Exact Distribution-Free Hypothesis Tests for the Regression Function of Binary Classification via Conditional Kernel Mean Embeddings

Ambrus Tamás[1,2] Balázs Csanád Csáji[1,2]

***Abstract*— In this paper we suggest two statistical hypothesis tests for the regression function of binary classification based on conditional kernel mean embeddings. The regression function is a fundamental object in classification as it determines both the Bayes optimal classifier and the misclassification probabilities.**

A resampling based framework is presented and combined with consistent point estimators of the conditional kernel mean map, in order to construct distribution-free hypothesis tests. These tests are introduced in a flexible manner allowing us to control the exact probability of type I error for any sample size. We also prove that both proposed techniques are consistent under weak statistical assumptions, namely, the type II error probabilities pointwise converge to zero.

I. Introduction

Binary classification [1] is a central problem in supervised learning with a lot of crucial applications, for example, in quantized system identification, signal processing and fault detection. Kernel methods [2] offer a wide range of tools to draw statistical conclusions by embedding datapoints and distributions into a (possibly infinite dimensional) reproducing kernel Hilbert space (RKHS), where we can take advantage of the geometrical structure. These nonparametric methods often outperform the standard parametric approaches [3]. A key quantity, for example in model validation, is the conditional distribution of the outputs given the inputs. A promising way to handle such conditional distributions is to apply conditional kernel mean embeddings [4] which are input dependent elements of an RKHS. In this paper we introduce distribution-free hypothesis tests for the regression function of binary classification based on these conditional embeddings. Such distribution-free guarantees are of high importance, since our knowledge on the underlying distributions is often limited in practical applications.

Let $(\mathbb{X}, \mathcal{X})$ be a measurable input space, where $\mathcal{X}$ is a σ-field on $\mathbb{X}$, and let $\mathbb{Y} = \{-1, 1\}$ be the output space. In binary classification we are given an independent and identically distributed (i.i.d.) sample $\{(X_i, Y_i)\}_{i=1}^n$ from an unknown distribution $P = P_{X,Y}$ on $\mathcal{X} \otimes \mathcal{Y}$. Measurable $\mathbb{X} \to \mathbb{Y}$ functions are called classifiers. Let $\mathbf{L} : \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}^+$ be a nonnegative measurable loss function. In this paper we restrict our attention to the archetypical $0/1$-loss given by the indicator $\mathbf{L}(y_1, y_2) \doteq \mathbb{I}(y_1 \neq y_2)$ for $y_1, y_2 \in \mathbb{Y}$. In general, our aim is to minimize the Bayes risk, which is $R(\phi) \doteq \mathbb{E}[\,\mathbf{L}(\phi(X), Y)\,]$ for classifier ϕ, i.e., the expected loss. It is known that for the $0/1$-loss, the Bayes risk is the misclassification probability $R(\phi) = \mathbb{P}(\,\phi(X) \neq Y\,)$ and a risk minimizer (P_X-a.e.) equals to the sign of the regression function $f_*(x) \doteq \mathbb{E}[\,Y \,|\, X = x\,]$, i.e., classifier[1] $\phi_*(x) = \operatorname{sign}\big(f_*(x)\big)$ reaches the optimal risk [1, Theorem 2.1]. It can also be proved that the conditional distribution of Y given X is encoded in f_* for binary outputs.

One of the main challenges in statistical learning is that distribution P is unknown, therefore the true risk cannot be directly minimized, only through empirical estimates [5]. Vapnik's theory quantifies the rate of convergence for several approaches (empirical and structural risk minimization), but these bounds are usually conservative for small samples. The literature is rich in efficient point estimates, but there is a high demand for distribution-free uncertainty quantification.

It is well-known that hypothesis tests are closely related to confidence regions. Distribution-free confidence regions for classification received considerable interest, for example, Sadinle *et al.* suggested set-valued estimates with guaranteed coverage confidence [6], Barber studied the limitations of such distribution-free estimation methods [7], while Gupta *et al.* analyzed score based classifiers and the connection of calibration, confidence intervals and prediction sets [8].

Our main contribution is that, building on the distribution-free resampling framework of [9] which was motivated by finite-sample system identification methods [10], we suggest conditional kernel mean embeddings based ranking functions to construct hypothesis tests for the regression function of binary classification. These tests have exact non-asymptotic guarantees for the probability of type I error and have strong asymptotic guarantees for the type II error probabilities.

II. Reproducing Kernel Hilbert Spaces

A. Real-Valued Reproducing Kernel Hilbert Spaces

Let $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ be a symmetric and positive-definite kernel, i.e., for all $n \in \mathbb{N}$, $x_1, \ldots, x_n \in \mathbb{X}, a_1, \ldots, a_n \in \mathbb{R}$:

$$\sum_{i,j=1}^{n} k(x_i, x_j)\, a_i\, a_j \geq 0. \tag{1}$$

Equivalently, kernel (or Gram) matrix $K \in \mathbb{R}^{n \times n}$, where $K_{i,j} \doteq k(x_i, x_j)$ for all $i, j \in [n] \doteq \{1, \ldots, n\}$, is required

*The research was supported by the Ministry of Innovation and Technology of Hungary NRDI Office within the framework of the Artificial Intelligence National Laboratory Program. Prepared with the professional support of the Doctoral Student Scholarship Program of the Cooperative Doctoral Program of the Ministry of Innovation and Technology financed from the National Research, Development and Innovation Fund.

[1]A. Tamás and B. Cs. Csáji are with SZTAKI: Institute for Computer Science and Control, Eötvös Loránd Research Network, Budapest, Hungary, ambrus.tamas@sztaki.hu, csaji@sztaki.hu

[2]A. Tamás and B. Cs. Csáji are also with the Institute of Mathematics, Eötvös Loránd University (ELTE), Budapest, Hungary, H-1117.

