Semi-Parametric Uncertainty Bounds for Binary Classification

Balázs Csanád Csáji[1] Ambrus Tamás[1]

***Abstract*— The paper studies binary classification and aims at estimating the underlying regression function which is the conditional expectation of the class labels given the inputs. The regression function is the key component of the Bayes optimal classifier, moreover, besides providing optimal predictions, it can also assess the risk of misclassification. We aim at building non-asymptotic confidence regions for the regression function and suggest three kernel-based semi-parametric resampling methods. All of them guarantee confidence regions with exact coverage probabilities and they are strongly consistent.**

I. Introduction

Classification is one of the principal problems of *statistical learning theory* [1], and it is widely applied across several fields [2], for example, in quantized identification [3]. A typical aim of classification is to minimize the *probability of misclassification*. If the (joint) probability distribution of the input-output pairs was known, the misclassification probability could be minimized by the Bayes optimal classifier. This classifier can be written as the sign of the *regression function* which is the conditional expectation of the labels given the inputs. The regression function can also help to assess the risk of misclassification. Estimating the regression function can be seen as identifying a (nonlinear) function from a sample of input and *quantized* (binary) output measurements.

Besides providing point-estimates of the regression function, for which there are several methods available [1], [4], it is also an important problem to bound the *uncertainty* of a candidate model. We will provide these bounds in the form of *confidence regions*. Note that such regions also induce confidence sets for the misclassification probabilities.

In this paper, inspired by recent developments in Finite-Sample System Identification (FSID) [5], [6], [7], [8], we suggest three semi-parametric *kernel-based* [2] resampling algorithms to build *non-asymptotic* confidence regions for the regression function of binary classification. We argue that each of these algorithms provides confidence sets with *exact* coverage probabilities, and they are *strongly consistent*, that is any false model will be (almost surely) excluded from the confidence regions, as the sample size tends to infinity. As the suggested algorithms build on distribution-free results and work directly with the samples, the constructions are not restricted to models parametrized by finite dimensional vectors, but also allow infinite dimensional model classes.

*This work was supported by the National Research, Dev. and Innovation Office (NKFIH), Hungary, grant numbers ED_18-2-2018-0006 and KH_17 125698. B. Cs. Csáji was supported by a János Bolyai Res. Fellowship.

[1] Balázs Csanád Csáji and Ambrus Tamás are with MTA SZTAKI: The Institute for Computer Science and Control, Hungarian Academy of Sciences, Budapest, Hungary, balazs.csaji@sztaki.mta.hu, ambrus.tamas@sztaki.mta.hu

II. Preliminaries

A. Binary Classification

Let $(\mathbb{X}, \mathcal{X})$ be a measurable space and $\mathbb{Y} \doteq \{+1, -1\}$. Let $\mathcal{D} \doteq \{(x_i, y_i)\}_{i=1}^n$ be an i.i.d. (independent and identically distributed) sample from an unknown probability distribution P on $\mathbb{X}\times\mathbb{Y}$, where $x_i \in \mathbb{X}$ is the input and $y_i \in \mathbb{Y}$ is the label of the ith observation. We call any (measurable) function $g : \mathbb{X} \to \mathbb{Y}$ a *classifier*. The *Bayes optimal classifier*, g_*, is a classifier which minimizes the risk, $R(g) \doteq \mathbb{E}\big[L(Y, g(X))\big]$, where L is an arbitrary loss function, that is a $\mathbb{Y}^2 \to [0, \infty)$ function penalizing label mismatch. Here, we focus on the $0/1$ loss which is an arch-typical choice [1]. It is defined by $L(y, g(x)) \doteq \mathbb{I}(g(x) \neq y)$, where $\mathbb{I}$ is the indicator function. The corresponding risk is simply $R(g) = \mathbb{P}(g(X) \neq Y)$.

Since the joint distribution of the inputs and outputs, P, is unknown, we typically aim at estimating g_*. At any $x \in \mathbb{X}$, $g_*(x) = \operatorname{sign}\big(\mathbb{E}\big[\, Y \,|\, X = x \,\big]\big)$, if it is feasible [4]. Note that the *regression function* $f_*(x) \doteq \mathbb{E}\big[\, Y \,|\, X = x \,\big]$ contains even more information than the Bayes optimal classifier, g_*, e.g., $f_*(x)$ also encodes the probability of misclassification of x. Therefore, it is of high importance to estimate f_*.

B. Reproducing Kernel Hilbert Spaces

A Hilbert space $\mathcal{H}$ of $f : \mathbb{X} \to \mathbb{R}$ type functions, with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, is a *Reproducing Kernel Hilbert Space* (RKHS) if the point evaluation functional, $\delta_x : f \to f(x)$, is bounded, or equivalently continuous, for all $x \in \mathbb{X}$ [2]. Then it can be proven, by applying the Riesz representation theorem, that there uniquely exists $k : \mathbb{X}^2 \to \mathbb{R}$, the *kernel* of $\mathcal{H}$, such that for all $x \in \mathbb{X}$, $k(\cdot, x) \in \mathcal{H}$ and $f(x) = \langle\, f, k(\cdot, x)\,\rangle_{\mathcal{H}}$, which is called the *reproducing property*. In particular $\langle\, k(\cdot, x), k(\cdot, y)\,\rangle_{\mathcal{H}} = k(x, y)$ thus k is symmetric and positive definite. The converse is also true by the Moore-Arnoszjan theorem [9]: for each symmetric, positive definite function there uniquely exists an RKHS. Typical examples of kernels are the Gaussian kernel, $k(x, y) = \exp(\frac{-\|x-y\|^2}{2\sigma^2})$ with $\sigma > 0$, and the polynomial kernel, $k(x, y) = (c + x^{\mathrm{T}} y)^d$ with $c \geq 0$ and $d \in \mathbb{N}$. For a given sample $\mathcal{D}$, the Gram matrix, $K \in \mathbb{R}^{n\times n}$, is defined as $K_{i,j} \doteq k(\, x_i, x_j\,)$, which is a (data-dependent) symmetric, positive semidefinite matrix.

