Optimal classification and nonparametric regression for functional data

ALEXANDER MEISTER

Institut für Mathematik, Universität Rostock, D-18051 Rostock, Germany.
E-mail: alexander.meister@uni-rostock.de

We establish minimax convergence rates for classification of functional data and for nonparametric regression with functional design variables. The optimal rates are of logarithmic type under smoothness constraints on the functional density and the regression mapping, respectively. These asymptotic properties are attainable by conventional kernel procedures. The bandwidth selector does not require knowledge of the smoothness level of the target mapping. In this work, the functional data are considered as realisations of random variables which take their values in a general Polish metric space. We impose certain metric entropy constraints on this space; but no algebraic properties are required.

Keywords: asymptotic optimality; kernel methods; minimax convergence rates; nonparametric estimation; topological data

1. Introduction

In many statistical applications, the empirical data cannot be described by random vectors in a Euclidean space $\mathbb{R}^d$. Still one can often reasonably define a distance between the possible realisations of the observations. Then parts of the data are supposed to take their values in a non-empty Polish metric space $(\mathcal{X}, \rho)$ where the corresponding probability measure has the corresponding Borel σ-field $\mathfrak{B}(\mathcal{X})$ as its domain. Note that a separable and complete metric space is called a Polish metric space.

Within that general framework, the analysis of functional data has attained increasing attention (see, e.g., the book of Ramsay and Silverman [22] for an introduction to the topic). Therein $\mathcal{X}$ denotes some appropriate function space, for example, the set of all continuous and bounded functions on [0, 1] or the set of all measurable and squared-integrable functions on that domain. The current work is mainly motivated by this research field; whereas, in general, the elements of $\mathcal{X}$ are not imposed to be functions or equivalence classes of functions, which opens up new perspectives for extensions to even more complex types of data. In particular, we use only topological properties of the set $\mathcal{X}$; but no algebraic structure on $\mathcal{X}$ is required (e.g., linear space, group, ring, etc.). Therefore, tools from principal component analysis (e.g., Benko, Härdle and Kneip [2]) or manifold representation (e.g., Chen and Müller [9]) cannot be applied in this setting. Instead, we use arguments based on covering and packing numbers from the approximation theory. Such techniques are frequently used in empirical process theory in order to study the parameter set of a statistical experiment, which consists of functions in nonparametric statistics (e.g., van der Vaart and Wellner [25], van de Geer [24]). Contrarily, they have been applied to learn about the sample set in quite few papers.

We focus on two widely studied problems in functional data analysis: nonparametric regression (Section 3) and classification (Section 4). A review on existing literature is provided in the corresponding sections. While a huge amount of literature is available on these topics, only little has been known about the aspect of asymptotic optimality of statistical procedures when the sample size n tends to infinity. The current note intends to advance the understanding of those problems by providing the minimax convergence rates for the statistical risks. The proofs are deferred to Section 5. Section 2 provides some essential topological tools which are used in both Sections 3 and 4.

2. Entropy condition

In the following, we recall two concepts from approximation theory (e.g., van der Vaart and Wellner [25], page 83, Definition 2.1.5 and page 98, Definition 2.2.3): by $\mathcal{N}_{\mathcal{X}}(\delta, \mathcal{Y}, \rho)$ we denote the covering number of some set $\mathcal{Y} \subseteq \mathcal{X}$, that is, the minimal number of open ρ-balls in $\mathcal{X}$ with the radius δ so that $\mathcal{Y}$ is a subset of the union of these balls. If we stipulate in addition that the centers of those balls lie in $\mathcal{Y}$, we call this quantity the intrinsic covering number $\mathcal{N}_{\mathcal{Y}}(\delta, \mathcal{Y}, \rho)$. The packing number $\mathcal{D}(\delta, \mathcal{Y}, \rho)$ of the set $\mathcal{Y}$ describes the maximal cardinality of a subset of $\mathcal{Y}$ such that $\rho(x, y) > \delta$ for all elements $x \neq y$ of this subset. Also, we learn from Kolmogorov and Tihomirov [19] and van der Vaart and Wellner [25], page 98, that

$$\mathcal{N}_{\mathcal{Y}}(\delta, \mathcal{Y}, \rho) \leq \mathcal{D}(\delta, \mathcal{Y}, \rho) \leq \mathcal{N}_{\mathcal{Y}}(\delta/2, \mathcal{Y}, \rho) \qquad \forall \delta > 0. \tag{2.1}$$

Also, we easily derive that

$$\mathcal{N}_{\mathcal{X}}(\delta, \mathcal{Y}, \rho) \leq \mathcal{N}_{\mathcal{Y}}(\delta, \mathcal{Y}, \rho) \leq \mathcal{N}_{\mathcal{X}}(\delta/2, \mathcal{Y}, \rho) \qquad \forall \delta > 0. \tag{2.2}$$

Now we classify a type of sets $\mathcal{Y}$ by their metric entropy, which we define by

$$\Phi(s, \mathcal{Y}, \rho) := \log \mathcal{N}_{\mathcal{X}}(s, \mathcal{Y}, \rho) \qquad \forall s > 0.$$

Concretely, we assume that

$$c_{x,0} s^{-\gamma} \leq \Phi(s, \mathcal{Y}, \rho) \leq c_{x,1} s^{-\gamma} \qquad \forall s \in (0, s_0), \tag{2.3}$$

for some fixed constants $s_0 > 0$, $0 < c_{x,0} < c_{x,1}$ and $\gamma > 0$. We write $B_{\mathcal{Y}}(x, r) := \{y \in \mathcal{Y} : \rho(x, y) < r\}$ for $x \in \mathcal{X}$ and $r > 0$. We easily see that $B_{\mathcal{Y}}(x, r) \in \mathfrak{B}(\mathcal{X})$ for all $x \in \mathcal{X}$, $r > 0$ and $\mathcal{Y} \in \mathfrak{B}(\mathcal{X})$. Condition (2.3) can be justified in many applications. Let us consider two examples of classes $\mathcal{Y}$ which satisfy this condition.