[1]Let the sign function be defined as $\operatorname{sign}(x) = \mathbb{I}(x \geq 0) - \mathbb{I}(x < 0)$.

to be positive semidefinite. Let $\mathcal{F}$ denote the corresponding reproducing kernel Hilbert space containing $\mathbb{X} \to \mathbb{R}$ functions, see [2], where $k_x(\cdot) = k(\cdot, x) \in \mathcal{F}$ and the reproducing property, $f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{F}}$, holds for all $x \in \mathbb{X}$ and $f \in \mathcal{F}$. Let $l : \mathbb{Y} \times \mathbb{Y} \to \mathbb{R}$ denote a symmetric and positive-definite kernel and let $\mathcal{G}$ be the corresponding RKHS.

B. Vector-Valued Reproducing Kernel Hilbert Spaces

The definition of conditional kernel mean embeddings [4] requires a generalization of real-valued RKHSs [11], [12].

Definition 1: Let $\mathcal{H}$ be a Hilbert space of $\mathbb{X} \to \mathcal{G}$ type functions with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, where $\mathcal{G}$ is a Hilbert space. $\mathcal{H}$ is a *vector-valued RKHS* if for all $x \in \mathbb{X}$ and $g \in \mathcal{G}$ the linear functional (on $\mathcal{H}$) $h \mapsto \langle g, h(x)\rangle_{\mathcal{G}}$ is bounded.

Then by the Riesz representation theorem, for all $(g, x) \in \mathcal{G} \times \mathbb{X}$ there exists a unique $\tilde{h} \in \mathcal{H}$ for which $\langle g, h(x)\rangle_{\mathcal{G}} = \langle \tilde{h}, h\rangle_{\mathcal{H}}$. Let Γ_x be a $\mathcal{G} \to \mathcal{H}$ operator defined as $\Gamma_x g = \tilde{h}$. The notation is justified because Γ_x is linear. Further, let $\mathcal{L}(\mathcal{G})$ denote the bounded linear operators on $\mathcal{G}$ and let $\Gamma : \mathbb{X} \times \mathbb{X} \to \mathcal{L}(\mathcal{G})$ be defined as $\Gamma(x_1, x_2)g \doteq (\Gamma_{x_2} g)(x_1) \in \mathcal{G}$. We will use the following result [11, Proposition 2.1]:

Proposition 1: Operator Γ satisfies for all $x_1, x_2 \in \mathbb{X}$:

1) $\forall g_1, g_2 \in \mathcal{G} : \langle g_1, \Gamma(x_1, x_2) g_2\rangle_{\mathcal{G}} = \langle \Gamma_{x_1} g_1, \Gamma_{x_2} g_2\rangle_{\mathcal{G}}$.
2) $\Gamma(x_1, x_2) \in \mathcal{L}(\mathcal{G})$, $\Gamma(x_1, x_2) = \Gamma^*(x_2, x_1)$, and for all $x \in \mathbb{X}$ operator $\Gamma(x, x)$ is positive.
3) For all $n \in \mathbb{N}$, $\{(x_i)\}_{i=1}^n \subseteq \mathbb{X}$ and $\{(g_j)\}_{j=1}^n \subseteq \mathcal{G}$:

$$\sum_{i,j=1}^{n} \langle g_i, \Gamma(x_i, x_j) g_j\rangle_{\mathcal{G}} \geq 0. \tag{2}$$

When properties $1) - 3)$ hold, we call Γ a *vector-valued reproducing kernel*. Similarly to the classical Moore-Aronszjan theorem [13, Theorem 3], for any kernel Γ, there uniquely exists (up to isometry) a vector-valued RKHS, having Γ as its reproducing kernel [11, Theorem 2.1].

III. Kernel Mean Embeddings

A. Kernel Means of Distributions

Kernel functions with a fixed argument are feature maps, i.e., they represent input points from $\mathbb{X}$ in Hilbert space $\mathcal{F}$ by mapping $x \mapsto k(\cdot, x)$. Let X be a random variable with distribution P_X, then $k(\cdot, X)$ is a random element in $\mathcal{F}$. The kernel mean embedding of distribution P_X is defined as $\mu_X \doteq \mathbb{E}[\, k(\cdot, X)\,]$, using a Bochner integral [14].

It can be proved that if kernel k is measurable as well as $\mathbb{E}\big[\, \sqrt{k(X, X)}\,\big] < \infty$ holds, then the kernel mean embedding of P_X exists and it is the representer of the bounded, linear expectation functional w.r.t. X, therefore $\mu_X \in \mathcal{F}$ and we have $\langle f, \mu_X\rangle_{\mathcal{F}} = \mathbb{E}\big[\, f(X)\,\big]$ for all $f \in \mathcal{F}$ [15]. Similarly, for variable Y let μ_Y be the kernel mean embedding of P_Y.

B. Conditional Kernel Mean Embeddings

If the kernel mean embedding of P_Y exists, then $l_Y \doteq l(\circ, Y) \in L_1(\Omega, \mathcal{A}, \mathbb{P}; \mathcal{G})$, that is l_Y is a Bochner integrable $\mathcal{G}$-valued random element, hence for all $\mathcal{B} \subseteq \mathcal{A}$ the conditional expected value can be defined. Let $\mathcal{B} \doteq \sigma(X)$ be the σ-field generated by random element X, then the conditional kernel mean embedding of $P_{Y|X}$ in RKHS $\mathcal{G}$ is defined as

$$\mu_{Y|X} = \mu_*(X) \doteq \mathbb{E}\big[\, l(\circ, Y) \,|\, X\,\big], \tag{3}$$

see [16], where μ_* is a P_X-a.e. defined (measurable) conditional kernel mean map. It is easy to see that for all $g \in \mathcal{G}$

$$\mathbb{E}\big[\, g(Y) \,|\, X\,\big] \overset{\text{a.s.}}{=} \langle\, g, \mathbb{E}[\, l(\circ, Y) \,|\, X\,]\,\rangle_{\mathcal{G}}, \tag{4}$$

showing that this approach is equivalent to the definition in [12]. We note that the original paper [4] introduced conditional mean embeddings as $\mathcal{F} \to \mathcal{G}$ type operators. The presented approach is more natural and has theoretical advantages as its existence and uniqueness is usually ensured.

