Conditions for Existence of Uniformly Consistent Classifiers

Agne Kazakeviciute, Vytautas Kazakevicius, and Malini Olivo

***Abstract*— We consider the statistical problem of binary classification, which means attaching a random observation X from a separable metric space E to one of the two classes, 0 or 1. We prove that the consistent estimation of conditional probability $p(X) = \mathsf{P}(Y = 1 \mid X)$, where Y is the true class of X, is equivalent to the consistency of a class of empirical classifiers. We then investigate for what classes P there exist an estimate $\hat{p}$ that is consistent uniformly in $p \in P$. We show that this holds if and only if P is a totally bounded subset of $L^1(E, \mu)$, where μ is the distribution of X. In the case, where E is countable, we give a complete characterization of classes Π, allowing consistent estimation of p, uniform in $(\mu, p) \in \Pi$.**

***Index Terms*— Consistency in uniform metric, classification, functional data analysis, pattern recognition, uniform consistency, universal consistency.**

I. Introduction

BINARY classification amounts to attaching an observation x from a separable metric space E to one of the two classes, $y = 0$ or $y = 1$. Formally, a *classifier* is a Borel function $h\colon E \to \{0, 1\}$. The pair (x, y), where $y \in \{0, 1\}$ is the true class of x, is considered as a realization of a random vector (X, Y). The quality of a classifier is measured by the false positive and false negative probabilities. If the distribution of (X, Y) is known and the costs for false positives and false negatives are given, then the form of the optimal, so-called Bayes, classifier is known. However, in practice the distribution of (X, Y) is unknown and is estimated from a training set $(X_1, Y_1), \ldots, (X_n, Y_n)$ of independent copies of (X, Y). The problem of classification then amounts to providing an estimate $\hat{h}_n$ of the Bayes classifier h^*, given that training set. We call $\hat{h}_n$ an *empirical classifier*. As we will discuss in Section II, the task of providing a good empirical estimator $\hat{h}_n$ is directly related to the task of providing a good estimate $\hat{p}_n$ of the conditional probability $p(x) = \mathsf{P}(Y = 1 \mid X = x)$. Therefore, we will now focus on the latter task.

The distribution of (X, Y) is uniquely defined by the distribution μ of X and the conditional probabilities $p(x)$. Let $M_{\max}$ denote the set of all distributions on E, $P_{\max}$ the set of all Borel functions $p\colon E \to [0; 1]$ and $\Pi_{\max} = M_{\max} \times P_{\max}$. The estimate $\hat{p}_n$ is called *consistent*, if $\hat{p}_n(X) \rightsquigarrow p(X)$, where $\rightsquigarrow$ denotes convergence in probability. The notion of consistency obviously depends on the distribution of (X, Y), that is, an estimate $\hat{p}_n$ may be consistent for one pair (μ, p) and may not be consistent for another. An estimate $\hat{p}_n$ is called *universally consistent*, if it is consistent for all $(\mu, p) \in \Pi_{\max}$.

In the case, where $E = \mathbb{R}^d$ the existence of universally consistent estimates was noticed by Stone [1] and since then universal consistency of various estimators has been proved (see [2]–[8]). As pointed out in [9], some proofs can be transferred to the more general spaces E such as separable metric spaces. The need in such generalization has recently emerged in relation with a growing attention to functional data analysis (see [10], [11] for the general theory, or [12] for a proof of consistency of a concrete estimator in Hilbert space). For the sake of completeness, we give a proof of universal consistency of histogram type estimates in the end of Section II. However, our main interest is on the notion stronger than that of consistency, namely, *uniform consistency*.

Definition 1: For $\Pi \subset \Pi_{\max}$, an estimate $\hat{p}_n$ is called Π-*uniformly consistent*, if the convergence $\hat{p}_n(X) \rightsquigarrow p(X)$ is uniform over $(\mu, p) \in \Pi$, that is, if for all $\varepsilon > 0$,

$$\sup_{(\mu,p)\in\Pi} \mathsf{P}_{\mu,p}(|\hat{p}_n(X) - p(X)| > \varepsilon) \to 0, \tag{1}$$

as $n \to \infty$ (we write $\mathsf{P}_{\mu,p}$ and $\mathsf{E}_{\mu,p}$ instead of P and E when we want to indicate explicitly the distribution of (X, Y)).

The notion of uniform consistency, as in Definition 1, should not be confused with the similar notion meaning that $\sup_x |\hat{p}_n(x) - p(x)| \rightsquigarrow 0$. We prefer to call the latter *consistency in uniform metric*, while the usual consistency can be referred as *consistency in L^1 metric* (because $\hat{p}_n(X) \rightsquigarrow p(X)$ if and only if $\int |\hat{p}_n - p| \mathrm{d}\mu \rightsquigarrow 0$). Consistency in uniform metric of regression function estimators has been first studied in [13] and [14] and various results on this topic are continuously emerging (see [15]–[19]).

It should be also noted that a lot of papers in machine learning theory adopt another approach to the problem of binary classification than that described above. Instead of estimating the function p they try to find a consistent empirical classifier $\hat{h}_n$ directly. The so-called *ERM principle* recommends using empirical estimators $\hat{h}_n$ that minimize *empirical*

Manuscript received July 14, 2015; revised April 19, 2016 and December 2, 2016; accepted March 28, 2017. Date of publication April 24, 2017; date of current version May 18, 2017. This work was supported in part by the Department of Statistical Science, University College London, U.K., and in part by the Singapore Bioimaging Consortium, Agency for Science, Technology and Research, Singapore. *(Corresponding author: Agne Kazakeviciute.)*

A. Kazakeviciute is with the Department of Statistical Science, University College London, WC1E 6BT London, U.K. (e-mail: a.kazakeviciute.12@ucl.ac.uk).

V. Kazakevicius is with the Department of Mathematical Statistics, Vilnius University, 01513 Vilnius, Lithuania.

M. Olivo is with the Agency for Science, Technology and Research, Singapore 138632, and also with the School of Physics, National University of Ireland, H91 CF50 Galway, U.K.

Communicated by I. Nikiforov, Associate Editor for Detection and Estimation.

Digital Object Identifier 10.1109/TIT.2017.2696961

risk, for example,

$$\hat{R}_n(h) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{h(X_i)\neq Y_i\}},$$

over some class of classifiers H. If the class is not too big then such an empirical estimator has also a small mean risk $\mathsf{E}R(\hat{h}_n)$, where $R(h) = \mathsf{P}(h(X) \neq Y)$. More precisely, if $\hat{R}_n(h) \rightsquigarrow R(h)$ for each $h \in H$ then $R(\hat{h}_n) \rightsquigarrow R^*_H = \inf_{h\in H} R(h)$. If H contains the Bayes classifier, $\hat{h}_n$ is consistent (its risk tends in probability to the risk of the Bayes classifier). If not, a sequence $\hat{h}_{kn}$, $k \geq 1$, of ERM classifiers, corresponding to some classes H_k, can be considered. If $\inf_k R^*_{H_k}$ is equal to the risk of the Bayes classifier, one may expect that some empirical classifier of the form $\hat{h}_{k_n n}$ would be consistent. The problem is in choosing an appropriate sequence k_n. It may be proved that such a sequence exists (and does not depend on unknown pair $(\mu, p) \in \Pi$) if each H_k has the following *uniform convergence property* (see [20, Definition 4.3]): for all ε

$$\sup_{(\mu,p)\in\Pi} \sup_{h\in H} \mathsf{P}_{\mu,p}(|\hat{R}_n(h) - R(h)| > \varepsilon) \to 0, \tag{2}$$

as $n \to \infty$.