Let $\mathbb{X}$ be a metric space and $Z \subseteq \mathbb{X}$ be a compact set. Let $\mathcal{C}(Z)$ denote the space of continuous $Z \to \mathbb{R}$ type functions with the uniform (sup) norm. Let us consider the subspace $\mathcal{H}(Z) \doteq \operatorname{span}\big\{K(\cdot, z) : z \in Z\big\} \subseteq \mathcal{H}$, which contains all finite linear combinations of $\{K(\cdot, z)\}$. A kernel is *universal* if for each compact set $Z \subseteq \mathbb{X}$, function $f \in \mathcal{C}(Z)$, and $\varepsilon > 0$, there exists an approximation $h \in \mathcal{H}(Z)$, such that $\sup_{z\in Z} |h(z) - f(z)| \leq \varepsilon$, i.e., $\mathcal{H}(Z)$ is *dense* in $\mathcal{C}(Z)$.

C. Kernel Mean Embedding of Distributions

The idea of *kernel mean embedding* is to map distributions to elements of an RKHS with the help of the kernel [10]. Let $(\mathbb{X}, \mathcal{X})$ be a measurable space and let $M_+(\mathbb{X})$ denote the space of all probability measures on it. The kernel mean embedding of these probability measures into an RKHS $\mathcal{H}$, endowed with a (measurable) kernel $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$, is

$$\mu : M_+(\mathbb{X}) \to \mathcal{H} \qquad \text{and} \qquad P \to \int k(x, \cdot) P(dx), \quad (1)$$

assuming the integral is well-defined (e.g., k is bounded).

A kernel is called *characteristic* if the embedding, μ, is *injective* (e.g., the Gaussian kernel). In this case the embedded element captures all informations about the distribution, e.g., for all $P, Q \in M_+(\mathbb{X})$, $\| \mu_P - \mu_Q \|_{\mathcal{H}} = 0$ if and only if $P = Q$. Hence, the embedding induces a *metric* on $M_+(\mathbb{X})$.

The kernel mean embedding has nice properties even when the kernel is not characteristic. For example, for polynomial kernels with degree d it holds that $\| \mu_P - \mu_Q \|_{\mathcal{H}} = 0$ if and only if the first d moments of P and Q are the same.

Furthermore, many fundamental operations can be performed in $\mathcal{H}$ instead of dealing with the distributions themselves, e.g., Smola showed [10] that $\mathbb{E}_P[f(X)] = \langle f, \mu_P \rangle_{\mathcal{H}}$.

The underlying probability distribution of the sample is typically unknown, therefore, the kernel mean embedding should be estimated from empirical data. An important tool to prove the validity of such approaches is the *Strong Law of Large Numbers* (SLLN) for random elements taking values in a *separable* Hilbert space $\mathcal{H}$. Let $\{X_n\}$ be a sequence of *independent* random elements taking values in $\mathcal{H}$. If

$$\sum_{n=1}^{\infty} \frac{\mathrm{Var}(X_n)}{n^2} < \infty \quad (2)$$

where $\mathrm{Var}(X) \doteq \mathbb{E}\big[\| X - \mathbb{E}[X] \|_{\mathcal{H}}^2 \big]$, then, as $n \to \infty$,

$$\frac{1}{n} \sum_{k=1}^{n} \big(X_k - \mathbb{E}[X_k] \big) \xrightarrow{a.s.} 0, \quad (3)$$

in the metric induced by $\| \cdot \|_{\mathcal{H}}$, see [11, Theorem 3.1.4].

III. Resampling Framework

In this section we develop a *framework* to provide non-asymptotically guaranteed uncertainty quantification *resampling* algorithms for the "regression function", namely, the conditional expectation of the labels given the inputs. The regression function is a fundamental object to study, for example, its signs at various inputs define the Bayes optimal classifier which achieves minimal *misclassification risk*.

Assume we have a distribution on $\mathbb{S} \doteq \mathbb{X} \times \mathbb{Y}$, where $\mathbb{X}$ and $\mathbb{Y}$ are the input and output spaces, respectively. Space $\mathbb{X}$ does not have to be a subset of $\mathbb{R}^d$, it can be a general measurable space, while $\mathbb{Y} \doteq \{+1, -1\}$, since we consider binary classification. Then, the *regression function* satisfies

$$\begin{aligned} f_*(x) &\doteq \mathbb{E}\big[Y \mid X = x \big] \\ &= \mathbb{P}(\, Y = +1 \mid X = x \,) - \mathbb{P}(\, Y = -1 \mid X = x \,) \\ &= 2 \cdot \mathbb{P}(\, Y = +1 \mid X = x \,) - 1. \end{aligned} \quad (4)$$

Given f_*, the *Bayes optimal classifier* is [4]

$$g_*(x) \doteq \mathrm{sign}(f_*(x)), \quad (5)$$

where "sign" denotes the signum function. Note that in (5), for simplicity, we assumed that $\mathbb{P}(\, f_*(X) \neq 0 \,) = 1$.