Example 2.1 (Classes of smooth functions). We assume that our functional data $X_1, \ldots, X_n$ are located in a class of smooth functions almost surely. We write $\lceil \alpha \rceil$ for the smallest integer which is larger or equal to $\alpha > 0$. Precisely, we impose the Hölder constraints that $\mathcal{Y}$ consists of functions f mapping from $[0, 1]^d$ to $\mathbb{R}$ such that all partial derivatives of f up to the order $\lceil \alpha \rceil - 1$ are bounded by a constant M; and that the $(\lceil \alpha \rceil - 1)$th partial derivatives satisfy the Hölder

condition with the exponent $\alpha - \lceil \alpha \rceil + 1$ and again the constant M. Also, we put $\mathcal{X} = C_0([0, 1]^d)$ and ρ equal to the supremum metric.

We learn from Theorem 2.7.1, page 155 in van der Vaart and Wellner [25] that the upper bound in condition (2.3) is satisfied with $\gamma = d/\alpha$. Also, the corresponding lower bound can be verified (see Kolmogorov and Tihomirov [19]).

Moreover, for any $\alpha > 0$, the Hölder class $\mathcal{Y}$ is relatively compact with respect to the supremum metric thanks to the Arzelà–Ascoli theorem, from what follows compactness of the closure $\overline{\mathcal{Y}}$. We deduce that

$$\lim_{\delta' \downarrow \delta} \mathcal{N}_{\mathcal{X}}\big(\delta', \overline{\mathcal{Y}}, \rho\big) \leq \mathcal{N}_{\mathcal{X}}(\delta, \mathcal{Y}, \rho) \leq \mathcal{N}_{\mathcal{X}}(\delta, \overline{\mathcal{Y}}, \rho) \qquad \forall \delta > 0,$$

since, for any cover of $\mathcal{Y}$ by the union of finitely many open balls, the union of the corresponding closed balls covers $\overline{\mathcal{Y}}$ (and so does the union of the corresponding open balls with arbitrarily enlarged radius). Therefore, condition (2.3) is extended from $\mathcal{Y}$ to $\overline{\mathcal{Y}}$; and the role of $\mathcal{Y}$ can be taken over by its closure.

In general, the technique of the last paragraph in Example 2.1, that is, switching to the closure of $\mathcal{Y}$, can be used to impose without loss of generality that $\mathcal{Y}$ is closed – and hence, $\mathcal{Y} \in \mathfrak{B}(\mathcal{X})$ – without any loss of generality when condition (2.3) is assumed.

Example 2.2 (Classes of monotonic functions). Now we consider the example of componentwise monotonic mappings from the cube $[0, 1]^d$ to $[0, 1]$. The collection of these functions is denoted by $\mathcal{Y}$. As the corresponding Polish metric space, we choose $\mathcal{X} = L_p([0, 1]^d)$, $p \geq 1$, that is, the Banach space of all Borel measurable functions f from $[0, 1]^d$ to $\mathbb{R}$ which satisfies $\int |f(x)|^p \, dx < \infty$. Clearly, ρ is the metric generated by the $L_p([0, 1]^d)$-norm.

Then Theorem 1.1 in Gao and Wellner [17] yields that $\mathcal{Y}$ satisfies condition (2.3) with $\gamma = \max\{d, (d-1)p\}$ for $d \geq 2$ and $(d-1)p \neq d$. In the univariate setting $d = 1$, the upper bound part of condition (2.3) with $\gamma = 1$ follows from Theorem 2.7.5 in van der Vaart and Wellner [25], page 159. Therein we use that the covering number is bounded from above by the bracketing number with doubled radius for the $L_p([0, 1]^d)$-metric ρ (see page 84, van der Vaart and Wellner [25]). On the other hand, the according lower bound can be established by Proposition 2.1 in Gao and Wellner [17].

The following lemma provides a useful result for the upper bound proofs in the following two sections.

Lemma 2.1. *Let $(\mathcal{X}, \rho)$ be a Polish metric space. Take some $\mathcal{Y} \in \mathfrak{B}(\mathcal{X})$ which satisfies (2.3), and let P be any probability measure on $\mathfrak{B}(\mathcal{X})$ with $P(\mathcal{Y}) = 1$. We set*

$$\psi(x, h) := P\big(B_{\mathcal{Y}}(x, h)\big), \qquad h > 0.$$

Then we have

$$P\big(\{x \in \mathcal{Y}: \psi(x, h) \leq \delta\}\big) \leq \delta \exp\big(c_{x,1} 4^\gamma h^{-\gamma}\big),$$

for all $\delta > 0$.

3. Nonparametric regression

We observe the data set $\mathcal{Z}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ where the X_j are i.i.d. random variables taking their values in the Polish metric space $(\mathcal{X}, \rho)$ equipped with the corresponding Borel σ-algebra $\mathfrak{B}(\mathcal{X})$. The Y_j are defined by

$$Y_j = g(X_j) + \varepsilon_j, \tag{3.1}$$

where g denotes some Borel measurable mapping from $\mathcal{X}$ to $\mathbb{R}$; and the ε_j are real-valued random variables which satisfy

$$E(\varepsilon_1|X_1) = 0, \qquad \mathrm{var}(\varepsilon_1|X_1) \leq c_v, \qquad P_X\text{-a.s.}, \tag{3.2}$$

for some uniform constant c_v where P_X denotes the probability measure on $\mathfrak{B}(\mathcal{X})$ which is generated by X_1. The random variables $(X_1, \varepsilon_1), \ldots, (X_n, \varepsilon_n)$ are assumed to be i.i.d. Moreover, we assume that $X_1 \in \mathcal{Y}$ holds almost surely for some subset $\mathcal{Y} \in \mathfrak{B}(\mathcal{X})$. Our goal is to estimate the regression function g based on the data set $\mathcal{Z}_n$.