Definition (2) resembles our definition of uniform consistency (1) (both notions claim uniform convergence in probability). So maybe the uniform convergence property is the analogue of uniform consistency in that data analytical approach to binary classification problem. However, the exact relation between the two notions are not yet clear to us.

Our notion of uniform consistency, as described in Definition 1, is standard in statistical estimation theory. For example, the local asymptotic minimax theorem ([21, Th. 8.11]) says that if a sample from a distribution, which smoothly depends on some parameter $\theta \in \Theta$ (Θ is an open subset of $\mathbb{R}^d$), and a smooth function $\theta \mapsto \psi_\theta$ on Θ are given then for any estimator $\hat{\psi}_n$ of ψ_θ based on that sample, any bowl-shaped loss function ℓ and any $\theta_0 \in \Theta$ the following inequality holds:

$$\sup_{c<\infty} \varliminf_{n\to\infty} \sup_{|\theta-\theta_0|<c/\sqrt{n}} \mathsf{E}_\theta \ell(\sqrt{n}(\hat{\psi}_n - \psi_\theta)) \geq \mathsf{E}\ell(Z_0), \tag{3}$$

where Z_0 is a random vector distributed according to the normal $N(0, \dot{\psi}_{\theta_0} I^{-1}_{\theta_0} \dot{\psi}^{\top}_{\theta_0})$ law, $\dot{\psi}_\theta$ denotes the derivative of ψ_θ and I_θ is the information matrix of the model. If the lower bound in (3) is attained for some estimate $\hat{\psi}_n$, it is called asymptotically efficient. Clearly, every asymptotically efficient estimator (for example, the maximum likelihood estimator), is (locally) uniformly consistent.

To the best of our knowledge, there are currently no papers on pattern recognition, where uniform consistency of estimates $\hat{p}_n$ would be analyzed. There is also no result analogous to that of the local asymptotic minimax theorem (3). This is due to the fact that usually only nonparametric classes P of possible functions p are considered and in that case there is no standard rate of convergence for the estimates $\hat{p}_n$, such as $n^{-1/2}$ in the parametric or semi-parametric estimation theory. One exception is the logistic regression model, where $E = \mathbb{R}^d$ and $P = \{p_\theta \mid \theta \in E\}$ with

$$p_\theta(x) = \frac{1}{1 + \mathrm{e}^{-\langle\theta,x\rangle}}. \tag{4}$$

In this case the maximum likelihood estimator $\hat{\theta}_n$ of the unknown parameter θ tends to the true value of that parameter at the rate $n^{-1/2}$, which yields the same convergence rate of $\hat{p}_n(X) = p_{\hat{\theta}_n}(X)$ to $p(X) = p_\theta(X)$.

It is known (see [9, Th. 7.2]) that for each estimate $\hat{p}_n$ we can find $(\mu, p) \in \Pi_{\max}$ with the arbitrary slow rate of convergence of $\hat{p}_n$ to p under the distribution (μ, p). We can bypass the latter fact by fixing μ and some $P \subset P_{\max}$ and by asking what is the asymptotic lower bound for, say,

$$\Delta_n(\mu, P) = \inf_{\hat{p}_n} \sup_{p\in P} \mathsf{E}_{\mu,p}|\hat{p}_n(X) - p(X)|.$$

However, this still seems to be a very difficult problem because even the rate of convergence of $\Delta_n(\mu, P)$ can depend on μ and P in an unpredictable manner. Therefore, we can first ask ourselves, for what pairs (μ, P) this quantity at least tends to 0.

It is easily seen that $\Delta_n(\mu, P) \to 0$ if and only if there exists a $(\mu \times P)$-uniformly consistent estimator (here $\mu \times P$ is a shorthand for $\{\mu\} \times P$). Therefore, the problem of characterizing pairs (μ, P) for which $(\mu \times P)$-uniformly consistent estimator exists can be thought as the first step towards the theory of asymptotically efficient binary classifiers. In Section III we give such a characterization. It appears that $(\mu \times P)$-uniformly consistent estimators exist if and only if the set P is totally bounded in $L^1(E, \mu)$.

This paper is organized as follows. In Section II we study the relation between $\hat{h}_n$ and $\hat{p}_n$ and prove that the consistency of the estimate $\hat{p}_n$ is equivalent to the consistency of a class of binary classifiers. We also give some additional arguments in favor of our definition of consistency of $\hat{h}_n$ by proving that it can be equivalently defined in a number of ways. Finally, we construct some histogram-type estimator in the general case of a separable metric space E and prove its universal consistency. In Section III we present our main result which is the characterization of pairs (μ, P) for which there exist $(\mu \times P)$-uniformly consistent estimators. We also prove that histogram-type estimator constructed in Section II is uniformly consistent in all such cases. Then we discuss the more difficult problem of Π-uniform consistency for arbitrary $\Pi \subset \Pi_{\max}$ and give its solution in the simplest case, where the set E is countable.

II. Relationship Between Binary Classification and Estimating the Conditional Probability

We always suppose that the values $\hat{p}_n(x)$ of any estimate $\hat{p}_n$ are in $[0; 1]$. In that case consistency of $\hat{p}_n$ can be equivalently defined in many other ways. For any loss function L on $[0; 1]^2$, the estimate $\hat{p}_n$ is called *L-consistent*, if $\mathsf{E}L(p(X), \hat{p}_n(X)) \to 0$, as $n \to \infty$.

Theorem 1 states that L-consistency is equivalent to consistency for a large class of loss functions. This result is used to prove Theorem 2.

Theorem 1: Let L be a continuous non-negative function on $[0; 1]^2$ *and, for all* $p, p' \in [0; 1]$, $L(p, p') = 0 \iff p = p'$. *Then* $\hat{p}_n$ *is L-consistent if and only if it is consistent.*

The proof of Theorem 1 is given in Appendix A.

The usual choices of the loss function are $L(p, p') = |p - p'|$ and $L(p, p') = (p - p')^2$. Hence consistency of $\hat{p}_n$ is equivalent to either of the two following conditions:

$$\mathsf{E}\big|\hat{p}_n(X) - p(X)\big| \to 0 \quad \text{or} \quad \mathsf{E}\big(\hat{p}_n(X) - p(X)\big)^2 \to 0.$$

Let u and $1-u$ (here $u \in (0; 1)$) denote the costs for false negative and false positive decisions, respectively. Then the *risk* of a classifier h is defined by

$$R_u(h) = u\mathsf{P}(h(X) = 1, Y = 0) + (1-u)\mathsf{P}(h(X) = 0, Y = 1),$$

and the problem of classification is to find a h with as small risk as possible. The value $u = 1/2$, corresponding to the case of equal costs for false positives and false negatives, leads to the risk

$$R_{1/2}(h) = \frac{1}{2}\mathsf{P}(h(X) \neq Y) = \frac{1}{2}R(h),$$

where R is the risk mentioned in the Introduction. This is the usual choice, however, for some reason that will be explained later in this Section, we need the risks R_u with an arbitrary u.