We assume that we are given an (indexed) *family* of possible regression functions that also contains f_*, that is

$$f_* \in \mathcal{F} \doteq \big\{ f_\theta : \mathbb{X} \to [\,-1, +1\,] \mid \theta \in \Theta \big\}. \quad (6)$$

For simplicity, we refer to $\theta \in \Theta$ as a *parameter*, but Θ can be an *arbitrary* set, even an infinite dimensional vector space. The true parameter is denoted by θ^*, that is $f_{\theta^*} = f_*$.

We assume that $\mathcal{F}$ contains *square integrable* functions w.r.t. the input distribution, and that the parametrization is *injective*, i.e., $\theta_1 \neq \theta_2$ implies $f_{\theta_1} \neq f_{\theta_2}$ on a set having nonzero measure w.r.t. the input distribution. In other words,

$$\| f_{\theta_1} - f_{\theta_2} \|_P^2 \doteq \int_{\mathbb{X}} (f_{\theta_1}(x) - f_{\theta_2}(x))^2 P_{\mathbb{X}}(dx) \neq 0, \quad (7)$$

if $\theta_1 \neq \theta_2$, where $P_{\mathbb{X}}$ is the distribution of the inputs.

Note that f_* in itself does *not* determine the joint probability distribution generating the observations, namely, it does not contain information about the (marginal) distribution of the inputs, therefore, our approach is *semi-parametric*.

As an example, consider the case where the "$+1$" class has probability density function φ_1, while the "-1" class has density φ_2. For each element of the sample, there is a p probability to see an element with "$+1$" label and a $1 - p$ probability to see a measurement with "-1" label. Then,

$$\mathbb{E}\big[Y \mid X = x \big] = \frac{p\, \varphi_1(x) - (1-p)\, \varphi_2(x)}{p\, \varphi_1(x) + (1-p)\, \varphi_2(x)}, \quad (8)$$

thus, if we have candidate densities for inputs with various labels and we know their mixing probability, then we can compute the regression function. However, observe that the regression function does not determine φ_1, φ_2 and p.

A. Resampling Labels

The observed i.i.d. input-output dataset is denoted by

$$\mathcal{D}_0 \doteq ((x_1, y_1), \ldots, (x_n, y_n)), \quad (9)$$

which can also be seen as a $\mathbb{S}^n$-valued random vector.

One of our core ideas is that if we are given a candidate θ, then we can generate (resample) alternative labels for the available inputs using the distribution induced by f_θ, that is

$$\begin{aligned} \mathbb{P}_\theta(\, Y = +1 \mid X = x \,) &= \frac{f_\theta(x) + 1}{2}, \\ \mathbb{P}_\theta(\, Y = -1 \mid X = x \,) &= \frac{1 - f_\theta(x)}{2}, \end{aligned} \quad (10)$$

which immediatelly follow from our observations in (4).

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

$$\mathcal{D}_i(\theta) \doteq ((x_1, y_{i,1}(\theta)), \ldots, (x_n, y_{i,n}(\theta))), \quad (11)$$

for $i = 1, \ldots, m-1$, where for all i, j, label $y_{i,j}(\theta)$ is generated randomly according to the conditional distribution $\mathbb{P}_\theta(\, Y \mid X = x_j \,)$. For notational simplicity, we extend this to $\mathcal{D}_0$, that is $\forall\, \theta : \mathcal{D}_0(\theta) \doteq \mathcal{D}_0$ and $\forall\, j : y_{0,j}(\theta) \doteq y_j$.

Naturally, for all i, dataset $\mathcal{D}_i(\theta)$ can also be identified with a random vector in $\mathbb{S}^n$, and $\mathcal{D}_1(\theta), \dots, \mathcal{D}_{m-1}(\theta)$ are always *conditionally i.i.d.*, given the inputs, $\{x_j\}$.

Observe that, in case $\theta \neq \theta^*$, the distribution of $\mathcal{D}_0$ is in general different than that of $\mathcal{D}_i(\theta)$, $\forall\, i \neq 0$; while $\mathcal{D}_0$ and $\mathcal{D}_i(\theta^*)$ have the same distribution for all possible i.

B. Ranking Functions

The proposed algorithms will be defined via rank statistics based on suitably defined orderings. A key concept will be the "ranking function" which, informally, computes the rank of its first argument among all of its arguments based on some underlying ordering principle. Let $(\mathbb{A}, \mathcal{A})$ be a measurable space. A (measurable) function $\psi : \mathbb{A}^m \to [\,m\,]$, where $[\,m\,] \doteq \{1, \dots, m\}$, is called a *ranking function* if for all $(a_1, \dots, a_m) \in \mathbb{A}^m$ it satisfies the two properties

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

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

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

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

$$\psi\big(\, a_i, \{a_k\}_{k\neq i} \,\big) \;\neq\; \psi\big(\, a_j, \{a_k\}_{k\neq j} \,\big), \tag{12}$$

where the simplified notation is justified by (P1).

We refer to the output of the ranking function ψ as the *rank*. An important observation about ranking *exchangeable* random elements is given by the following lemma. (Recall that if a sample is i.i.d., then it is also exchangeable.)

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

The proofs, except that of Theorem 1, are omitted due to lack of space, but will be included in an extended version.