C. Empirical Estimates of Conditional Kernel Mean Maps

The advantage of dealing with kernel means instead of the distributions is that we can use the structure of the Hilbert space. In statistical learning, the underlying distributions are unknown, thus their kernel mean embeddings are needed to be estimated. A typical assumption for classification is that:

A0 Sample $\mathcal{D}_0 \doteq \{(X_i, Y_i)\}_{i=1}^n$ is i.i.d. with distribution P.

The empirical estimation of conditional kernel mean map $\mu_* : \mathbb{X} \to \mathcal{G}$ is challenging in general, because its dependence on $x \in \mathbb{X}$ can be complex. The standard approach defines estimator $\widehat{\mu}$ as a regularized empirical risk minimizer in a vector-valued RKHS, see [12], which is equivalent to the originally proposed operator estimates in [4].

By (4) it is intuitive to estimate μ_* with a minimizer of the following objective over some space $\mathcal{H}$ [12, Equation 5]:

$$\mathcal{E}(\mu) = \sup_{\|g\|_{\mathcal{G}} \leq 1} \mathbb{E}\big[\, \big|\, E[g(Y) \,|\, X] - \langle g, \mu(X)\rangle_{\mathcal{G}}\,\big|^2\,\big]. \tag{5}$$

Since $\mathbb{E}[\, g(Y) \,|\, X\,]$ is not observable, the authors of [12] have introduced the following surrogate loss function:

$$\mathcal{E}_s(\mu) \doteq \mathbb{E}\Big[\, \|l(\circ, Y) - \mu(X)\|_{\mathcal{G}}^2\,\Big]. \tag{6}$$

It can be shown [12] that $\mathcal{E}(\mu) \leq \mathcal{E}_s(\mu)$ for all $\mu : \mathbb{X} \to \mathcal{G}$, moreover under suitable conditions [12, Theorem 3.1] the minimizer of $\mathcal{E}(\mu)$ P_X-a.s. equals to the minimizer of $\mathcal{E}_s(\mu)$, hence the surrogate version can be used. The main advantage of $\mathcal{E}_s$ is that it can be estimated empirically as:

$$\widehat{\mathcal{E}}_s(\mu) \doteq \frac{1}{n} \sum_{i=1}^{n} \|l(\circ, Y_i) - \mu(X_i)\|_{\mathcal{G}}^2\,. \tag{7}$$

To make the problem tractable, we minimize (7) over a vector-valued RKHS, $\mathcal{H}$. There are several choices for $\mathcal{H}$. An intuitive approach is to use the space induced by kernel $\Gamma(x_1, x_2) \doteq k(x_1, x_2)\mathbf{Id}_{\mathcal{G}}$, where $x_1, x_2 \in \mathbb{X}$ and $\mathbf{Id}_{\mathcal{G}}$ is the identity map on $\mathcal{G}$. Henceforth, we will focus on this kernel, as it leads to the same estimator as the one proposed in [4]. A regularization term is also used to prevent overfitting and to ensure well-posedness, hence, the estimator is defined as

$$\widehat{\mu} = \widehat{\mu}_{\mathcal{D}} = \widehat{\mu}_{n,\lambda} \doteq \operatorname*{arg\,min}_{\mu \in \mathcal{H}} \widehat{\mathcal{E}}_\lambda(\mu) \tag{8}$$

where $\widehat{\mathcal{E}}_\lambda(\mu) \doteq \big[\widehat{\mathcal{E}}_s(\mu) + {}^{\lambda}\!/_{n}\, \|\mu\|_{\mathcal{H}}^2\,\big]$. An explicit form of $\widehat{\mu}$ can be given by [11, Theorem 4.1] (cf. representer theorem):

Theorem 1: If $\widehat{\mu}$ minimizes $\mathcal{E}_\lambda$ in $\mathcal{H}$, then it is unique and admits the form of $\widehat{\mu} = \sum_{i=1}^n \Gamma_{X_i} c_i$, where coefficients $\{(c_i)\}_{i=1}^n$, $c_i \in \mathcal{G}$ for $i \in [n]$, are the unique solution of

$$\sum_{j=1}^{n} \big(\Gamma(X_i, X_j) + \lambda \mathbb{I}(i=j) \mathbf{Id}_{\mathcal{G}} \big) c_j = l(\circ, Y_i) \quad \text{for} \;\; i \in [n].$$

By Theorem 1 we have: $c_i = \sum_{j=1} W_{i,j} l(\circ, Y_j)$ for $i \in [n]$, with $W = (K + \lambda \mathbf{I})^{-1}$, where $\mathbf{I}$ is the identity matrix.

IV. Distribution-Free Hypothesis Tests

For binary classification, one of the most intuitive kernels on the output space is $l(y_1, y_2) \doteq \mathbb{I}(y_1 = y_2)$ for $y_1, y_2 \in \mathbb{Y}$, which is called the naïve kernel. It is easy to prove that l is symmetric and positive definite. Besides, we can describe its induced RKHS $\mathcal{G}$ as $\{a_1 \cdot l(\circ, 1) + a_2 \cdot l(\circ, -1) \,|\, a_1, a_2 \in \mathbb{R}\}$. Hereafter, l will denote this kernel for the output space.