By definition of conditional probability, for any function f,

$$\begin{aligned}\mathsf{E}f(X, Y) &= \mathsf{E}\mathsf{E}^X f(X, Y)\\ &= \mathsf{E}\mathsf{E}^X(f(X, 0)\mathsf{1}_{\{Y=0\}} + f(X, 1)\mathsf{1}_{\{Y=1\}})\\ &= \mathsf{E}f(X, 0)(1 - p(X)) + f(X, 1)p(X),\end{aligned}$$

(here E^X denotes conditional expectation, given X). Taking

$$\begin{aligned}f(X, Y) &= \mathsf{1}_{\{h(X)=1, Y=0\}} + \mathsf{1}_{\{h(X)=0, Y=1\}}\\ &= h(X)\mathsf{1}_{\{Y=0\}} + (1 - h(X))\mathsf{1}_{\{Y=1\}}\end{aligned}$$

we get

$$\begin{aligned}R_u(h) &= u\mathsf{E}(1 - p(X))h(X) + (1 - u)\mathsf{E}p(X)(1 - h(X))\\ &= \mathsf{E}(u - p(X))h(X) + (1 - u)\mathsf{E}p(X).\end{aligned}$$

The risk takes the least possible value, if $h(x) = 1$, when $u < p(x)$, and $h(x) = 0$, otherwise. Hence, the optimal classifier is given by

$$h_u^*(x) = \begin{cases}1, & \text{if } p(x) > u,\\ 0, & \text{otherwise,}\end{cases} \tag{5}$$

Its risk is denoted by R_u^*, so

$$R_u^* = (1 - u)\mathsf{E}p(X) - \mathsf{E}\big(p(X) - u\big)\mathsf{1}_{\{p(X)>u\}}.$$

The form of the Bayes classifier suggests the following scheme for obtaining empirical classifiers. We should estimate the unknown function p by, say, $\hat{p}_n$ and use

$$\hat{h}_{un}(x) = \begin{cases}1, & \text{if } \hat{p}_n(x) > u,\\ 0, & \text{otherwise.}\end{cases} \tag{6}$$

Note that the estimate $\hat{p}_n$ itself does not depend on u. Nevertheless, it provides an entire class of empirical classifiers containing, for each value of u, a possibly good approximation to Bayes classifier h_u^*. It seems very likely that the classification problem described above is equivalent to the problem of estimating the unknown function p of conditional probabilities. To formulate this principle more precisely, we need the following definition.

Definition 2: An empirical classifier $\hat{h}_n$ is called *u-consistent*, if $\mathsf{E}R_u(\hat{h}_n) \to R_u^*$, as $n \to \infty$.

The mean risk of the classifier (6) is

$$\mathsf{E}R_u(\hat{h}_{un}) = (1 - u)\mathsf{E}p(X) - \mathsf{E}\big(p(X) - u\big)\mathsf{1}_{\{\hat{h}_n(X)>u\}}.$$

Therefore,

$$\begin{aligned}\mathsf{E}R_u(\hat{h}_{un}) - R_u^* &= \mathsf{E}\big(p(X) - u\big)\big(\mathsf{1}_{\{p(X)>u\}} - \mathsf{1}_{\{\hat{p}_n(X)>u\}}\big)\\ &= \mathsf{E}\big(p(X) - u\big)\mathsf{1}_{\{p(X)>u\geq\hat{p}_n(X)\}}\\ &\quad + \mathsf{E}\big(u - p(X)\big)\mathsf{1}_{\{\hat{p}_n(X)>u\geq p(X)\}},\end{aligned}$$

which yields

$$0 \leq \mathsf{E}R_u(\hat{h}_{un}) - R_u^* \leq \mathsf{E}\big|\hat{p}_n(X) - p(X)\big| \tag{7}$$

and also

$$\begin{aligned}&\int_0^1 \big(\mathsf{E}R_u(\hat{h}_{un}) - R_u^*\big)\mathrm{d}u\\ &\quad= \mathsf{E}\mathsf{1}_{\{\hat{p}_n(X)<p(X)\}}\int_{\hat{p}_n(X)}^{p(X)}\big(p(X) - u\big)\mathrm{d}u\\ &\qquad+ \mathsf{E}\mathsf{1}_{\{p(X)<\hat{p}_n(X)\}}\int_{p(X)}^{\hat{p}_n(X)}\big(u - p(X)\big)\mathrm{d}u\\ &\quad= \frac{1}{2}\mathsf{E}\big(p(X) - \hat{p}_n(X)\big)^2.\end{aligned} \tag{8}$$

It is well-known that consistency of $\hat{p}_n$ implies u-consistency of $\hat{h}_{un}$ (see [22]). This also easily follows from (7). Conversely, if $\hat{h}_{un}$ is u-consistent for all u, then by dominated convergence and (8), $\mathsf{E}\big(p(X) - \hat{p}_n(X)\big)^2 \to 0$, that is, $\hat{p}_n$ is consistent. We thus proved the following theorem.

Theorem 2: Let $\hat{h}_{un}$ be defined by (6)*. The following two statements are then equivalent:*

1) $\hat{p}_n$ is consistent.

2) $\hat{h}_{un}$ is u-consistent for each u.

We end this Section by constructing a universally consistent histogram-type estimate of p. It is based on the idea of approximating conditional probability $p(x) = \mathsf{P}(Y = 1 \mid X = x)$ by

$$\mathsf{P}(Y = 1 \mid X \in A_n(x)) = \frac{\mathsf{P}(Y = 1,\ X \in A_n(x))}{\mathsf{P}(X \in A_n(x))},$$

where $A_n(x)$ is some set containing x, which should be small enough for approximation to be good enough, but with probability $\mathsf{P}(X \in A_n(x)) > 0$ large enough to allow good estimation by the sample of size n. Although they are consistent, the drawbacks of histogram-type estimates are well known and we do not recommend it for practical use. Our only aim here is to provide a proof of consistency, which is valid in the general case, where E is an arbitrary separable metric space, and which can be simply transformed into a proof of uniform consistency.

Let $\mathcal{P}_n = \{A_{nj} \mid j \geq 1\}$ be a sequence of finite partitions of E such that $\mathcal{P}_n$ is finer than $\mathcal{P}_{n-1}$ for any $n \geq 2$ and $\bigcup_n \mathcal{P}_n$ generates the Borel σ-algebra on E. Define

$$N_{nj} = \sum_{i=1}^n \mathsf{1}_{\{X_i \in A_{nj}\}}, \quad S_{nj} = \sum_{i=1}^n \mathsf{1}_{\{X_i \in A_{nj}, Y_i = 1\}}$$

and let $N_n(x) = N_{nj}$, $S_n(x) = S_{nj}$ for $x \in A_{nj}$. Then our estimate is defined by

$$\hat{p}_n(x) = \begin{cases} \frac{S_n(x)}{N_n(x)} & \text{if } N_n(x) > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Let k_n denote the number of non-empty elements in $\mathcal{P}_n$ ($k_n = \infty$ if there are infinitely many j with $A_{nj} \neq \varnothing$).