As a usual condition in nonparametric regression, we impose some smoothness constraints on the regression function g. Precisely, we introduce the class $\mathcal{G} = \mathcal{G}_{\beta,C}$ of all Borel measurable mappings g from $\mathcal{X}$ to $\mathbb{R}$ such that $\sup_{y \in \mathcal{Y}} |g(y)| \leq C$ and

$$\left| g(y) - g(z) \right| \leq C\rho(y, z)^\beta \qquad \forall y, z \in \mathcal{Y}$$

with $C > 0$, $\beta \in (0, 1]$. Critically, we remark that our framework is restricted to smoothness degrees β which are smaller or equal to one. An extension to higher smoothness levels seems difficult as $\mathcal{X}$ is not equipped with any algebraic structure so that no common definitions of Taylor series can be applied. Approaches to local linear methods, which should capture all smoothness levels smaller than two, are provided in Berlinet, Elamine and Mas [3] and Mas [20]; while, in these papers, $\mathcal{X}$ is assumed to be a Hilbert space – transferred to our notation.

Whereas linear models for g (along with generalizations) are popular in functional regression problems (e.g., Hall and Horowitz [18], Meister [21]), fully nonparametric approaches to the regression function have also received considerable attention. We refer to the book of Ferraty and Vieu [14] for a comprehensive review on kernel methods for functional covariates. In Ferraty *et al.* [15], a generic upper bound is derived for the uniform rate of convergence. Recently, Forzani, Fraiman and Llop [16] consider consistency of nonparametric functional regression estimation in the setting of a metric space without any imposed algebraic structure. In a similar setting, Biau, Cérou and Guyader [5] establish upper bounds on an integrated risk for the convergence rates of the functional k-nearest neighbor estimator when $\beta = 1$ (in our notation). The convergence rates used in that paper are of logarithmic type. However, minimax optimality is apparently not studied in this work.

To our best knowledge, the only approach to rate-optimal nonparametric functional regression estimation is given by Mas [20], who uses principal component analysis on $\mathcal{X}$ and specific conditions on these components. The attained rates are faster than any logarithmic rates but slower than any polynomial rate in the non-Gaussian case. In our setting where the design distribution obeys the condition (2.3), the minimax convergence rates are different. We consider estimators

$\hat{g}$ of g which are Borel measurable mappings from $\mathcal{X}^{n+1}$ to $\mathbb{R}$ and which are squared integrable with respect to the design measure P_X after inserting the data, regardless of their realization. Also, we impose that $g \in L_2(P_X)$, that is, the Hilbert space of all squared integrable and measurable functions with respect to P_X. Then we are guaranteed that $\|\hat{g}(\cdot, \mathcal{Z}_n) - g\|_{P_X}^2$ is a real-valued random variable where $\|\cdot\|_{P_X}$ denotes the $L_2(P_X)$-norm.

We take the Nadaraya–Watson estimator for functional data,

$$\hat{g}(x) := \begin{cases} \hat{A}(x)/\hat{B}(x), & \text{if } \hat{B}(x) > \delta_n, \\ 0, & \text{otherwise,} \end{cases} \tag{3.3}$$

where

$$\hat{A}(x) := \frac{1}{n} \sum_{j=1}^{n} Y_j K_h\big(\rho(x, X_j)\big),$$

$$\hat{B}(x) := \frac{1}{n} \sum_{j=1}^{n} K_h\big(\rho(x, X_j)\big).$$

However, we have modified the concept by adding the truncation to the denominator $\hat{B}(x)$ where the ridge parameter $\delta_n > 0$ remains to be selected. Moreover, $h > 0$ denotes a bandwidth parameter and $K : \mathbb{R} \to \mathbb{R}$ a kernel function. We employ the notation $K_h := K(\cdot/h)$ (without dividing by h). For simplicity, we choose that $K = 1_{[0,1)}$. We provide the following asymptotic result.

Theorem 3.1. *Let* $\mathcal{Y} \in \mathfrak{B}(\mathcal{X})$ *such that* (2.3) *holds true. We consider model* (3.1) *under the condition* (3.2). *Then, for any sequence* $\{P_{X,n}\}_n$ *of design measures on* $\mathfrak{B}(\mathcal{X})$ *with* $P_{X,n}(\mathcal{Y}) = 1$ *for all* n, *the estimator* $\hat{g}$ *in* (3.3) *satisfies*

$$\sup_{g \in \mathcal{G}} \int E\big|\hat{g}(x) - g(x)\big|^2 \, dP_X(x) = \mathcal{O}\big(\{\log n\}^{-2\beta/\gamma}\big),$$

under the kernel choice $K = 1_{[0,1)}$ *and the parameter selection* $\delta_n = n^{-\eta}$, $\eta \in (0, 1/2)$ *and* $h = \{d \log n\}^{-1/\gamma}$ *with* $d \in (0, \eta c_{x,1}^{-1} 4^{-\gamma})$.