A. Resampling Framework

We consider the following hypotheses:

$$H_0: \;\; f_* = f \;\; (P_X\text{-a.s.}) \;\; \text{and} \;\; H_1 : \neg H_0 \tag{9}$$

for a given candidate regression function f, where $\neg H_0$ denotes the negation of H_0. For the sake of simplicity, we will use the slightly inaccurate notation $f_* \neq f$ for H_1, which refers to inequality in the $L_2(P_X)$-sense. To avoid misunderstandings, we will call f the "candidate" regression function and f_* the "true" regression function.

One of our main observations is that in binary classification the regression function determines the conditional distribution of Y given X, i.e., by [1, Theorem 2.1] we have

$$\mathbb{P}_*(Y = 1 \,|\, X) = \frac{f_*(X) + 1}{2} = p_*(X). \tag{10}$$

Notation $\mathbb{P}_*$ is introduced to emphasize the dependence of the conditional distribution on f_*. Similarly, candidate function f can be used to determine a conditional distribution given X. Let $\bar{Y}$ be such that $\mathbb{P}_f(\bar{Y} = 1 \,|\, X) = (f(X) + 1)\,/\,2 = p(X)$. Observe that if H_0 is true, then (X, Y) and $(X, \bar{Y})$ have the same joint distribution, while when H_1 holds, then $\bar{Y}$ has a "different" conditional distribution w.r.t. X than Y. Our idea is to imitate sample $\mathcal{D}_0 = \{(X_i, Y_i)\}$ by generating alternative outputs for the original input points from the conditional distribution induced by the candidate function f, i.e., let $m > 1$ be a user-chosen integer and let us define samples

$$\mathcal{D}_j \doteq \{(X_i, \bar{Y}_{i,j})\}_{i=1}^n \quad \text{for} \quad j \in [m-1]. \tag{11}$$

An uninvolved way to produce $\bar{Y}_{i,j}$ for $i \in [n]$, $j \in [m-1]$ is as follows. We generate i.i.d. uniform variables from $(-1, 1)$. Let these be $U_{i,j}$ for $i \in [n]$ and $j \in [m-1]$. Then we take

$$\bar{Y}_{i,j} = \mathbb{I}(U_{i,j} \leq f(X_i)) - \mathbb{I}(U_{i,j} > f(X_i)), \tag{12}$$

for $(i, j) \in [n] \times [m-1]$. The following remark highlights one of the main advantages of this scheme.

Remark 1: If H_0 holds, then $\{(\mathcal{D}_j)\}_{j=0}^{m-1}$ are conditionally i.i.d. w.r.t. $\{(X_i)\}_{i=1}^n$, hence they are also exchangeable.

The suggested distribution-free hypothesis tests are carried out via rank statistics as described in [9], where our resampling framework for classification was first introduced. That is, we define a suitable ordering on the samples and accept the nullhypothesis when the rank of the original sample is not "extremal" (neither too low nor too high), i.e., the original sample does not differ significantly from the alternative ones. More abstractly, we define our tests via ranking functions:

Definition 2 (ranking function): Let $\mathbb{A}$ be a measurable space. A (measurable) function $\psi : \mathbb{A}^m \to [\,m\,]$ is called a *ranking function* if for all $(a_1, \ldots, a_m) \in \mathbb{A}^m$ we have:

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

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

that is the function is invariant w.r.t. 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

$$\psi\big(\,a_i, \{a_k\}_{k \neq i}\,\big) \;\neq\; \psi\big(\,a_j, \{a_k\}_{k \neq j}\,\big), \tag{13}$$

where the simplified notation is justified by P1.

Because of P2 when $a_1, \ldots, a_n \in \mathbb{A}$ are pairwise different ψ assigns a unique *rank* in $[m]$ to each a_i by $\psi(a_i, \{a_k\}_{k \neq i}))$. We would like to consider the rank of $\mathcal{D}_0$ w.r.t. $\{\mathcal{D}_i\}_{I=1}^{m-1}$, hence we apply ranking functions on $\mathcal{D}_0, \ldots, \mathcal{D}_{m-1}$. One can observe that these datasets are not necessarily pairwise different causing a technical challenge. To resolve ties in the ordering we extend each sample with the different values of a uniformly generated (independently from every other variable) random permutation, π, on set $[m]$, i.e., we let

$$\mathcal{D}_j^\pi \doteq (\mathcal{D}_j, \pi(j)) \quad \text{for} \quad j = 1, \ldots, m-1 \tag{14}$$

and $\mathcal{D}_0^\pi \doteq (\mathcal{D}_0, \pi(m))$. Assume that a ranking function $\psi : (\mathbb{X} \times \mathbb{Y})^n \times [m] \to [m]$ is given. Then, we define our tests as follows. Let p and q be user-chosen integers such that $1 \leq p < q \leq m$. We accept hypothesis H_0 if and only if

$$p \leq \psi\big(\mathcal{D}_0^\pi, \ldots, \mathcal{D}_{m-1}^\pi\big) \leq q, \tag{15}$$

i.e., we reject the nullhypothesis if the computed rank statistic is "extremal" (too low or too high). Our main tool to determine exact type I error probabilities is Theorem 2, originally proposed in [9, Theorem 1].