Theorem 3: If $k_n = o(n)$, as $n \to \infty$, then estimate (9) *is universally consistent.*

The proof of Theorem 3 is given in Appendix B.

We now switch to the main topic of our paper, uniform consistency defined by (1). Inspecting the proofs of Theorems 1 and 3 we easily see that both statements remain valid if wording 'consistent' is replaced by 'Π-uniformly consistent'. The analogue of Theorem 2 could also be proved, were the notion of Π-uniform u-consistency defined in the obvious way. However, we will not need the latter result in the next Section.

III. Uniform Consistency

If $\Pi \subset \Pi_{\max}$, an estimate $\hat{p}_n$ is called Π-uniformly consistent, if (1) holds. Inspecting the proof of Theorem 1, we see that uniform consistency can be equivalently defined by

$$\sup_{(\mu,p)\in\Pi} \mathsf{E}_{\mu,p} L(p(X), \hat{p}_n(X)) \to 0,$$

where L is an appropriate loss function. For instance, $\hat{p}_n$ is Π-uniformly consistent if either of the following two conditions holds:

$$\sup_{(\mu,p)\in\Pi} \mathsf{E}_{\mu,p} \left|\hat{p}_n(X) - p(X)\right| \to 0$$

or

$$\sup_{(\mu,p)\in\Pi} \mathsf{E}_{\mu,p} (\hat{p}_n(X) - p(X))^2 \to 0.$$

We denote by $L^1(E,\mu)$ the set of all μ-integrable functions $f\colon E \to \mathbb{R}$, endowed with the norm $\|f\| = \int |f|\, \mathrm{d}\mu$. More precisely, the elements of $L^1(E,\mu)$ are not the functions f themselves, but equivalence classes of such functions, where two functions are meant equivalent if they agree almost everywhere with respect to μ. Note that conditional probabilities $p(x) = \mathsf{P}(Y = 1 \mid X = x)$ are bounded and defined up to that equivalence, therefore they belong to the space $L^1(E,\mu)$.

Recall also that a subset of a metric space is called *totally bounded* if, for each $\varepsilon > 0$, it can be covered by a finite number of balls of diameter less than ε.

The following Theorem is the main result of our paper.

Theorem 4: 1. If $k_n = o(n)$, then estimate (9) *is $(\mu \times P)$-uniformly consistent for each $\mu \in M_{\max}$ and $P \subset P_{\max}$ such that P is totally bounded as a subset of $L^1(E,\mu)$.*

2. If there exists a $(\mu \times P)$-uniformly consistent estimate of p, then P is totally bounded as a subset of $L^1(E,\mu)$.

The proof of Theorem 4 is given in Appendix C.

The conditions for a subset of $L^1(E,\mu)$ to be totally bounded are known, but rarely used, because they are very confusing (as compared with the similar conditions for, say, the space $C[a;b]$). In our case P consists of bounded functions, but this fact eliminates only one (unfortunately, less complicated) condition. So P is totally bounded as a subset of $L^1(E,\mu)$ if and only if for each $\varepsilon > 0$ there exists a measurable partition $(A_1,\dots,A_k)$ of E such that, for all p in P there exists a measurable $B \subset E$ with $\mu(B^c) < \varepsilon$ and

$$\forall j \quad \forall x, x' \in B \cap A_j \ |p(x) - p(x')| < \varepsilon. \tag{10}$$

In practice, however, we can use the well-known fact that P is totally bounded if and only if it is precompact, that is, if each sequence $(p_n) \subset P$ contains a converging subsequence. In [23] we used this criterion to establish (in the case, where $E = \mathbb{R}^d$) total boundedness of the set $P = \{p_\theta \mid \theta \in \mathbb{R}^d\}$, where p_θ are given by (4).

It seems to be a real challenge to prove a generalization of Theorem 4 for arbitrary subsets $\Pi \subset \Pi_{\max}$ instead of $\mu \times P$, or at least for the sets Π of the form $M \times P$, where $M \subset M_{\max}$ and $P \subset P_{\max}$. In the latter case, we can expect that $(M \times P)$-uniformly consistent estimators will exist if and only if P is totally bounded in each $L^1(E,\mu)$ and if this holds uniformly, in some way, over $\mu \in M$. Criterion (10) gives several suggestions about how this condition could look. For example, the set B, or the partition $(A_1,\dots,A_k)$ could not depend on μ. However, we could not prove any statement of such kind.

In our opinion, the first step in solving this problem should be the analysis of a number of specific models. Moreover, some 'natural' set P_0 should exist in these models so that we could focus only on the sets M for which $(M \times P_0)$-uniformly consistent estimators exist. The simplest such model is that with the countable set E. In this case $P_{\max}$ is totally bounded as a subset of each $L^1(E,\mu)$ and therefore there exists an estimate of p, which is $(\mu \times P_{\max})$-uniformly consistent for all μ. It is interesting to characterize the sets $M \subset M_{\max}$, for which there exist $(M \times P_{\max})$-uniformly consistent estimates. It appears that these are exactly the sets satisfying the following condition:

$$\sup_{\mu\in M} \mathsf{P}_\mu(X_1 \neq X, \dots, X_n \neq X) \to 0, \tag{11}$$

as $n \to \infty$. Here $X, X_1, ..., X_n$ are independent random variables identically distributed according to the law μ.

Theorem 5: 1. If M satisfies condition (11)*, then histogram-type estimate* (9)*, corresponding to the sequence of partitions $\mathcal{P}_n = \{\{j\} \mid j \geq 1\}$ is $(M \times P_{\max})$-uniformly consistent.*

2. If there exists an $(M \times P_{\max})$-uniformly consistent estimate of p, then M satisfies condition (11)*.*

The proof of Theorem 5 is given in Appendix D.

IV. Conclusions and Discussion

The main result of Section II is Theorem 2, which relates consistency of empirical classifiers to consistency in L^1 metric of estimators of conditional probabilities. In our opinion, it also shows that L^1 metric is more natural in this context than, say, uniform metric.

The main result of Section III, as well as of the whole paper, is Theorem 4, which gives necessary and sufficient conditions on (μ, P) for the existence of $(\mu \times P)$-uniformly consistent estimates $\hat{p}_n$. As we have explained in Introduction,

we consider this result as an analogue of the local asymptotic minimax theorem in statistical estimation theory. If an estimator $\hat{p}_n$ is $(\mu \times P)$-consistent for all pairs (μ, P) such that P is totally bounded in $L^1(E, \mu)$, we think of it as asymptotically efficient in some sense (say, $o(1)$-efficient). For instance, histogram type estimators described in Section II are $o(1)$-efficient.

This suggests the following programme of revising the existing classification procedures: each estimator $\hat{p}_n$ should be tested on efficiency and modified, if needed (by using a histogram-type estimator as a model, for example). The modified estimator would probably perform better even for finite sample size n.