Remark 3.1. Under the additional assumption $P_X \in \mathcal{R}_X$, which says that

$$P_X\big(B_{\mathcal{Y}}(y, \delta)\big) \ge c_{x,3} \delta \exp\big(-c_{x,4} \delta^{-\gamma}\big) \qquad \forall \delta \in (0, 1), \, y \in \mathcal{Y},$$

which can be shown to be non-empty for some positive constants $c_{x,3}$ and $c_{x,4}$ and any compact $\mathcal{Y}$, we can also derive the following upper bound on the pointwise risk:

$$\sup_{P_X \in \mathcal{R}_X} \sup_{g \in \mathcal{G}} \sup_{x \in \mathcal{Y}} E\big|\hat{g}(x) - g(x)\big|^2 = \mathcal{O}\big(\{\log n\}^{-2\beta/\gamma}\big),$$

under the same conditions on K, δ_n and h as in Theorem 3.1 except that $d \in (0, \eta/c_{x,4})$.

We consider model (3.1) with the additional condition that the ε_j are i.i.d. random variables with a continuously differentiable density function f_ε with finite Fisher information, that is,

$$\int |f'_\varepsilon(x)|^2 / f_\varepsilon(x)\,\mathrm{d}x < \infty. \tag{3.4}$$

Moreover, all the $X_1, \varepsilon_1, \ldots, X_n, \varepsilon_n$ are independent. Also we impose compactness of the set $\mathcal{Y}$ from Theorem 3.2. The following theorem provides an asymptotic lower bound for the estimation of g with respect to the pointwise estimation error as well as the integrated risk.

Theorem 3.2. *Let $\mathcal{Y} \in \mathfrak{B}(\mathcal{X})$ be compact and assume that (2.3) holds true. We consider model (3.1) under independent additive regression errors ε_j, $j = 1, \ldots, n$ with a density f_ε which satisfies (3.4).*

(a) *Then there exists a sequence of design measures $P_{X,n}$ on $\mathfrak{B}(\mathcal{X})$ with $P_{X,n}(\mathcal{Y}) = 1$ for all n, such that no sequence of estimators $\{\hat{g}_n\}_n$ based on the data $\mathcal{Z}_n$ satisfies*

$$\sup_{g \in \mathcal{G}} \int E|\hat{g}(x) - g(x)|^2 \,\mathrm{d}P_{X,n}(x) = \mathrm{o}\big(\{\log n\}^{-2\beta/\gamma}\big).$$

(b) *For any sequence of design measures $P_{X,n}$ on $\mathfrak{B}(\mathcal{X})$ with $P_{X,n}(\mathcal{Y}) = 1$ for all n and for any sequence of estimators $\{\hat{g}_n\}_n$ based on the data $\mathcal{Z}_n$, we have*

$$\liminf_{n \to \infty} \sup_{g \in \mathcal{G}} \sup_{x \in \mathcal{Y}} P\big[|\hat{g}(x) - g(x)|^2 > c \cdot \{\log n\}^{-2\beta/\gamma}\big] > 0,$$

for some constant c depending on C and β.

Theorem 3.2 establishes minimax optimality of the convergence rate attained in Theorem 3.1 in two views. Part (a) shows that there exists a sequence of design measures such that the integrated risk does not converge with faster rates. Obviously, we cannot obtain such a result for any design measure: if $P_{X,n}$ was a one-point measure then just the average of the $Y_1, \ldots, Y_n$ would be a consistent estimator with the usual parametric rate. In part (b), we prove that no matter what the design measure looks like, one is not able to obtain faster pointwise convergence rates simultaneously for all $x \in \mathcal{Y}$, even with respect to the weak rates.

An important and widely studied issue in nonparametric regression is bandwidth selection. The minimax convergence rates are of slow logarithmic type. However, the bandwidth selector in Theorem 3.1 leads to the optimal rates while it can be used without knowing the smoothness degree β. This selector is fully deterministic, which means that no data-driven procedure (e.g., cross validation, Lepski's method, etc.) is required in order to achieve the optimal convergence rates. It is remarkable that the same effects occur in nonparametric deconvolution from super-smooth error distributions (see, e.g., Fan [13]) and other severely ill-posed inverse problems. We face a bias-dominating problem, that is, the variance term is asymptotically negligible under the optimal bandwidth selection. In other bias-dominating problems, sharp asymptotics have been studied (Butucea and Tsybakov [7]). It is an interesting question for future research if those results apply to the current problem as well.

4. Classification

The problem of classifying functional data has also stimulated great research activity (e.g., Ferraty and Vieu [14], Carroll, Delaigle and Hall [8], Delaigle and Hall [10,11], Biau, Bunea and Wegkamp [4]). It has its applications in the fields of biometrics, genetics, recognition of sounds, technometrics, etc. Classification problems are closely linked to the field of statistical learning theory (e.g., Vapnik [26]). We choose the model of supervised classification. Concretely, we observe some random variable Z taking its values in some Polish metric space $\mathcal{X}$ – and in $\mathcal{Y} \in \mathfrak{B}(\mathcal{X})$ almost surely. We assume that we have two groups 0 and 1 and our goal is to decide whether Z should be categorized as a member of group 0 or 1. The groups 0 and 1 are characterized by the probability measures P_X and P_Y on $\mathfrak{B}(\mathcal{X})$, respectively. One does not know these measures; however, i.i.d. a training sample (Z_j, W_j), $j = 1, \ldots, n$ is available where the W_j are binary random variables and $W_j = b$, $b = 0, 1$, indicates that Z_j has the probability measure P_X and P_Y, respectively. Moreover, Z is independent of all training data.

