepsilon}(x, y, \alpha_0) = \varepsilon \). \( \square \)