C. Confidence Regions

Inspired by FSID methods [5], [6], [7], the core idea of the proposed algorithms is to compare the original dataset with alternative samples which are randomly generated according to a given hypothesis. The comparison will be based on the rank of the original dataset among all the available samples, therefore, the ranking function is in the heart of all proposed algorithms. The differences between various algorithms primarily come from the various ways they rank.

Lemma 1 will be one of our main technical tools, however, it requires almost surely different elements, which is not guaranteed for $\{\mathcal{D}_k(\theta)\}$. This will be resolved by *random tie-breaking*, similarly to the solution of [6]. To make this precise, consider a permutation π of the set $\{0, \dots, m-1\}$, generated randomly with *uniform* distribution, and independently of $\{\mathcal{D}_k(\theta)\}$. Then, obviously $\pi(0), \dots, \pi(m-1)$ are almost surely different, *exchangeable* random variables.

We extend datasets $\{\mathcal{D}_k(\theta)\}$ with $\{\pi(k)\}$. As a shorthand notation we introduce, for $k = 0, \dots, m-1$, the sample

$$\mathcal{D}^\pi_k(\theta) \;\doteq\; \big(\mathcal{D}_k(\theta), \pi(k)\big), \tag{13}$$

which now takes values in $\mathbb{A} \doteq \mathbb{S}^n \times \{0, \dots, m-1\}$.

Given a ranking function ψ, defined on the codomain (range) of the extended datasets, and hyper-parameters $p, q \in [\,m\,]$ with $p \leq q$, a *confidence region* can be defined by

$$\Theta^\psi_\varrho \;\doteq\; \big\{\, \theta \in \Theta : \, p \;\leq\; \psi\big(\, \mathcal{D}^\pi_0, \{\mathcal{D}^\pi_k(\theta)\}_{k\neq 0} \,\big) \;\leq\; q \,\big\}, \tag{14}$$

where $\varrho \doteq (m, p, q)$ denotes the applied hyper-parameters, with $m \geq 1$ being the total number of available samples, including the original one as well as the generated ones.

Our main abstract result about the *coverage probability* of the true parameter of such confidence regions is

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

$$\mathbb{P}\big(\, \theta^* \in \Theta^\psi_\varrho \,\big) \;=\; \frac{q - p + 1}{m}. \tag{15}$$

Proof: First note that $\mathcal{D}_0, \mathcal{D}_1(\theta^*), \dots, \mathcal{D}_{m-1}(\theta^*)$ are conditionally i.i.d., given the inputs, $\{x_k\}$, therefore they are also exchangeable. As $\pi(0), \dots, \pi(m-1)$ are exchangeable, as well, and π is generated independently of the datasets, we have that $\mathcal{D}^\pi_0, \mathcal{D}^\pi_1(\theta^*), \dots, \mathcal{D}^\pi_{m-1}(\theta^*)$ are *exchangeable*, too, furthermore, they are almost surely pairwise different.

Then, the theorem follows directly from Lemma 1, as the lemma implies that the rank of D^π_0 takes each value in $[\,m\,]$ with probability exactly $1/m$, therefore, the probability that its rank is between p and q is exactly $(\,q - p + 1\,)\,/\,m$. ∎

Theorem 1 shows that the confidence regions constructed as (14) have *exact* coverage probabilities, independently of the underlying probability distribution generating the (i.i.d.) data and for all ranking functions (satisfying P1 and P2). Observe that it is a *non-asymptotic* result, the exact coverage probability is valid irrespective of the sample size, n. Also note that the hyper-parameters are user-chosen, therefore, any (rational) confidence probability in $(0, 1)$ can be achieved.

This theorem is very general and hence also allows some degenerate constructions, like the ones that do not depend on the data at all, only on the tie-breaking random permutation, π. Such regions are called *purely randomized*. In order to avoid such constructions, we should analyze other properties of the methods. Besides having guaranteed confidence, one of the most important properties an algorithm can have is (strong) consistency, namely, the property that, as the sample size tends to infinity, any (fixed) false parameter will be eventually (a.s.) excluded from the confidence region.

Formally, a method is *strongly consistent* if

$$\mathbb{P}\left(\, \bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} \big\{\, \theta \in \Theta^\psi_{\varrho,n} \,\big\} \,\right) \;=\; 0, \tag{16}$$

for all parameter $\theta \neq \theta^*$, $\theta \in \Theta$, where $\Theta^\psi_{\varrho,n}$ denotes the confidence region constructed based on a sample of size n. Obviously, purely randomized regions are not consistent.

IV. Kernel-Based Constructions

In this section we propose three kernel-based algorithms to construct exact, non-asymptotic confidence regions based on the framework of Section III. All of these methods have *exact* coverage probabilities and are *strongly consistent*.

A. Algorithm I (Neighborhood Based)

The main idea of Algorithm 1 is that we can estimate the regression function, f_*, based on the available (quantized) dataset, $\mathcal{D}_0$, by the kNN (k-nearest neighbors) algorithm. We can similarly do so based on the alternative datasets, $\{\mathcal{D}_k(\theta)\}_{k\neq 0}$. Then, we can compare the estimate based on $\mathcal{D}_0$ to the ones coming from the alternative samples.

For Algorithm I, we assume that $\mathbb{X} \subseteq \mathbb{R}^d$, $\mathbb{X}$ is *compact*, the *support* of the (marginal) distribution of the inputs, $P_{\mathbb{X}}$, is the whole $\mathbb{X}$, furthermore, $P_{\mathbb{X}}$ is *absolutely continuous*.