In order to specify all admitted probability measures P_X and P_Y, we impose that

$$(P_X, P_Y) \in \mathcal{P}_\kappa := \big\{(P, Q)\colon P \text{ and } Q \text{ are probability measures on } \mathfrak{B}(\mathcal{X}) \text{ so that}$$
$$P(\mathcal{Y}) = Q(\mathcal{Y}) = 1, \mathrm{TV}(P, Q) \geq \kappa\big\}, \tag{4.1}$$

for some $\kappa > 0$ where $\mathrm{TV}(P, Q)$ denotes the total variation distance between some measures P and Q,

$$\mathrm{TV}(P, Q) := \sup_{A \in \mathfrak{B}(\mathcal{X})} \big|P(A) - Q(A)\big|.$$

With respect to the set $\mathcal{Y}$, we assume condition (2.3).

Unlike in classification problems for data in $\mathbb{R}^d$, $d \in \mathbb{N}$, we face the problem that no spatially homogeneous measure (e.g., Lebesgue–Borel measure, Haar measure) exists on $\mathfrak{B}(\mathcal{X})$ so that no density of P_X and P_Y can be defined with respect to such a measure. Nevertheless, P_X and P_Y are dominated by their sum measure $Q := P_X + P_Y$. We write p_X and p_Y for the Radon–Nikodym derivatives $p_X := \mathrm{d}P_X/\mathrm{d}Q$ and $p_Y := \mathrm{d}P_Y/\mathrm{d}Q = 1 - p_X$. We impose some smoothness constraints on both p_X and p_Y via

$$(P_X, P_Y) \in \mathcal{P}_{C,\beta,\kappa} := \big\{(P_X, P_Y) \in \mathcal{P}_\kappa \colon \exists \mathcal{Y}_0 \in \mathcal{P}(\mathcal{Y}) \cap \mathfrak{B}(\mathcal{X}) \text{ with } [P_X + P_Y](\mathcal{Y}_0) = 2 \text{ s.t.}$$
$$\big|p_X(y) - p_X(z)\big| \leq C\rho^\beta(y, z), \forall y, z \in \mathcal{Y}_0\big\}, \tag{4.2}$$

with $C > 0$ and $\beta \in (0, 1]$ – analogously as in Section 3 in the regression setting. Therein $\mathcal{P}(\mathcal{Y})$ denotes the power set of $\mathcal{Y}$.

A (supervised) classifier φ is defined as a Borel measurable mapping from $\mathcal{X}^n \times \{0, 1\}^n \times \mathcal{X}$ to $\{0, 1\}$. Clearly, the sample $(Z_1, \ldots, Z_n, W_1, \ldots, W_n, Z)$ is inserted into φ and $\varphi = b$, $b = 0, 1$, means categorizing Z as a member of group b. We define the excess risk of classification by

$$\mathcal{E}_n(\varphi) := \sup_{(P_X, P_Y) \in \mathcal{P}_{C,\beta,\kappa}} \big(P_{X,Y,X}[\varphi = 1] + P_{X,Y,Y}[\varphi = 0] - 1 + \mathrm{TV}(P_X, P_Y)\big),$$

in order to evaluate the accuracy of some classifier φ. The excess risk is the sum of the probabilities of misclassification into group 0 and 1, respectively, reduced by $1 - \mathrm{TV}(P_X, P_Y)$. Therein $P_{X,Y,X}$ and $P_{X,Y,Y}$ indicate that Z has the probability measure P_X or P_Y, respectively. It is well known that the excess risk of the Bayes classifier

$$\varphi_B(\mathbf{z}, \mathbf{w}, z) := \begin{cases} 0, & \text{if } p_X(z) \geq 1/2, \\ 1, & \text{otherwise,} \end{cases}$$

vanishes if $\mathcal{P}_{C,\beta,\kappa}$ was replaced by some two-element set $\{P_X, P_Y\}$, that is, if P_X and P_Y were known.

Our goal is to find a classifier φ which minimizes the excess risk asymptotically as n, m tend to infinity. To our best knowledge optimal convergence rates for classification of functional data have been unexplored so far; whereas for finite-dimensional data they have been studied, for example, in Yang [27,28] and Audibert and Tsybakov [1]. Considering the Bayes classifier, it is reasonable to mimic the unknown densities p_X and p_Y by some appropriate estimators based on the data $Z_1, W_1, \ldots, Z_n, W_n$ (also see, e.g., Biau, Bunea and Wegkamp [4] or Ferraty and Vieu [14]). We employ the classifier

$$\varphi(Z_1, \ldots, Z_n, W_1, \ldots, W_n, Z) = \begin{cases} 0, & \text{if } \hat{p}_X(Z) \geq \hat{p}_Y(Z), \\ 1, & \text{otherwise,} \end{cases} \tag{4.3}$$

where

$$\hat{p}_X(z) := \sum_{j=1}^{n} (1 - W_j) \cdot K\big(\rho(z, Z_j)/h\big) \Big/ \sum_{j=1}^{n} (1 - W_j),$$

$$\hat{p}_Y(z) := \sum_{j=1}^{n} W_j \cdot K\big(\rho(z, Z_j)/h\big) \Big/ \sum_{j=1}^{n} W_j,$$

if $\sum_{j=1}^{n}(1 - W_j) \in (0, n)$; otherwise put $\hat{p}_X(z) = 0$ or $\hat{p}_Y(z) = 0$ by convention. Therein we apply some kernel K and bandwidth parameter $h > 0$ as in Section 3. We stipulate that enough data Z_j from both P_X and P_Y are available; concretely, we impose

$$P[W_1 = 1] = w \qquad \text{for some fixed value } w \in (0, 1). \tag{4.4}$$

The asymptotic performance of the classifier (4.3) is studied in the following theorem.

Theorem 4.1. *We consider the model of supervised classification. Let $\mathcal{Y} \in \mathfrak{B}(\mathcal{X})$ such that (2.3) holds true. Moreover, we assume (4.4). Then the excess risk of the classifier φ in (4.3) attains the following uniform upper bound*:

$$\mathcal{E}_n(\varphi) = \mathcal{O}\big((\log n)^{-\beta/\gamma}\big),$$

under the kernel choice and the bandwidth selection from Theorem 3.1.