Theorem 2: Assume that $\mathcal{D}_0$ is an i.i.d. sample (A0). For all ranking function ψ, if H_0 holds true, then we have

$$\mathbb{P}\Big(\, p \leq \psi\big(\mathcal{D}_0^\pi, \ldots, \mathcal{D}_{m-1}^\pi\,\big) \leq q \Big) = \frac{q - p + 1}{m}. \tag{16}$$

The intuition behind this result is that if H_0 holds true then the original dataset behaves similarly to the alternative ones, consequently, its rank in an ordering admits a (discrete) uniform distribution. The main power of this thoerem comes from its distribution-free and non-asymptotically guaranteed nature. Furthermore, we can observe that parameters p, q and m are user-chosen, hence the probability of the acceptance region can be controlled exactly when H_0 holds true, that is the probability of the type I error is exactly quantified.

The main statistical assumption of Theorem 2 is very mild, that is we only require the data sample, $\mathcal{D}_0$, to be i.i.d. Even

though we presupposed that a ranking function is given, one can observe that our definition for ψ is quite general. Indeed, it also allows some degenerate choices that only depend on the ancillary random permutation, π, which is attached to the datasets. Our intention is to exclude such options, therefore we examine the type II error probabilities of our hypothesis tests. We present two new ranking functions that are endowed with strong asymptotic bounds for their type II errors.

B. Conditional Kernel Mean Embedding Based Tests

The proposed ranking functions are defined via conditional kernel mean embeddings. The main idea is to compute the empirical estimate of the conditional kernel mean map based on all available samples, both the original and the alternatively generated ones, and compare these to an estimate obtained from the regression function of the nullhypothesis. The main observation is that the estimates based on the alternative samples always converge to the theoretical conditional kernel mean map, which can be deduced from f, while the estimate based on the original sample, $\mathcal{D}_0$, converges to the theoretical one only if H_0 holds true. We assume that:

- A1 Kernel l is the naïve kernel.
- A2 Kernel k is real-valued, measurable, C_0-universal [17] and bounded by C_k as well as $\Gamma = k\,\mathbf{Id}_{\mathcal{G}}$.

One can easily guarantee A2 by choosing a "proper" kernel k, e.g., a suitable choice is the Gaussian kernel, if $\mathbb{X} = \mathbb{R}^d$.

We can observe that if a regression function, f, is given in binary classification, then the exact conditional kernel mean embedding $\mu_{\bar{Y}|X}$ and mean map μ_f can be obtained as

$$\begin{aligned} &\mu_{\bar{Y}|X} = (1-p(X))l(\circ,-1) + p(X)l(\circ,1) \;\text{ and} \\ &\mu_f(x) = (1-p(x))l(\circ,-1) + p(x)l(\circ,1) \quad P_X\text{-a.s.}, \end{aligned} \tag{17}$$

because by the reproducing property for all $g \in \mathcal{G}$ we have

$$\begin{aligned} &\langle (1-p(X))l(\circ,-1) + p(X)l(\circ,1), g\rangle_{\mathcal{G}} \\ &= (1-p(X))g(-1) + p(X)g(1) \overset{\text{a.s.}}{=} \mathbb{E}\big[g(\bar{Y}) \,|\, X\big]. \end{aligned} \tag{18}$$

For simplicity, we denote μ_{f_*} by μ_*. We propose two methods to empirically estimate μ_f. First, we use the regularized risk minimizer, $\widehat{\mu}$, defined in Theorem 1, and let

$$\widehat{\mu}_j^{(1)} \doteq \widehat{\mu}_{\mathcal{D}_j} \quad \text{for } j = 0,\dots,m-1. \tag{19}$$

Second, we rely on the intuitive form of (17) and estimate the conditional probability function p directly by any standard method (e.g., k-nearest neighbors) and let

$$\widehat{\mu}_j^{(2)}(x) \doteq (1-\widehat{p}_j(x))l(\circ,-1) + \widehat{p}_j(x)l(\circ,1), \tag{20}$$

for $j = 0,\dots,m-1$, where $\widehat{p}_j = \widehat{p}_{\mathcal{D}_j}$ denotes the estimate of p based on sample $\mathcal{D}_j$. The first approach follows our motivation by using a vector-valued RKHS and the user-chosen kernel Γ lets us adaptively control the possibly high-dimensional scalar product. The second technique highly relies on the used conditional probability function estimator, hence we can make use of a broad range of point estimators available for this problem. For brevity, we call the first approach vector-valued kernel test (VVKT) and the second approach point estimation based test (PET). The main advantage of VVKT comes from its nonparametric nature, while PET can be favorable when a priori information on the structure of f is available.

Let us define the ranking functions with the help of reference variables, which are estimates of the deviations between the empirical estimates and the theoretical conditional kernel mean map in some norm. An intuitive norm to apply is the expected loss in $\|\cdot\|_{\mathcal{G}}$, i.e., for $\mathbb{X} \to \mathcal{G}$ type functions μ_f, $\widehat{\mu} \in L_2(P_X;\mathcal{G})$ we consider the expected loss

$$\int_{\mathbb{X}} \|\mu_f(x) - \widehat{\mu}(x)\|_{\mathcal{G}}^2 \,\mathrm{d}P_X(x). \tag{21}$$