Let us introduce functions, for $i = 0, \dots, m-1$, as

$$f^{(i)}_{\theta,n}(x) \doteq \frac{1}{k_n} \sum_{j=1}^{n} y_{i,j}(\theta)\, \mathbb{I}\big(\, x_j \in N(x, k_n)\,\big), \qquad (17)$$

where $\mathbb{I}$ is an indicator function (its value is 1 if its argument is true, and 0 otherwise), $N(x, k_n)$ denotes the k_n closest neighbors of x from $\{x_j\}_{j=1}^n$, and $k_n \leq n$ is a constant (window size), which can depend on n. We use the standard Euclidean distance as a metric on $\mathbb{X}$ (to define neighbors). Since the inputs, $\{x_j\}$, have a distribution that is absolutely continuous, there is zero probability of ties in $N(x, n_k)$.

Let us denote the $\mathcal{L}^2(\mathbb{X})$ distance of the tested function, f_θ, from its empirical estimate based on sample $\mathcal{D}_i(\theta)$ by

$$Z^{(i)}_n(\theta) \doteq \| f^{(i)}_{\theta,n} - f_\theta \|_2^2 = \int_{\mathbb{X}} \big(f^{(i)}_{\theta,n}(x) - f_\theta(x)\big)^2 dx. \quad (18)$$

We can define the *rank* of $Z^{(0)}_n(\theta)$ among $\{Z^{(i)}_n(\theta)\}$ as

$$\mathcal{R}_n(\theta) \doteq 1 + \sum_{i=1}^{m-1} \mathbb{I}\big(\, Z^{(0)}_n(\theta) \prec_\pi Z^{(i)}_n(\theta)\,\big), \qquad (19)$$

where $\mathbb{I}$ is an indicator function, and binary relation "$\prec_\pi$" is the standard "$<$" with random tie-breaking. More precisely, as before, 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 $\pi(k) < \pi(j)$.

Therefore, in case of Algorithm I, the ranking function is

$$\psi\big(\, \mathcal{D}^\pi_0, \{\mathcal{D}^\pi_k(\theta)\}_{k\neq 0}\,\big) = \mathcal{R}_n(\theta). \qquad (20)$$

It can be shown that for any fixed false parameter θ, $Z^{(0)}_n(\theta)$ tends to have the largest rank as the sample size increases, thus, we fix $p = 1$ and only exclude parameters which lead to high ranks. That is, using (14), the confidence set is

$$\Theta^{(1)}_{\varrho,n} \doteq \big\{\, \theta \in \Theta :\, \mathcal{R}_n(\theta) \leq q \,\big\}, \qquad (21)$$

where $\varrho \doteq (m, q)$ are hyper-parameters with $1 \leq q < m$.

The main theoretical results can be summarized as

Theorem 2: The coverage probability of the region is

$$\mathbb{P}\big(\, \theta^* \in \Theta^{(1)}_{\varrho,n}\,\big) = q\,/\,m, \qquad (22)$$

for any sample size n. Moreover, if $\{k_n\}$ are chosen such that $k_n \to \infty$ and $k_n/n \to 0$, as $n \to \infty$, then the confidence regions are strongly consistent, as defined by (16)*.*

Regarding the computation aspects of Algorithm I note that $\{f^{(i)}_{\theta,n}\}$ can be calculated exactly based on the available data, as they are piece-wise constant functions. The distance $\| f^{(i)}_{\theta,n} - f_\theta \|_2^2$ can also be calculated from the available data. Nevertheless, one may use the Monte Carlo approximation

$$\| f^{(i)}_{\theta,n} - f_\theta \|_2^2 \approx \frac{1}{\ell_n} \sum_{k=1}^{\ell_n} \big(f^{(i)}_{\theta,n}(\bar{x}_k) - f_\theta(\bar{x}_k)\big)^2, \qquad (23)$$

where ℓ_n is a constant and $\{\bar{x}_k\}$ are i.i.d. random variables having uniform distribution on $\mathbb{X}$. Note that we know from the *strong law of large numbers* (SLLN) that the sum in (23) almost surely converges to $\| f^{(i)}_{\theta,n} - f_\theta \|_2^2$, as $\ell_n \to \infty$.

It is relatively easy to see that using the approximation in (23) does not affect the *exact* coverage probability of the algorithm. Moreover, if $\ell_n \to \infty$ as $n \to \infty$, then one can also show the strong consistency of the approximated variant. Hence, the theoretical properties of Theorem 2 remain valid, but the sizes of regions are affected by the approximation.

The kNN estimator, which is in the core of Algorithm I, is a simple kernel method that uses a variable bandwidth rectangular window. A natural generalization of this approach is to apply other kernels for local averaging. Given any kernel $k(\cdot,\cdot)$, e.g., Gaussian, we can redefine functions $\{f^{(i)}_{\theta,n}\}$ as

$$f^{(i)}_{\theta,n}(x) \doteq \frac{1}{\sum_{l=1}^n k(x, x_l)} \sum_{j=1}^{n} y_{i,j}(\theta)\, k(x, x_j), \qquad (24)$$

which leads to alternative confidence region constructions.

These variants typically also build confidence regions with *exact* coverage probabilities. Moreover, as a wide variety of such kernel estimates are strongly consistent, under some technical conditions [4], and the generalized Algorithm I inherits these properties, the resulting confidence sets are also *strongly consistent*. The corresponding coverage and consistency theorems could be stated analogously to Theorem 2.