We have already started this programme in [23], where we consider the logistic classifier in $E = \mathbb{R}^d$. In that case the set $P = \{p_\theta \mid \theta \in \mathbb{R}^d\}$, where p_θ is given by (4), is totally bounded in $L^1(E, \mu)$, for each distribution μ. However, we could prove $(\mu \times P)$-uniform consistency of the logistic classifier only under some assumption on μ, which can be roughly described in the following way. Let $\bar{P}$ denote the set of functions that are point-wise limits of sequences $(p_{\theta_n}) \subset P$. For $p, p' \in \bar{P}$ denote also $d(p, p') = \int |p - p'| \mathrm{d}\mu$ and $N_{pp'} = \{p \neq p'\}$. Clearly, $d(p, p') = 0$ implies that $N_{pp'}$ is a null set, i.e. $\mu(N_{pp'}) = 0$. The above mentioned assumption says that, for any $p \in \bar{P}$, the union $\bigcup_{p':d(p,p')=0} N_{pp'}$ is also a null set. The assumption seems very technical and at the moment we do not know if it is crucial (it is needed only in the case $d \geq 3$). However, if the logistic classifier turned out to indeed not be uniformly consistent, it would be interesting to find out how its modification would look.

APPENDIX A
PROOF OF THEOREM 1

Proof: The proof of Theorem 1 is based on the following two formulas:

$$\forall \varepsilon\ \exists \delta\ \forall p, p' \in [0; 1]\ (L(p, p') < \delta \Rightarrow |p - p'| < \varepsilon),$$
$$\forall \varepsilon\ \exists \delta\ \forall p, p' \in [0; 1]\ (|p - p'| < \delta \Rightarrow L(p, p') < \varepsilon).$$

Let us prove the first one, the second one is proved analogously. Suppose the contrary and find an ε and $p_n, p'_n \in [0; 1]$ such that $L(p_n, p'_n) < 1/n$ and $|p_n - p'_n| \geq \varepsilon$ for all n. Without loss of generality we can assume that $p_n \to p$ and $p'_n \to p'$ for some $p, p' \in [0; 1]$. Then, by continuity, $L(p_n, p'_n) \to L(p, p')$ and therefore $L(p, p') = 0$, and $p = p'$. Hence $\varepsilon \leq |p_n - p'_n| \to 0$, a contradiction.

If $\hat{p}_n$ is L-consistent, then for any ε there exists δ such that

$$\mathsf{P}(|\hat{p}_n(X) - p(X)| \geq \varepsilon) \leq \mathsf{P}(L(p(X), \hat{p}_n(X)) \geq \delta).$$

By Chebyshev inequality,

$$\mathsf{P}(L(p(X), \hat{p}_n(X)) \geq \delta) \leq \delta^{-1} \mathsf{E} L(p(X), \hat{p}_n(X)) \to 0,$$

therefore $\hat{p}_n$ is consistent. Conversely, let $\hat{p}_n$ be consistent. By continuity, L is bounded. Denote $c = \sup_{0 \leq p, p' \leq 1} L(p, p')$. Then, for any ε there exists δ such that

$$\mathsf{E} L(p(X), \hat{p}_n(X)) \leq \varepsilon + c\mathsf{P}(\big|\hat{p}_n(X) - p(X)\big| \geq \delta) < 2\varepsilon,$$

for n sufficiently large, that is, $\hat{p}_n$ is L-consistent. ∎

APPENDIX B
PROOF OF THEOREM 3

Proof: Let μ be the distribution of X and T_n denote the operator of conditional expectation with respect to μ, given the σ-algebra $\mathcal{B}_n$, generated by the partition $\mathcal{P}_n$ (that is, $(T_n q)(X) = \mathsf{E}^{\mathcal{B}_n} q(X)$). It is well known that T_n is a continuous linear operator from $L^1(E, \mu)$ to itself with $\|T_n\| \leq 1$. Moreover, by our assumptions on $\mathcal{P}_n$, $\|T_n q - q\| \to 0$, as $n \to \infty$, for every $q \in L^1(E, \mu)$. Set $\tilde{p}_n = T_n p$. Then $\mathsf{E}\big|\tilde{p}_n(X) - p(X)\big| \to 0$ and it remains to prove that $\mathsf{E}\big|\hat{p}_n(X) - \tilde{p}_n(X)\big| \to 0$, as $n \to \infty$. Since our partitions do not depend on $(X_1, \ldots, X_n)$, we can proceed similarly to the proof of the analogous part of in [9, Th. 6.1] (in the case, where partitions depend on data, their proof contains a gap).

For short, denote $a_{nj} = \mu(A_{nj})$ and $b_{nj} = \int_{A_{nj}} p \, \mathrm{d}\mu$. Then

$$\mathsf{E}\big|\hat{p}_n(X) - \tilde{p}_n(X)\big| = \sum_{a_{nj} > 0} a_{nj} \mathsf{E} 1_{\{N_{nj} > 0\}} \Big|\frac{S_{nj}}{N_{nj}} - \frac{b_{nj}}{a_{nj}}\Big| + \sum_{a_{nj} > 0} b_{nj} \mathsf{P}(N_{nj} = 0).$$

Moreover,

$$\mathsf{E} 1_{\{N_{nj} > 0\}} \Big|\frac{S_{nj}}{N_{nj}} - \frac{b_{nj}}{a_{nj}}\Big| = \sum_{I \neq \varnothing} \mathsf{E} 1_{W_{njI}} \Big|\frac{1}{|I|} \sum_{i \in I} 1_{\{Y_i = 1\}} - \frac{b_{nj}}{a_{nj}}\Big|,$$

where I denotes various subsets of $\{1, \ldots, n\}$, $|I|$ is the number of elements in I and

$$W_{njI} = \{\forall i \in I\ X_i \in A_{nj},\ \forall i \notin I\ X_i \notin A_{nj}\}.$$

Conditionally on W_{njI}, the sum $Z_{nj} = \sum_{i \in I} 1_{\{Y_i = 1\}}$ is distributed according to the binomial law with parameters $|I|$ and $\frac{b_{nj}}{a_{nj}}$, therefore it is independent of W_{njI} and

$$\begin{aligned}\mathsf{E} 1_{W_{njI}} \Big|\frac{1}{|I|} \sum_{i \in I} 1_{\{Y_i = 1\}} - \frac{b_{nj}}{a_{nj}}\Big| &= \mathsf{P}(W_{njI}) \frac{1}{|I|} \mathsf{E}|Z_{njI} - \mathsf{E} Z_{njI}| \\ &\leq \mathsf{P}(W_{njI}) \frac{1}{|I|} \sqrt{\mathsf{Var} Z_{njI}} \\ &\leq \mathsf{P}(W_{njI}) |I|^{-1/2}.\end{aligned}$$

Hence

$$\begin{aligned}&\mathsf{E}\big|\hat{p}_n(X) - \tilde{p}_n(X)\big| \\ &\leq \sum_{a_{nj} > 0} a_{nj} \sum_{I \neq \varnothing} \mathsf{P}(W_{njI}) |I|^{-1/2} + \sum_{a_{nj} > 0} b_{nj} \mathsf{P}(N_{nj} = 0) \\ &\leq \frac{1}{\sqrt{k}} + \sum_{a_{nj} > 0} a_{nj} \sum_{|I| < k} \mathsf{P}(W_{njI}) \\ &= \frac{1}{\sqrt{k}} + \mathsf{P}(N_n(X) < k)\end{aligned}$$

and it remains to show that $\mathsf{P}(N_n(X) = s) \to 0$ for any fixed s.