The usage of this "metric" is justified by [18, Lemma 2.1], where it is proved that for any estimator $\widehat{\mu}$ and conditional kernel mean map μ_f, we have

$$\int_{\mathbb{X}} \|\mu_f(x) - \widehat{\mu}(x)\|_{\mathcal{G}}^2 \,\mathrm{d}P_X(x) = \mathcal{E}_s(\widehat{\mu}) - \mathcal{E}_s(\mu_f), \tag{22}$$

where the right hand side is the excess risk of $\widehat{\mu}$. The distribution of X is unknown, thus the reference variables and the ranking functions are constructed as[2]

$$\begin{aligned} Z_j^{(r)} &\doteq \frac{1}{n}\sum_{i=1}^{n} \Big\|\mu_f(X_i) - \widehat{\mu}_j^{(r)}(X_i)\Big\|_{\mathcal{G}}^2, \\ \mathcal{R}_n^{(r)} &\doteq 1 + \sum_{j=1}^{m-1} \mathbb{I}\big(\, Z_j^{(r)} \prec_\pi Z_0^{(r)} \,\big) \end{aligned} \tag{23}$$

for $j = 0,\dots,m-1$ and $r \in \{1,2\}$, where r refers to the two conditional kernel mean map estimators, (19) and (20). The acceptance regions of the proposed hypothesis tests are defined by (15) with $\psi(\mathcal{D}_0^\pi,\dots,\mathcal{D}_{m-1}^\pi) = \mathcal{R}_n^{(r)}$. The idea is to reject f when $Z_0^{(r)}$ is too high in which case our original estimate is far from the theoretical map given f. Hence setting p to 1 is favorable. The main advantage of these hypothesis tests is that we can adjust the exact type I error probability to our needs for any finite n, irrespective of the sample distribution. Moreover, asymptotic guarantees can be ensured for the type II probabilities. We propose the following assumption to provide asymptotic guarantees:

- B1 For the conditional kernel mean map estimates we have

$$\frac{1}{n}\sum_{i=1}^{n} \Big\|\mu_*(X_i) - \widehat{\mu}^{(1)}(X_i)\Big\|_{\mathcal{G}}^2 \xrightarrow[n\to\infty]{a.s.} 0.$$

That is we assume that the used regularized risk minimizer is consistent in the sense above. Condition B1, although nontrivial, is however key in proving that a hypothesis test can preserve the favorable asymptotic behaviour of the point estimator while also non-asymptotically guaranteeing a user-chosen type I error probability.

Theorem 3: Assume that A0, A1, A2 and H_0 hold true, then for all sample size $n \in \mathbb{N}$ we have

$$\mathbb{P}\Big(\mathcal{R}_n^{(1)} \le q\Big) = \frac{q}{m}. \tag{24}$$

If A0, A1, A2, B1, $q < m$ and H_1 hold, then

$$\mathbb{P}\Big(\bigcap_{N=1}^{\infty}\bigcup_{n=1}^{N}\big\{\mathcal{R}_n^{(1)} \le q\big\}\Big) = 0. \tag{25}$$

[2] $Z_j^{(r)} \prec_\pi Z_0^{(r)} \iff Z_j^{(r)} < Z_0^{(r)}$ or $\big(Z_j^{(r)} = Z_0^{(r)}$ and $\pi(j) < \pi(m)\big)$

The tail event in (25) is often called the "lim sup" of events $\{\mathcal{R}_n^{(1)} \leq q\}$, where $n \in \mathbb{N}$. In other words, the theorem states that the probability of type I error is exactly $1 - q/m$. Moreover, under H_1, $\mathcal{R}_n^{(1)} \leq q$ happens infinitely many times with zero probability, equivalently a "false" regression function is (a.s.) accepted at most finitely many times. The pointwise convergence of the type II error probabilities to zero (as $n \to \infty$) is a straightforward consequence of (25).

A similar theorem holds for the second approach, where we assume the consistency of $\widehat{p}$ in the following sense:

C1 For conditional probability function estimator $\widehat{p}$ we have

$$\frac{1}{n}\sum_{i=1}^{n}(p(X_i) - \widehat{p}(X_i))^2 \xrightarrow[n\to\infty]{a.s.} 0. \tag{26}$$

Condition C1 holds for a broad range of conditional probability estimators (e.g., kNN, various kernel estimates), however most of these techniques make stronger assumptions on the data generating distribution than we do. As in Theorem 3 the presented stochastic guarantee for the type I error is non-asymptotic, while for the type II error it is asymptotic.

Theorem 4: Assume that A0, A1, A2 and H_0 hold true, then for all sample size $n \in \mathbb{N}$ we have

$$\mathbb{P}\Big(\mathcal{R}_n^{(2)} \leq q\Big) = \frac{q}{m}. \tag{27}$$

If A1, C1, $q < m$ and H_1 hold, then

$$\mathbb{P}\Big(\bigcap_{N=1}^{\infty}\bigcup_{n=1}^{N}\{\mathcal{R}_n^{(2)} \leq q\}\Big) = 0. \tag{28}$$

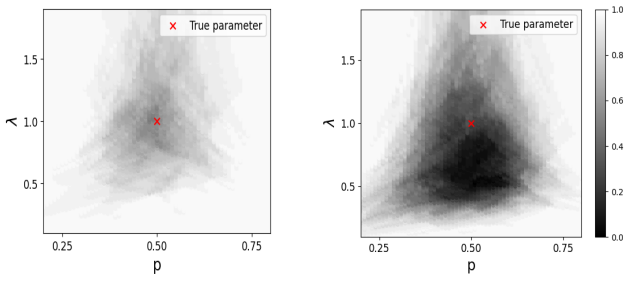
The proof of both theorems can be found in the appendix.