B. Algorithm II (Embedding Based)

The core idea of Algorithm II is to embed the distribution of the original sample and that of the alternative ones in an RKHS using a characteristic kernel. If the underlying distributions are different, then the original dataset results in a different element than the one the alternative datasets are being mapped to, which can be detected statistically.

For Algorithm II, let $(\mathbb{X}, \mathcal{X})$ be a measurable space, and $\mathcal{H}$ be a *separable* RKHS containing $\mathbb{S} \to \mathbb{R}$ type functions with a (measurable) *bounded* and *characteristic* kernel. If $\mathbb{X} = \mathbb{R}^d$, then $\mathbb{S} = \mathbb{R}^d \times \{+1, -1\}$, and we can use, e.g., Gaussian or Laplacian kernels, which are characteristic [10].

Let us introduce the following kernel mean embeddings

$$h_*(\cdot) \doteq \mathbb{E}\big[\, k(\cdot, S_*)\,\big] \quad \text{and} \quad h_\theta(\cdot) \doteq \mathbb{E}\big[\, k(\cdot, S_\theta)\,\big], \quad (25)$$

where S_* and S_θ are a random elements from $\mathbb{S}$; variable S_* has the "true" distribution of the observations, while S_θ has a distribution where the output, Y, is generated according to the conditional probability (10), parametrized by θ, while the marginal distribution of the input, X, remains the same.

Since the kernel is bounded, $\mathbb{E}\big[\,\sqrt{k(S_\theta, S_\theta)}\,\big] < \infty$, for all θ, which ensures that $\{h_\theta\}$ exist and belong to $\mathcal{H}$ [10].

Because the kernel is characteristic, we know that $h_\theta = h_*$ if and only if $\theta = \theta^*$. Now, let us introduce the following empirical versions of the embedded distributions,

$$h^{(i)}_{\theta,n}(\cdot) \doteq \frac{1}{n}\sum_{j=1}^{n} k(\cdot, s_{i,j}(\theta)), \quad (26)$$

for $i = 0, \ldots, m-1$, where $s_{i,j}(\theta) \doteq (x_j, y_{i,j}(\theta))$; and recall that for $i = 0$ (original sample), we have $y_{i,j}(\theta) = y_j$. In other words, $s_{i,j}(\theta)$ has the same distribution as S_θ for $i \neq 0$ and its distribution is the same as that of S_* for $i = 0$.

Let $\beta \in \mathbb{R}$ be a constant that satisfies $|\, k(x,y)\,| \leq \beta$ for all x, y. Then, obviously $|\, h_\theta(x)\,| \leq \beta$ for all x, as well. Now, applying the reproducing property, we have the bound

$$\begin{aligned}
\mathrm{Var}\big(k(\cdot,S)\big) &= \mathbb{E}\big[\,\|\, k(\cdot,S) - h(\cdot)\,\|^2_{\mathcal{H}}\,\big] \\
&\leq \mathbb{E}\big[\,\|\, k(\cdot,S)\,\|^2_{\mathcal{H}}\,\big] + \mathbb{E}\big[\,\|\, h(\cdot)\,\|^2_{\mathcal{H}}\,\big] \\
&\quad + 2\,\mathbb{E}\big[\,|\,\langle\, k(\cdot,S), h(\cdot)\,\rangle_{\mathcal{H}}\,|\,\big] \\
&\leq \mathbb{E}\big[\,\|\, k(\cdot,S)\,\|^2_{\mathcal{H}}\,\big] + \|\, h(\cdot)\,\|^2_{\mathcal{H}} + 2\,\mathbb{E}\big[\,|\, h(S)\,|\,\big] \\
&\leq \mathbb{E}\big[\,\langle\, k(\cdot,S), k(\cdot,S)\,\rangle_{\mathcal{H}}\,\big] + \|\, h(\cdot)\,\|^2_{\mathcal{H}} + 2\,\beta \\
&= \mathbb{E}\big[\, k(S,S)\,\big] + \|\, h(\cdot)\,\|^2_{\mathcal{H}} + 2\,\beta \\
&\leq 3\,\beta + \|\, h(\cdot)\,\|^2_{\mathcal{H}} < \infty, \qquad (27)
\end{aligned}$$

where S is either S_* or S_θ, and $h(\cdot) \doteq \mathbb{E}\big[\, k(\cdot, S)\,\big]$.

Then, we know from the SLLN for Hilbert space valued elements that $\|\, h^{(i)}_{\theta,n} - h_\theta\,\|_{\mathcal{H}} \xrightarrow{a.s.} 0$, as $n \to \infty$, for $i \neq 0$, and we also have that $\|\, h^{(0)}_{\theta,n} - h_*\,\|_{\mathcal{H}} \xrightarrow{a.s.} 0$, as $n \to \infty$.

Now, we can define the $\{Z^{(i)}_n(\theta)\}$ variables as

$$Z^{(i)}_n(\theta) \doteq \sum_{j=0}^{m-1} \|\, h^{(i)}_{\theta,n} - h^{(j)}_{\theta,n}\,\|^2_{\mathcal{H}}, \quad (28)$$

i.e., the total (cumulative) distance of $h^{(i)}_{\theta,n}$ from all other functions. Then, we can construct the confidence set as (21).