Suppose that $A_{nj} = \varnothing$ for $j > k_n$. Then for all $n \geq s$,

$$\begin{aligned}
\mathsf{P}(N_n(X) = s) &= \sum_{j=1}^{k_n} a_{nj} \mathsf{P}(N_{nj} = s) \\
&= \sum_{j=1}^{k_n} \binom{n}{s} a_{nj}^{s+1} (1 - a_{nj})^{n-s} \\
&\leq \sum_{j=1}^{k_n} \frac{n^s}{s!} a_{nj}^{s+1} \mathrm{e}^{-(n-s)a_{nj}} \\
&\leq \frac{\mathrm{e}^s}{s! n} \sum_{j=1}^{k_n} (n a_{nj})^{s+1} \mathrm{e}^{-n a_{nj}} \\
&\leq \frac{\mathrm{e}^s c_s}{s!} \frac{k_n}{n} = o(1),
\end{aligned}$$

where $c_s = \max_{u \geq 0} u^{s+1} \mathrm{e}^{-u} = (s+1)^{s+1} \mathrm{e}^{-(s+1)}$. ∎

Appendix C
Proof of Theorem 4

Proof: 1. Inspecting the proof of Theorem 3 shows that it suffices to prove that

$$\sup_{p \in P} \| T_n p - p \| \to 0, \tag{12}$$

as $n \to \infty$, where $T_n p$ denotes the conditional expectation of p with respect to σ-algebra, generated by the partition $\mathcal{P}_n$. Suppose the contrary and find an ε and a sequence $(p_n) \subset P$ such that $\| T_n p_n - p_n \| \geq \varepsilon$ for all n. Since P is totally bounded (= precompact), we can suppose that $p_n \to p \in P_{\max}$ (in $L^1(E, \mu)$). Then

$$\begin{aligned}
\| T_n p_n - p \| &\leq \| T_n p_n - T_n p \| + \| T_n p - p \| \\
&\leq \| T_n \| \, \| p_n - p \| + \| T_n p - p \| \\
&\leq \| p_n - p \| + \| T_n p - p \| \to 0.
\end{aligned}$$

2. We use one theorem of Yatracos [24], which extends the results of [25]. He considers the problem of estimation of the unknown distribution π using an i.i.d. sample $(Z_1, \ldots, Z_n)$ from π. The set $\mathcal{M}$ of possible distributions π is supposed to be separable with respect to the total variance metric ρ and the estimate $\hat{\pi}_n$ is called uniformly consistent, if $\sup_{\pi \in \mathcal{M}} \mathsf{E}_\pi \rho(\hat{\pi}_n, \pi) \to 0$, as $n \to \infty$. Reference [24, Th. 2] says that, if $\mathcal{M}$ is uniformly dominated by some probability π_0 (that is, $\sup_{\pi \in \mathcal{M}} \pi(A)$ is arbitrary small, whenever $\pi_0(A)$ is small enough) and if there exists a uniformly consistent estimate $\hat{\pi}_n$ of π, then $\mathcal{M}$ is totally bounded.

In our case $Z_i = (X_i, Y_i)$ and $\mathcal{M} = \{\pi_p \mid p \in P\}$, where π_p is the distribution on $E \times \{0, 1\}$ defined by

$$\pi_p(A \times \{0\}) = \int_A (1 - p) \, \mathrm{d}\mu \quad \text{and} \quad \pi_p(A \times \{1\}) = \int_A p \, \mathrm{d}\mu,$$

$A \subset E$ measurable. Since $0 \leq p \leq 1$, the set $\mathcal{M}$ is uniformly dominated by $\mu \times (\nu/2)$, where ν is the counting measure on $\{0, 1\}$. It is easily seen that $\rho(\pi_p, \pi_{p'}) = \int |p - p'| \, \mathrm{d}\mu$. Therefore, if $\hat{p}_n$ is a $(\mu \times P)$-uniformly consistent estimate of p, then $\pi_{\hat{p}_n}$ is uniformly consistent estimate of π_p. By [24, Th. 2], $\mathcal{M}$ is totally bounded, which implies that P is totally bounded as a subset of $L^1(E, \mu)$. ∎

Appendix D
Proof of Theorem 5

For simplicity, we suppose that $E = \{1, 2, 3, \ldots\}$ and write $\mu(j)$ instead of $\mu(\{j\})$ for $\mu \in M_{\max}$ and $j \in E$. First we show that condition (11) can be equivalently formulated in some other way.

Lemma 1: For each $M \subset M_{\max}$ the following three statements are equivalent:

1) $\sup_{\mu \in M} \mu\{j \mid \mu(j) < \delta\} \to 0$, *as* $\delta \to 0$,

2) for all $s \geq 0$, $\sup_{\mu \in M} \sum_j \binom{n}{s} \mu(j)^{s+1} (1 - \mu(j))^{n-s} \to 0$, *as* $n \to \infty$,

3) $\sup_{\mu \in M} \sum_j \mu(j)(1 - \mu(j))^n \to 0$, *as* $n \to \infty$.

Proof: $(1 \Rightarrow 2)$ For each $\mu \in M$ and $\delta > 0$,

$$\begin{aligned}
&\sum_j \binom{n}{s} \mu(j)^{s+1} (1 - \mu(j))^{n-s} \\
&\quad \leq \sum_{\mu(j) < \delta} \mu(j) + \sum_{\mu(j) \geq \delta} \binom{n}{s} (1 - \delta)^{n-s} \\
&\quad \leq \mu\{j \mid \mu(j) < \delta\} + n^s (1 - \delta)^{n-s}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\overline{\lim_{n \to \infty}} \sup_{\mu \in M} \sum_j \binom{n}{s} \mu(j)^{s+1} (1 - \mu(j))^{n-s} \\
&\quad \leq \sup_{\mu \in M} \mu\{j \mid \mu(j) < \delta\}
\end{aligned}$$

and, by assumption 1, the term on the right-hand side can be made arbitrary small.

$(2 \Rightarrow 3)$ Take $s = 0$.

$(3 \Rightarrow 1)$ For each $\mu \in M$ and $n \geq 1$

$$\begin{aligned}
\sum_j \mu(j)(1 - \mu(j))^n &\geq \frac{1}{2} \sum_{\mu(j) < 1 - 2^{-1/n}} \mu(j) \\
&= \frac{1}{2} \mu\{j \mid \mu(j) < 1 - 2^{-1/n}\}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sup_{\mu \in M} \mu\{j \mid \mu(j) < 1 - 2^{-1/n}\} \\
&\quad \leq 2 \sup_{\mu \in M} \sum_j \mu(j)(1 - \mu(j))^n.
\end{aligned}$$

Fix ε. By assumption 2, the term on the right-hand side $< \varepsilon$ for some n. Then, for all $\delta < 1 - 2^{-1/n}$,

$$\sup_{\mu \in M} \mu\{j \mid \mu(j) < \delta\} < \varepsilon.$$

Hence, $\sup_{\mu \in M} \mu\{j \mid \mu(j) < \delta\} \to 0$, as $\delta \to 0$. □

Note that condition 3 of Lemma 1 is exactly condition (11). Now we are ready to prove Theorem 5. The proof is based on the same idea as the proof of the so-called No-Free-Lunch Theorem (see [20, Th. 5.1]).