While Theorem 4.1 can be proved directly, it follows from Theorem 3.1 by the general argument that the excess mass is bounded from above by the integrated squared regression risk (see, e.g., Devroye, Györfi and Lugosi [12], page 104). Furthermore, we mention that, in the setting of Theorem 4.1, we could relax the assumptions contained in $\mathcal{P}_{C,\beta,\kappa}$ to $\kappa = 0$. Still the condition $\kappa > 0$ is realistic as P_X and P_Y should not become too close to each other; otherwise, the classification problem makes no sense. Finally, in Theorem 4.2 we will establish optimality of the convergence rates from Theorem 4.1 with respect to an arbitrary sequence of classifiers.

Theorem 4.2. *We consider the model of supervised classification. Let $\mathcal{Y} \in \mathfrak{B}(\mathcal{X})$ such that (2.3) holds true. Moreover, we assume (4.4). Fix some $\kappa > 0$ sufficiently small (but independent of n). Let $\{\varphi_n\}_n$ be an arbitrary sequence of (supervised) classifiers where φ_n is based on the data $(Z_1, \ldots, Z_n, W_1, \ldots, W_n, Z)$. Then we have*

$$\liminf_{N \to \infty}(\log n)^{\beta/\gamma} \mathcal{E}_n(\varphi_n) > 0.$$

The optimal convergence rates in Theorems 4.1 and 4.2 correspond to those established in Section 3 in the regression problem. Note that there we consider the squared risk. Again, we realize that the bandwidth selector in Theorem 4.1 does not require knowledge of the smoothness level β and, still, it leads to the optimal speed of convergence.

5. Proofs

Proof of Lemma 2.1. Let X, Y be some independent random variables with the induced measure P. Note that $\psi(X, h)$ can be viewed as the conditional expectation of the random variable $1_{[0,h)}(\rho(X, Y))$ given X so that the random mapping $\psi(X, h)$ is measurable, thus a random variable. By the factorization lemma of the conditional expectation, the mapping $x \mapsto \psi(x, h)$, $x \in \mathcal{X}$, is measurable so that $\mathcal{Y}_{h,\delta} := \{x \in \mathcal{Y}: \psi(x, h) \leq \delta\}$ lies in $\mathfrak{B}(\mathcal{X})$. Furthermore, we obtain that

$$\mathcal{Y}_{h,\delta} = \bigcup_{j=1}^{\mathcal{N}_{\mathcal{Y}}(h/2, \mathcal{Y}, \rho)} \{y \in B_{\mathcal{Y}}(y_j, h/2): \psi(y, h) \leq \delta\},$$

where $\{y_1, \ldots, y_{\mathcal{N}_{\mathcal{Y}}(h/2, \mathcal{Y}, \rho)}\} \subseteq \mathcal{Y}$ denotes an intrinsic $h/2$-cover of $\mathcal{Y}$ with respect to the metric ρ. By J we denote the collection of all $j = 1, \ldots, \mathcal{N}_{\mathcal{Y}}(h/2, \mathcal{Y}, \rho)$ such that the set $\{y \in B_{\mathcal{Y}}(y_j, h/2): \psi(y, h) \leq \delta\}$ is not empty. For any $j \in J$, there exists some $y \in B_{\mathcal{Y}}(y_j, h/2)$ with $P(B_{\mathcal{Y}}(y, h)) \leq \delta$. We have $B_{\mathcal{Y}}(y_j, h/2) \subseteq B_{\mathcal{Y}}(y, h)$ so that $P(B_{\mathcal{Y}}(y_j, h/2)) \leq \delta$. We deduce that

$$P(\mathcal{Y}_{h,\delta}) \leq \sum_{j \in J} P(B_{\mathcal{Y}}(y_j, h/2)) \leq \delta \mathcal{N}_{\mathcal{Y}}(h/2, \mathcal{Y}, \rho) \leq \delta \exp(c_{x,1} 4^{\gamma} h^{-\gamma}),$$

when combining (2.2) and (2.3). □

Proof of Theorem 3.1. For any $g \in \mathcal{G}$ we derive that

$$E\{|\hat{g}(x) - g(x)|^2 | X_1, \dots, X_n\}$$

$$\leq 1\{\hat{B}(x) > \delta_n\}\hat{B}^{-2}(x)E\{|\hat{A}(x) - g(x)\hat{B}(x)|^2 | X_1, \dots, X_n\} + g^2(x) \cdot 1\{\hat{B}(x) \leq \delta_n\}$$

$$\leq 2C^2h^{2\beta} + 2c_v \cdot 1\{\hat{B}(x) > \delta_n\}\hat{B}^{-2}(x)n^{-2}\sum_{j=1}^{n} 1_{[0,h)}(\rho(X_j, x)) + C^2 \cdot 1\{\hat{B}(x) \leq \delta_n\} \quad (5.1)$$

$$\leq 2C^2h^{2\beta} + 2c_v n^{-1}\delta_n^{-2} + C^2 \cdot 1\{\hat{B}(x) \leq \delta_n\},$$

holds almost surely under the convention $0 \cdot \infty = 0$. We realize that $E\hat{B}(x) = \psi_n(x, h) := P_{X,n}(B_{\mathcal{Y}}(x, h))$. By the inequality,

$$1\{\hat{B}(x) \leq \delta_n\} \leq 1\{|\hat{B}(x) - \psi_n(y, h)| \geq \delta_n\} + 1\{\psi_n(x, h) \leq 2\delta_n\},$$