Theorem 3: The confidence regions of Algorithm II have

$$\mathbb{P}\big(\,\theta^* \in \Theta^{(2)}_{\varrho,n}\,\big) = q\,/\,m, \quad (29)$$

for all n; and they are strongly consistent if $m > 2$.

The squared distance of the empirical versions of the embeddings $\|\, h^{(i)}_{\theta,n} -, h^{(j)}_{\theta,n}\,\|^2_{\mathcal{H}}$ can be computed by applying the reproducing property of the kernel and the Gram matrix of the sample $s_{i,1}(\theta), \ldots, s_{i,n}(\theta), s_{j,1}(\theta), \ldots, s_{j,n}(\theta)$.

Algorithm II has a nice theoretical interpretation as comparing embedded distributions in an RKHS. However, as the Gram matrices required to compute the $\{Z^{(i)}_n(\theta)\}$ variables depend on θ, this method has a large computational burden, hence the importance of Algorithm II is mainly theoretical. Nevertheless, motivated by its ideas, in the next section we suggest a computationally much lighter algorithm.

C. Algorithm III (Discrepancy Based)

Algorithm III follows the intuitions behind Algorithm II, but ensures that we can work with the same Gram matrix for all θ. Moreover, it has a simpler construction for $\{Z^{(i)}_n(\theta)\}$, which also makes it computationally more appealing.

For Algorithm III, let $(\mathbb{X}, d)$ be a *compact* Polish metric space (i.e., complete and separable), and assume that each $f \in \mathcal{F}$ is *continuous* (additionally to the assumptions of Section III). Let $\mathcal{H}$ be a *separable* RKHS containing $\mathbb{X} \to \mathbb{R}$ functions with a (measurable) *bounded* and *universal* kernel.

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

The fundamental objects of Algorithm III are

$$Z^{(i)}_n(\theta) \doteq \left\|\, \frac{1}{n}\sum_{j=1}^{n} \varepsilon_{i,j}(\theta)\, k(\cdot, x_j)\,\right\|^2_{\mathcal{H}}, \quad (30)$$

for $i = 0, \ldots, m-1$. Observe that $Z^{(i)}_n(\theta)$ can be easily computed using the Gram matrix $K_{i,j} \doteq k(x_i, x_j)$, as

$$Z^{(i)}_n(\theta) = \frac{1}{n^2}\, \varepsilon_i^{\mathrm{T}}(\theta)\, K\, \varepsilon_i(\theta), \quad (31)$$

using the notation $\varepsilon_i(\theta) \doteq (\varepsilon_{i,1}(\theta), \ldots, \varepsilon_{i,n}(\theta))^{\mathrm{T}}$.

From this point, we follow the construction of Algorithms I and II, namely, we define the ranking function as (19), and the confidence region as (21), but naturally we apply our new functions (30) as the definition of the $\{Z^{(i)}_n(\theta)\}$ variables.

Theorem 4: The confidence regions of Algorithm III have

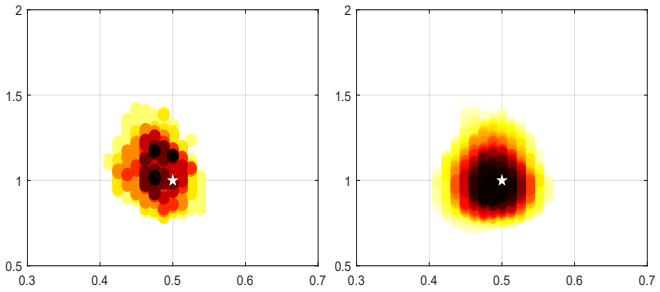
$$\mathbb{P}\big(\,\theta^* \in \Theta^{(3)}_{\varrho,n}\,\big) = q\,/\,m, \quad (32)$$

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

V. Numerical Experiments

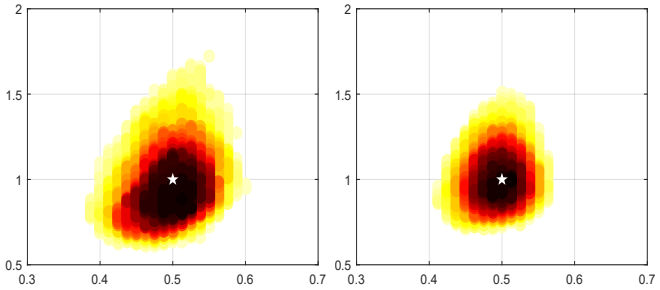
Numerical experiments were carried out to demonstrate the proposed algorithms. In the presented test scenario the joint probability distribution of the data was assumed to be the mixture of two Laplace distributions with different locations, μ_1, μ_2, but with the same scale λ. It was assumed that with probability p we observe the "$+1$" class, and with $1-p$ we see an element of the "-1" class. Selecting p, μ_1, μ_2 and λ induces a regression function, e.g., see (8).

The confidence regions were built for parameters p and λ, while the location parameters were fixed, $\mu_1 = 1$ and $\mu_2 = -1$, to allow two dimensional figures. Figure 1 demonstrates the obtained ranks, $\{\mathcal{R}_n(\theta)\}$, for various $\theta = (p, \lambda)$ using (a) Algorithm I with kNN (15 neighbors), (b) Algorithm I with a kernel, (c) Algorithm II, and (d) Algorithm III. The kernel was always Gaussian with $\sigma = 0.125$. Darker colors indicate smaller ranks, therefore, the darker the color is, the more likely the parameter is included in a confidence region.