Proof: 1. Inspecting the proof of Theorem 3 shows that it suffices to prove that, for any fixed $s \geq 0$,

$$\sup_{\mu \in M} \mathsf{P}_\mu(N_n(X) = s) \to 0$$

as $n \to \infty$. This is exactly condition 2 of Lemma 1.

2. First note that, for any bounded measurable function f,

$$\begin{aligned}&\mathsf{E}_{\mu,p} f(X_1, Y_1, \ldots, X_n, Y_n)\mathbf{1}_{\{X_i \neq j, \ldots, X_n \neq j\}} \\ &= \sum_{\substack{x_1,\ldots,x_n \neq j \\ y_1,\ldots,y_n \in \{0,1\}}} f(x_1, y_1, \ldots, x_n, y_n) \\ &\quad \times \prod_{i=1}^{n} \mu(x_i) p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}\end{aligned}$$

and the term on the right-hand side does not depend on $p(j)$. This means that

$$\begin{aligned}&\mathsf{E}_{\mu,p} f(X_1, Y_1, \ldots, X_n, Y_n)\mathbf{1}_{\{\forall i \leq n \;\; X_i \neq j\}} \\ &\quad = \mathsf{E}_{\mu,p'} f(X_1, Y_1, \ldots, X_n, Y_n)\mathbf{1}_{\{\forall i \leq n \;\; X_i \neq j\}},\end{aligned}$$

provided $p(k) = p'(k)$ for all $k \neq j$.

Now suppose that $\hat{p}_n$ is an $(M \times P_{\max})$-uniformly consistent estimate of p. We need to prove that M satisfies condition (11). Suppose the contrary. Then there exists an ε and a sequence $(\mu_n) \subset M$ such that

$$\sum_j \mu_n(j)(1 - \mu_n(j))^n \geq 3\varepsilon,$$

for all n. For each n find a k_n such that

$$\sum_{j \leq k_n} \mu_n(j)(1 - \mu_n(j))^n \geq 2\varepsilon.$$

By uniform consistency, there exists an n_0 such that, for all $n \geq n_0$,

$$\sup_{p \in P_{\max}} \mathsf{E}_{\mu_n,p} \left| \hat{p}_n(X) - p(X) \right| < \varepsilon.$$

But

$$\begin{aligned}&\sup_{p \in P_{\max}} \mathsf{E}_{\mu_n,p} \left| \hat{p}_n(X) - p(X) \right| \\ &\quad \geq \sup_{p \in P_{\max}} \sum_j \mu(j) \mathsf{E}_{\mu_n,p} \left| \hat{p}_n(j) - p(j) \right| \mathbf{1}_{\{X_1 \neq j, \ldots, X_n \neq j\}} \\ &\quad \geq \sup_{p \in P_{\max}} \sum_j \mu(j) \Big| \mathsf{E}_{\mu_n,p} \hat{p}_n(j) \mathbf{1}_{\{X_1 \neq j, \ldots, X_n \neq j\}} \\ &\qquad - p(j)(1 - \mu(j))^n \Big|.\end{aligned}$$

Therefore, for $n \geq n_0$,

$$\begin{aligned}&\sup_{p \in P_{\max}} \sum_j \mu(j) \Big| \mathsf{E}_{\mu_n,p} \hat{p}_n(j) \mathbf{1}_{\{X_1 \neq j, \ldots, X_n \neq j\}} \\ &\qquad - p(j)(1 - \mu(j))^n \Big| < \varepsilon.\end{aligned}$$

Let P_n^* denote the subset of $P_{\max}$, consisting of 2^{k_n} functions p with the following two properties:

(a) $p(j) \in \{0, 1\}$ for all $j \leq k_n$,

(b) $p(j) = 0$ for all $j > k_n$.

Denote also $\varphi_{nj}(p) = \mathsf{E}_{\mu_n,p} \hat{p}_n(j) \mathbf{1}_{\{X_1 \neq j, \ldots, X_n \neq j\}}$. If $p \in P_n^*$ then $\left| \varphi_{nj}(p) - p(j)(1 - \mu(j))^n \right|$ equals either $\varphi_{nj}(p)$ (if $p(j) = 0$) or $(1 - \mu(j))^n - \varphi_{nj}(p)$ (if $p(j) = 1$). Therefore, for all $n \geq n_0$ and $p \in P_n^*$,

$$\begin{aligned}&\sum_{p(j)=0} \mu(j)\varphi_{nj}(p) + \sum_{p(j)=1} \mu(j)(1 - \mu(j))^n \\ &\quad - \sum_{p(j)=1} \mu(j)\varphi_{nj}(p) < \varepsilon.\end{aligned}$$

But $p(j) = 0$ for $j > k_n$, therefore, for $n \geq n_0$, the following 2^{k_n} inequalities corresponding to different $p \in P_n^*$ hold:

$$\begin{aligned}&\sum_{p(j)=0, j \leq k_n} \mu(j)\varphi_{nj}(p) + \sum_{p(j)=1, j \leq k_n} \mu(j)(1 - \mu(j))^n \\ &\quad - \sum_{p(j)=1, j \leq k_n} \mu(j)\varphi_{nj}(p) < \varepsilon.\end{aligned}$$

Let us sum up all these inequalities. Of course,

$$\begin{aligned}&\sum_{p \in P_n^*} \sum_{p(j)=1, j \leq k_n} \mu(j)(1 - \mu(j))^n \\ &\quad = \sum_{j=1}^{k_n} \mu(j)(1 - \mu(j))^n \sum_{p \in P_n^*, p(j)=1} 1 \\ &\quad = 2^{k_n - 1} \sum_{j=1}^{k_n} \mu(j)(1 - \mu(j))^n.\end{aligned}$$

Moreover,

$$\begin{aligned}&\sum_{p \in P_n^*} \sum_{p(j)=0, j \leq k_k} \mu(j)\varphi_{nj}(p) - \sum_{p \in P_n^*} \sum_{p(j)=1, j \leq k_k} \mu(j)\varphi_{nj}(p) \\ &= \sum_{j=1}^{k_n} \mu(j) \Big(\sum_{p \in P_n^*, p(j)=0} \varphi_{nj}(p) - \sum_{p \in P_n^*, p(j)=1} \varphi_{nj}(p) \Big) = 0\end{aligned}$$

because $\varphi_{nj}(p)$ does not depend on $p(j)$. Hence, for all $n \geq n_0$,

$$2^{k_n - 1} \sum_{j=1}^{k_n} \mu(j)(1 - \mu(j))^n < 2^{k_n}\varepsilon,$$

that is,

$$\sum_{j=1}^{k_n} \mu(j)(1 - \mu(j))^n < 2\varepsilon.$$

We got a contradiction. ∎

ACKNOWLEDGMENT

The authors would like to thank the Associate Editor, prof. Igor Nikiforov, and anonymous reviewers for constructive comments that helped to improve this work.