V. Numerical Simulations

We made numerical experiments on a synthetic dataset to illustrate the suggested hypothesis tests. We considered the one dimensional, bounded input space $[-1, 1]$ with binary outputs. The marginal distribution of the inputs were uniform. The true regression function was the following model:

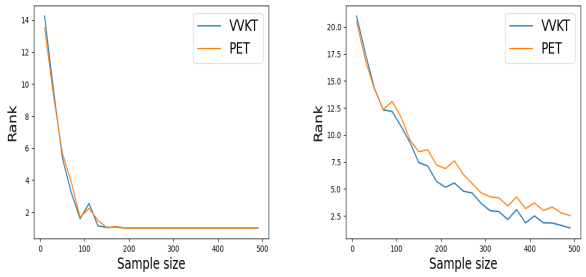
$$f_*(x) \doteq \frac{p_* \cdot e^{-(x-\mu_1)^2/\lambda_*} - (1-p_*)e^{-(x-\mu_2)^2/\lambda_*}}{p_* \cdot e^{-(x-\mu_1)^2/\lambda_*} + (1-p_*)e^{-(x-\mu_2)^2/\lambda_*}}, \tag{29}$$

where $p_* = 0.5$, $\lambda_* = 1$, $\mu_1 = 1$ and $\mu_2 = -1$. This form of the regression function is the reparametrization of a logistic regression model, which is an archetypical approach for binary classification. We get the same formula if we mix together two Gaussian classes. The translation parameters (μ_1 and μ_2) were considered to be known to illustrate the hypothesis tests with two dimensional pictures. The sample size was $n = 50$ and the resampling parameter was $m = 40$. We tested parameter pairs of (p, λ) on a fine grid with stepsize 0.01 on $[0.2, 0.8] \times [0.5, 1.5]$. The two hypothesis tests are illustrated with the generated rank statistics for all tested parameters on Figures 1(a) and 1(b). These normalized values are indicated with the colors of the points. Kernel k was a Gaussian with parameter $\sigma = 1/2$ for the VVKTs and we used kNN-estimates for PETs with $k = \lfloor\sqrt{n}\rfloor$ neighbors. We illustrated the consistency of our algorithm by plotting the ranks of the reference variables for parameters



(a) VVKT (b) PET

Fig. 1. The normalized ranks of VVKTs and PETs are represented with the darkness of points on a grid based on a sample of size $n = 50$ with resampling parameter $m = 40$. For VVKTs we used a Gaussian kernel with $\sigma = 1/2$. For PETs we used kNN with $k = \lfloor\sqrt{n}\rfloor$ neighbors.



(a) Ranks ($p = 0.3$, $\lambda = 1.3$) (b) Ranks ($p = 0.4$, $\lambda = 1.2$)

Fig. 2. The ranks of the reference variables are shown as functions of the sample size for two arbitrarily chosen "false" candidate functions.

$(0.3, 1.3)$ and $(0.4, 1.2)$ for various sample sizes in Figures 2(a) and 2(b). We took the average rank over several runs for each datasize. We can see that the reference variable corresponding to the original sample rapidly tends to be the greatest, though the rate of convergence depends on the particular hypothesis.

VI. Conclusions

In this paper we have introduced two new distribution-free hypothesis tests for the regression functions of binary classification based on conditional kernel mean embeddings.

Both proposed methods incorporate the idea that the output labels can be resampled based on the candidate regression function we are testing. The main advantages of the suggested hypothesis tests are as follows: (1) they have a user-chosen exact probability for the type I error, which is non-asymptotically guaranteed for any sample size; furthermore, (2) they are also consistent, i.e., the probability of their type II error converges to zero, as the sample size tends to infinity.

Our approach can be used to quantify the uncertainty of classification models, it can form a basis of confidence region constructions and thus can support robust decision making.

References

[1] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, vol. 31. Springer, 2013.

[2] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optim., and Beyond*. The MIT Press, 2001.

[3] G. Pillonetto, F. Dinuzzo, T. Chen, G. De Nicolao, and L. Ljung, "Kernel Methods in System Identification, Machine Learning and Function Estimation: A Survey," *Automatica*, vol. 50, no. 3, pp. 657–682, 2014.

[4] L. Song, J. Huang, A. Smola, and K. Fukumizu, "Hilbert Space Embeddings of Conditional Distributions with Applications to Dynamical Systems," in *26th Annual International Conference on Machine Learning (ICML), Montreal, Quebec, Canada*, p. 961–968, 2009.
[5] V. Vapnik, *Statistical Learning Theory*. Wiley-Interscience, 1998.
[6] M. Sadinle, J. Lei, and L. Wasserman, "Least Ambiguous Set-Valued Classifiers with Bounded Error Levels," *Journal of the American Statistical Association*, vol. 114, no. 525, pp. 223–234, 2019.
[7] R. F. Barber, "Is Distribution-Free Inference Possible for Binary Regression?," *Electronic Journal of Statistics*, pp. 3487–3524, 2020.
[8] C. Gupta, A. Podkopaev, and A. Ramdas, "Distribution-Free Binary Classification: Prediction Sets, Confidence Intervals and Calibration," in *34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada*, 2020.
[9] B. Cs. Csáji and A. Tamás, "Semi-Parametric Uncertainty Bounds for Binary Classification," in *58th IEEE Conference on Decision and Control (CDC), Nice, France*, pp. 4427–4432, 2019.
[10] A. Carè, B. Cs. Csáji, M. C. 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, 2017.
[11] C. A. Micchelli and M. Pontil, "On Learning Vector-Valued Functions," *Neural Computation*, vol. 17, no. 1, pp. 177–204, 2005.
[12] S. Grünewälder, G. Lever, L. Baldassarre, S. Patterson, A. Gretton, and M. Pontil, "Conditional Mean Embeddings as Regressors," in *Proceedings of the 29th International Coference on Machine Learning*, ICML'12, p. 1803–1810, Omnipress, 2012.
[13] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer, 2004.
[14] T. Hytönen, J. Van Neerven, M. Veraar, and L. Weis, *Analysis in Banach Spaces*, vol. 12. Springer, 2016.
[15] A. Smola, A. Gretton, L. Song, and B. Schölkopf, "A Hilbert Space Embedding for Distributions," in *International Conference on Algorithmic Learning Theory*, pp. 13–31, Springer, 2007.
[16] J. Park and K. Muandet, "A Measure-Theoretic Approach to Kernel Conditional Mean Embeddings," in *Advances in Neural Information Processing Systems 33*, Dec. 2020.
[17] C. Carmeli, E. De Vito, A. Toigo, and V. Umanitá, "Vector valued reproducing kernel hilbert spaces and universality," *Analysis and Applications*, vol. 8, no. 01, pp. 19–61, 2010.
[18] J. Park and K. Muandet, "Regularised Least-Squares Regression with Infinite-Dimensional Output Space," *arXiv:2010.10973*, 2020.