(a) Algorithm I (kNN) Ranks

(b) Algorithm I (Gauss) Ranks



(c) Algorithm II (Gauss) Ranks

(d) Algorithm III (Gauss) Ranks

Fig. 1. The ranks of the reference element for various parameters, indicating how likely they are included in a confidence set. The model was a mixture of Laplace distributions. The true mixing probability $p^* = 1/2$ (x-axis) and the scale parameter $\lambda^* = 1$ (y-axis), they are denoted by a "$\star$", were estimated from $n = 500$ observations. The kernel was Gaussian with $\sigma = 0.125$.

The true parameter was $\theta^* = (p^*, \lambda^*)$ with $p^* = 1/2$ (x-axis) and $\lambda^* = 1$ (y-axis). The sample size was $n = 500$ and $m = 40$ samples were used. The regions were evaluated on a grid. The algorithms provide comparable confidence sets with Algorithm I (kNN) being the best for the current setup.

Note that in this special example it is possible to construct individual confidence regions for true parameter values p^* and λ^* based on standard results. One can use, e.g., Hoeffding's inequality to get confidence intervals for probability p^*, and λ^* can be estimated based on the fact that the variance of the observations, for both classes, is $2(\lambda^*)^2$. Nevertheless, such approaches need the *specific interpretations* of the parameters: on how they influence the observations. Furthermore, even in this very special case it is not obvious how to construct a *joint* confidence region for (p^*, λ^*). Simply intersecting the two confidence tubes (i.e., if we extend the confidence intervals for p^* and λ^* to $\mathbb{R}^2$, they define two infinite "stripes", a vertical and a horizontal one) produces a set with a lower confidence than that of the original sets, and hence it ultimately leads to *conservative* regions.

On the other hand, the suggested three algorithms do not presuppose any interpretation of the tested parameters, apart from the fact that they determine a regression function. They do not need a fully parametrized joint distribution, indeed, the regression function is compatible with infinitely many joint distributions having widely different (marginal) input distributions. Furthermore, if $\theta^* \in \mathbb{R}^d$, then the algorithms automatically build *joint* and *non-conservative* confidence sets. Hence, another advantage of the presented framework, apart from its strong theoretical guarantees, is its *flexibility*.

VI. Conclusions

In this paper we addressed the problem of building *non-asymptotic* confidence regions for the *regression* function of binary classification, which is a key object defined as the conditional expectation of the class labels given the inputs.

The main idea was to test candidate models by generating *alternative samples* based on them, and then computing the performance of a kernel-based algorithm on all samples. If the candidate model is wrong, then the algorithm behave differently on the alternatively generated samples than on the original one, which can be detected statistically by *ranking*.

Three constructions were proposed and we argued that all of them build confidence regions with *exact* coverage probabilities, for any sample size, and are *strongly consistent*. The rigorous proofs will be available in an extended paper.

The proposed framework is *semi-parametric*, because the regression function does not determine the (joint) probability distribution of the data, it does not contain information about the (marginal) distribution of the inputs (and that is why only the outputs are resampled in the alternative datasets).

Note that the introduced algorithms are not restricted to specific classes of regressor functions, they can work with any such function, as its primary role is to generate alternative (perturbed) datasets. Consequently, the family of regression functions can be *arbitrary*. It could even be the set of all possible regression functions which satisfy (7) and the theoretical results are still valid. If we work with an infinite dimensional class of functions, then of course the confidence regions cannot be explicitly constructed in practice. Nevertheless, it is still possible to *test* any candidate regression function to check whether it is included in a confidence set, or in other words, to *quantify its uncertainty* by computing how compatible it is with the available observations.

References

[1] V. N. Vapnik, *Statistical Learning Theory*. Wiley-Interscience, 1998.

[2] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *The Annals of Statistics*, pp. 1171–1220, 2008.

[3] A. Goudjil, M. Pouliquen, E. Pigeon, O. Gehan, and M. M'Saad, "Identification of systems using binary sensors via support vector machines," in *54th IEEE Conference on Decision and Control (CDC), Osaka, Japan*, pp. 3385–3390, 2015.

[4] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

[5] A. Carè, B. Cs. Csáji, M. Campi, and E. Weyer, "Finite-sample system identification: An overview and a new correlation method," *IEEE Control Systems Letters*, vol. 2, no. 1, pp. 61 – 66, 2018.

[6] B. Cs. Csáji, M. C. Campi, and E. Weyer, "Sign-Perturbed Sums: A new 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.

[7] S. Kolumbán, *System Identification in Highly Non-Informative Environment*. PhD thesis, Budapest University of Technology and Economics, Hungary, and Vrije Univesiteit Brussels, Belgium, 2016.

[8] G. Pillonetto, A. Carè, and M. C. Campi, "Kernel-based SPS," in *Proceedings of the 18th IFAC Symposium on System Identification (SYSID 2018), Stockholm, Sweden, July 9-11, 2018*, pp. 31–36, Elsevier, 2018.

[9] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, no. 3, pp. 337–404, 1950.

[10] K. Muandet, K. Fukumizu, B. Sriperumbudur, and B. Schölkopf, "Kernel mean embedding of distributions: A review and beyond," *Foundations and Trends in Machine Learning*, pp. 1–141, 2017.

[11] R. L. Taylor, *Stochastic Convergence of Weighted Sums of Random Elements in Linear Spaces*, vol. 672. Springer, 1978.