REFERENCES

[1] C. J. Stone, "Consistent nonparametric regression," *Ann. Statist.*, vol. 5, no. 4, pp. 595–645, 1977.

[2] L. P. Devroye and T. J. Wagner, "Distribution-free consistency results in nonparametric discrimination and regression function estimation," *Ann. Statist.*, vol. 8, no. 2, pp. 231–239, 1980.

[3] C. Spiegelman and J. Sacks, "Consistent window estimation in nonparametric regression," *Ann. Statist.*, vol. 8, pp. 240–246, Sep. 1980.

[4] L. Györfi, "Universal consistencies of a regression estimate for unbounded regression functions," in *Nonparametric Functional Estimation and Related Topics*, G. Roussas, Ed. Dordrecht, The Netherlands: Kluwer, 1991, pp. 329–338.

[5] M. Kohler, "On the universal consistency of a least squares spline regression estimator," *Math. Methods Statist.*, vol. 6, no. 3, pp. 349–364, 1997.

[6] L. Györfi, M. Kohler, and H. Walk, "Weak and strong universal consistency of semi-recursive partitioning and kernel regression estimates," *Statist. Decisions*, vol. 16, pp. 1–18, Sep. 1998.

[7] M. Kohler, "Universally consistent regression function estimation using hierarchial B-splines," *J. Multivariate Anal.*, vol. 67, pp. 138–164, Sep. 1999.
[8] M. Kohler, "Universal consistency of local polynomial kernel regression estimates," *Ann. Inst. Statist. Math.*, vol. 54, no. 4, pp. 879–899, 2002.
[9] L. Devroye, L. Györfy, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York, NY, USA: Springer, 1996.
[10] J. Ramsay and B. W. Silverman, *Functional Data Analysis*. New York, NY, USA: Springer-Verlag, 2005.
[11] F. Ferraty and P. Vieux, *Non-Parametric Functional Data Analysis*. New York, NY, USA: Springer-Verlag, 2006.
[12] G. Biau, F. Bunea, and M. H. Wegkamp, "Functional classification in Hilbert spaces," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 2163–2172, Jun. 2005.
[13] H. J. Bierens, "Uniform consistency of kernel estimators of a regression function under generalized conditions," *J. Amer. Stat. Assoc.*, vol. 78, no. 383, pp. 699–707, 1983.
[14] W. Hardle and L. Luckhaus, "Uniform consistency of a class of regression function estimators," *Ann. Statist.*, vol. 12, no. 2, pp. 612–623, 1984.
[15] D. Wied and R. Weisbach, "Consistency of kernel density estimator: A survey," *Statist. Lett.*, vol. 53, no. 1, pp. 1–21, 2012.
[16] N. L. Kudraszow and P. Vieu, "Uniform consistency of kkNN regressors for functional variables," *Statist. Probab. Lett.*, vol. 83, pp. 1863–1870, 2013.
[17] J. Gao, D. Li, and D. Tjostheim, "Uniform consistency for nonparametric estimators in null recurrent time series," *Econ. Theory*, vol. 31, no. 5, pp. 911–952, 2015.
[18] D. Li, P. C. B. Phillips, and J. Gao. (May 2015). *Uniform Consistency of Nonstationary Kernel-Weighted Sample Covariances for Nonparametric Regression*. [Online]. Available: http://dx.doi.org/10.1017/S0266466615000109
[19] M. Chaouch, N. Laib, and D. Louani, "Rate of uniform consistency for a class of mode regression on functional stationary ergodic data," *Statist. Methods Appl.*, vol. 26, no. 1, pp. 1–29, 2016.
[20] S. Shalev-Schwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2014.
[21] A. W. van der Vaart, *Asymptotic Statistics*. Cambridge, U.K.: Cambridge Univ. Press, 1998.
[22] J. Van Ryzin, "Bayes risk consistency of classification procedures using density estimation," *Sankhya, Indian J. Statist. A*, vol. 28, nos. 2–3, pp. 261–270, 1966.
[23] A. Kazakeviciute and M. Olivo, "A study of logistic classifier: Uniform consistency in finite-dimensional linear spaces," *J. Math. Statist. Operations Res. (JMSOR)*, vol. 3, no. 2, pp. 12–18, 2016.
[24] Y. G. Yatracos, "On the existence of uniformly consistent estimates," *Proc. Amer. Math. Soc.*, vol. 94, no. 3, pp. 479–486, 1985.
[25] L. LeCam and L. Schwartz, "A necessary and sufficient condition for the existence of consistent estimates," *Ann. Math. Stat.*, vol. 31, no. 1, pp. 140–150, 1960.

Agne Kazakeviciute received a Bsc degree in Statistics from Vilnius University, Lithuania, in 2012 and a MSc degree in Medical Statistics from University College London, United Kingdom, in 2013. In 2013 she was awarded the joint UCL-A*STAR scholarship to both pursue PhD studies in Statistical Science at UCL and take a two-year research assistant position at A*STAR in Singapore. She returned to UCL in 2016 and is currently a last year PhD student. She is particularly interested in statistical and machine learning pattern recognition as well as functional data theory and methodology. On top of this, her research interests include high-dimensional image processing as well as video compression and enhancement.

Vytautas Kazakevicius received a MSc degree in Mathematics in 1980 and a PhD in Mathematics in 1985 both from Vilnius University, Lithuania. Since 1985 he has been working in the Faculty of Mathematics and Informatics at Vilnius University where he currently works as a Professor in Probability and Statistics in the Department of Mathematical Statistics. His main research interests are in mathematical statistics and dynamical systems.

Malini Olivo is currently Head of Bio-Optical Imaging Group, Singapore Bioimaging Consortium, A*STAR, Singapore. She holds an adjunct Professorship at the National University of Ireland, Royal College of Surgeons, Ireland and is also a visiting Professor at Harvard Medical school. She obtained a Ph.D. in Bio-Medical Physics in 1990 from University Malaya/University College London, and did her post-doctoral training between 1991-1995 in University College London, UK and both McMaster University and University of Toronto, Canada. She was a Principal Investigator at the Singapore National Cancer Centre and Singhealth from 1996 to 2009, heading biophotonics research for clinical translation in cancer. In 2009, she took a Stokes Professorship Science Foundation Ireland Chair in the National University of Ireland and returned to Singapore in 2012 to head bio-optical imaging in SBIC, A*STAR. She is recognized as a pioneer in the area of clinical applications of photomedicine in optical diagnostics and therapeutics in Singapore. Her current research interests are in nano-biophotonics and its applications in translational medicine. Malini Olivo is well recognized internationally in her field and serves in numerous scientific advisory boards in the area of Photonics in Medicine. She currently serves in the EU research commission in Photonics 21 shaping research in photonics for Life sciences till 2015.