APPENDIX

A. Proof of Theorem 3

Proof: The first equality follows from Theorem 2. When the alternative hypothesis holds true, i.e., $f \neq f_*$, let $S_n^{(j)} = \sqrt{Z_j^{(1)}}$ for $j = 0, \dots, m-1$. It is sufficient to show that $S_n^{(0)}$ tends to be the greatest in the ordering as $n \to \infty$, because the square root function is order-preserving. For $j = 0$ by the reverse triangle inequality we have

$$\begin{aligned} S_n^{(0)} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n} \Big\| \mu_f(X_i) - \widehat{\mu}_0^{(1)}(X_i) \Big\|_{\mathcal{G}}^2} \\ &\geq \sqrt{\frac{1}{n}\sum_{i=1}^{n} \| \mu_f(X_i) - \mu_*(X_i) \|_{\mathcal{G}}^2} \\ &\quad - \sqrt{\frac{1}{n}\sum_{i=1}^{n} \Big\| \mu_*(X_i) - \widehat{\mu}_0^{(1)}(X_i) \Big\|_{\mathcal{G}}^2}. \end{aligned} \tag{30}$$

The first term converges to a positive number, as

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} \| \mu_f(X_i) - \mu_*(X_i) \|_{\mathcal{G}}^2 = \\ &= \frac{1}{n}\sum_{i=1}^{n} \| (p(X_i) - p_*(X_i)) l(\circ, 1) \\ &\qquad + (p_*(X_i) - p(X_i)) l(\circ, -1) \|_{\mathcal{G}}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \big[\big(p(X_i) - p_*(X_i)\big)^2 l(1,1) \\ &\qquad + \big(p_*(X_i) - p(X_i)\big)^2 l(-1,-1) \big] \\ &= \frac{2}{n}\sum_{i=1}^{n} \Big[\big(p(X_i) - p_*(X_i)\big)^2 \Big] \to 2\,\mathbb{E}\Big[\big(p(X) - p_*(X)\big)^2 \Big], \end{aligned} \tag{31}$$

where we used the SLLN and that $l(1,-1) = l(-1,1) = 0$. When $f \neq f_*$ we have that $\kappa \doteq \mathbb{E}\Big[\big(p(X) - p_*(X)\big)^2 \Big] > 0$. The second term almost surely converges to zero by B1, hence we can conclude that $S_n^{(0)} \xrightarrow{a.s.} \sqrt{2\kappa}$.

For $j \in [m-1]$ variable $Z_j^{(1)}$ has a similar form as the second term in (30), thus its (a.s.) convergence to 0 follows from B1, i.e., we get $Z_j^{(1)} \xrightarrow{a.s.} 0$. Hence, $Z_0^{(1)}$ (a.s.) tends to become the greatest, implying (25) for $q < m$. ∎

B. Proof of Theorem 4

Proof: The first part of the theorem follows from Theorem 2 with $p = 1$. For the second part let $f \neq f_*$. We transform the reference variables as

$$\begin{aligned} Z_j^{(2)} &= \frac{1}{n}\sum_{i=1}^{n} \| (p(X_i) l(\circ, 1) + (1 - p(X_i)) l(\circ, -1) \\ &\qquad - \big(\widehat{p}_j(X_i) l(\circ, 1) + (1 - \widehat{p}_j(X_i)) l(\circ, -1) \big) \|_{\mathcal{G}}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \| (p(X_i) - \widehat{p}_j(X_i)) l(\circ, 1) \\ &\qquad + (\widehat{p}_j(X_i) - p(X_i)) l(\circ, -1) \|_{\mathcal{G}}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \big((p(X_i) - \widehat{p}_j(X_i))^2 l(1,1) \\ &\qquad + (\widehat{p}_j(X_i) - p(X_i))^2 l(-1,-1) \big) \\ &= \frac{2}{n}\sum_{i=1}^{n} (p(X_i) - \widehat{p}_j(X_i))^2 \end{aligned} \tag{32}$$

for $j = 0, \dots, m-1$. From C1 it follows that $Z_j^{(2)}$ goes to zero a.s. for $j \in [m-1]$. For $j = 0$ we argue that $Z_0^{(2)} \to \kappa$ for some $\kappa > 0$. Notice that

$$\begin{aligned} &\frac{2}{n}\sum_{i=1}^{n} (p(X_i) - \widehat{p}_0(X_i))^2 \\ &= \frac{2}{n}\sum_{i=1}^{n} (p(X_i) - p_*(X_i))^2 + \frac{2}{n}\sum_{i=1}^{n} (p_*(X_i) - \widehat{p}_0(X_i))^2 \\ &+ \frac{4}{n}\sum_{i=1}^{n} \Big(\big(p(X_i) - p_*(X_i) \big) \big(p_*(X_i) - \widehat{p}_0(X_i) \big) \Big) \end{aligned}$$

holds. By the SLLN the first term converges to a positive number, $\mathbb{E}\big[(p_*(X) - p_0(X))^2\big] > 0$. By C1 the second term converges to 0 (a.s.). The third term also tends to 0 by the Cauchy-Schwartz inequality and C1, as for $x \in \mathbb{X}$ we have $| p(x) - p_*(x) | \leq 1$. We conclude that if $f \neq f_*$, $Z_0^{(2)}$ (a.s.) tends to be the greatest in the ordering implying (28). ∎