applying the expectation to both sides of (5.1) leads to

$$E|\hat{g}(x) - g(x)|^2 \leq 2C^2h^{2\beta} + 2c_v n^{-1}\delta_n^{-2} + C^2 \cdot \delta_n^{-2}\operatorname{var}\hat{B}(x) + C^2 \cdot 1\{\psi_n(x, h) \leq 2\delta_n\}, \quad (5.2)$$

where $\operatorname{var}\hat{B}(x) \leq n^{-1}\psi_n(x, h)$. Putting $x = X_{n+1}$ (i.e., an independent copy of $X_1, \dots, X_n$) and applying the expectation to both sides of (5.2) leads to

$$E|\hat{g}(X_{n+1}) - g(X_{n+1})|^2 \leq 2C^2h^{2\beta} + (2c_v + C^2)n^{-1}\delta_n^{-2} + C^2 P[\psi_n(X_{n+1}, h) \leq 2\delta_n].$$

Putting $\psi_n = \psi$, $X_{n+1} = X$ and $2\delta_n = \delta$, Lemma 2.1 yields that

$$P[\psi_n(X_{n+1}, h) \leq 2\delta_n] \leq 2n^{4^\gamma c_{x,1}d - \eta}.$$

Due to the constraint on d the term $2C^2h^{2\beta}$ is asymptotically dominating, which provides the desired upper bound on the considered risk with uniform constants on $g \in \mathcal{G}$. $\qquad \square$

Proof of Theorem 3.2. (a) We introduce some sequence $(\delta_n) \downarrow 0$. As $\mathcal{Y}$ satisfies (2.3), the packing number has the lower bound

$$\mathcal{D}(\delta_n, \mathcal{Y}, \rho) \geq m_n := \exp(c_{x,0}\delta_n^{-\gamma}),$$

due to (2.1) and (2.2). This implies the existence of some $z_{1,n}, \dots, z_{m_n,n} \in \mathcal{Y}$ such that the balls $B_{j,n} := B_{\mathcal{Y}}(z_{j,n}, \delta_n/4)$, $j = 1, \dots, m_n$ are pairwise disjoint. This statement can be strengthened to the result that the ρ-distance between the sets $B_{j,n}$ and $\bigcup_{k \neq j} B_{k,n}$ is even bounded from below by $\delta_n/2$. We specify $P_X = P_{X,n}$ as the discrete uniform distribution on the grid $\{z_{1,n}, \dots, z_{m_n,n}\}$.

We use the function $\vartheta(t) = \exp\{1/(t^2 - 1)\} \cdot 1_{(-1,1)}(t)$, $t \in \mathbb{R}$. Thus, ϑ is differentiable infinitely often on the whole real line, yielding that

$$|\vartheta(t) - \vartheta(s)| \leq \min\{\|\vartheta\|_\infty, \|\vartheta'\|_\infty|t - s|\} \leq \max\{\|\vartheta\|_\infty, \|\vartheta'\|_\infty\} \cdot |t - s|^\beta.$$

We construct the regression curves

$$g_\theta(x) = \sum_{j=1}^{m_n} \theta_j d h_n^\beta \vartheta\big(\rho(z_{j,n}, x)/h_n\big),$$

with the vector $\theta = (\theta_1, \ldots, \theta_{m_n}) \in \{0, 1\}^{m_n}$ and $h_n := \delta_n/4$, for some $d > 0$. As $(h_n)_n$ is bounded from above, the constraint $\sup_{g \in \mathcal{G}} \|g\|_\infty \leq C$ can be satisfied by choosing $d > 0$ small enough. For all $y_1, y_2 \in \mathcal{Y}$ there exist at most one j_1 and one j_2 such that $y_l \in B_{j_l, n}$, $l = 1, 2$. Therefore, we have

$$\big|g_\theta(y_1) - g_\theta(y_2)\big| \leq d \sum_{j=1}^{m_n} h_n^\beta \big|\vartheta\big(\rho(z_{j,n}, y_1)/h_n\big) - \vartheta\big(\rho(z_{j,n}, y_2)/h_n\big)\big|$$

$$\leq 2d\rho(y_1, y_2)^\beta \max\big\{\|\vartheta\|_\infty, \|\vartheta'\|_\infty\big\},$$

so that a sufficiently small choice of d guarantees that $g_\theta \in \mathcal{G}$ uniformly for all $\theta \in \{0, 1\}^{m_n}$.

Now we use Assouad's lemma, which is based on the common Bayesian approach of imposing the uniform distribution on $\{0, 1\}^{m_n}$ to be the a-priori distribution of θ. We refer to the book of Tsybakov [23], in particular, Section 2.7.2 for a detailed review and proof of these results. From there, it follows that

$$\sup_{g \in \mathcal{G}} E_g \|\hat{g} - g\|_{P_X}^2$$

$$\geq \frac{1}{4} d^2 h_n^{2\beta} \sum_{j=1}^{m_n} \int_{B_{j,n}} \vartheta^2\big(\rho(z_{j,n}, x)/h_n\big) \, dP_X(x) \big\{1 - E H^2\big(E_\theta f_{\theta,j,0}(y|\mathbf{X}_n), E_\theta f_{\theta,j,1}(y|\mathbf{X}_n)\big)\big\},$$

where $H^2(f_1, f_2) := \int (\sqrt{f_1} - \sqrt{f_2})^2$ denotes the squared Hellinger distance between two densities f_1 and f_2. We consider that

$$\int_{B_{j,n}} \vartheta^2\big(\rho(z_{j,n}, x)/h_n\big) \, dP_X(x) = \vartheta^2(0) m_n^{-1}.$$

Therefore, we realize that the uniform squared risk is bounded from below by a global constant times $h_n^{2\beta}$ whenever we can show that

$$\lim_{n \to \infty} \max_{j=1,\ldots,m_n} E H^2\big(E_\theta f_{\theta,j,0}(y|\mathbf{X}_n), E_\theta f_{\theta,j,1}(y|\mathbf{X}_n)\big) = 0. \qquad (5.3)$$

By the Cauchy–Schwarz inequality with respect to E_θ we deduce that

$$E H^2\big(E_\theta f_{\theta,j,0}(y|\mathbf{X}_n), E_\theta f_{\theta,j,1}(y|\mathbf{X}_n)\big) \leq E_\theta E H^2\big(f_{\theta,j,0}(y|\mathbf{X}_n), f_{\theta,j,1}(y|\mathbf{X}_n)\big).$$

We consider that

$$H^2\big(f_{\theta,j,0}(y|\mathbf{X}_n), f_{\theta,j,1}(y|\mathbf{X}_n)\big)$$

$$= 2 - 2 \prod_{k=1}^n \Big(1 - \frac{1}{2} H^2\big(f_\varepsilon(\cdot - g_{\theta,j,0}(X_k)), f_\varepsilon(\cdot - g_{\theta,j,1}(X_k))\big)\Big),$$

almost surely. Applying the expectation yields that

$$EH^2\big(E_\theta f_{\theta,j,0}(y|\mathbf{X}_n), E_\theta f_{\theta,j,1}(y|\mathbf{X}_n)\big)$$

$$\leq 2 - 2E_\theta\left(1 - \frac{1}{2}EH^2\big(f_\varepsilon\big(\cdot - g_{\theta,j,0}(X_1)\big), f_\varepsilon\big(\cdot - g_{\theta,j,1}(X_1)\big)\big)\right)^n$$

$$\leq 2 - 2E_\theta\left(1 - \frac{1}{8}E\big|g_{\theta,j,1}(X_1) - g_{\theta,j,0}(X_1)\big|^2 \cdot \int \big|f_\varepsilon'(t)\big|^2/f_\varepsilon(t)\,dt\right)^n,$$

where, for all $\theta \in \{0, 1\}^{m_n}$ and $j = 1, \ldots, m_n$, we have

$$E\big|g_{\theta,j,1}(X_1) - g_{\theta,j,0}(X_1)\big|^2 = d^2 h_n^{2\beta} \int_{B_{j,n}} \vartheta^2\big(\rho(z_{j,n}, x)/h_n\big)\,dP_X(x) = d^2 h_n^{2\beta}\vartheta^2(0)m_n^{-1}.$$

Therefore, recalling that $h_n = \delta_n/4$ we put $\delta_n = \{c_h \log n\}^{-1/\gamma}$ for some $c_h > 1/c_{x,0}$ so that $h_n^{2\beta} m_n^{-1} = o(1/n)$ and (5.3) is fulfilled. This provides the desired lower bound.

(b) We take m_n, the $z_{1,n}, \ldots, z_{m_n,n} \in \mathcal{Y}$ and the balls $B_{j,n}$ from the proof of part (a). As the $B_{j,n}$ are pairwise disjoint we have that

$$\sum_{j=1}^{m_n} P_{X,n}(B_{j,n}) = P_{X,n}\left(\bigcup_{j=1}^{m_n} B_{j,n}\right) \leq 1,$$

so that, for at least one $k_n \in \{1, \ldots, m_n\}$, we have $P_{X,n}(B_{k_n,n}) \leq 1/m_n$ where $z_{k_n,n} \in \mathcal{Y}$. To simplify the notation, we write $w_n := z_{k_n,n}$.

We consider the mappings $g_0: \equiv 0$ and $g_n(z) := dh_n^\beta \vartheta(\rho(w_n, z)/h_n)$ with $h_n = \delta_n/4$ on the domain $\mathcal{X}$ with ϑ as in the proof of (a). Again, choosing d small enough ensures that $g_0, g_n \in \mathcal{G}$ for all n.

We define

$$\alpha_n := \big|g_n(w_n) - g_0(w_n)\big|/2 = d\vartheta(0)h_n^\beta/2,$$

and the events

$$H_n(g) := \big\{\omega \in (\mathcal{X} \times \mathbb{R})^n : \big|\hat{g}_n(w_n, \omega) - g(w_n)\big| \geq \alpha_n\big\}.$$

Also we write $\mathbf{X}_n := (X_1, \ldots, X_n)$. We deduce that

$$\sup_{g \in \mathcal{G}} \sup_{y \in \mathcal{Y}} P_g\big[\big|\hat{g}_n(y, \mathcal{Z}_n) - g(y)\big| > d\vartheta(0)h_n^\beta/2\big]$$

$$\geq \sup_{g \in \mathcal{G}} E P_g\big(H(g)|\mathbf{X}_n\big)$$

$$\geq \frac{1}{2} \cdot E\big\{P_{g_0}\big(H(g_0)|\mathbf{X}_n\big) + P_{g_n}\big(H(g_n)|\mathbf{X}_n\big)\big\}$$

$$\geq \frac{1}{2} - \frac{1}{2} \cdot E\,\mathrm{TV}(P_{g_0}|\mathbf{X}_n, P_{g_n}|\mathbf{X}_n),$$

(5.4)

where $\mathrm{TV}(P, Q)$ denotes the total variation distance between some probability measures P and Q. Note that the conditional probability measure $P_g|\mathbf{X}_n$ just turns out to be the probability measure of independent random variables δ_j, $j = 1, \ldots, n$, with the density $f_\varepsilon(\cdot - g(X_j))$ conditionally on $\mathbf{X}_n$. By LeCam's inequality, we have

$$\mathrm{TV}(P_{g_0}|\mathbf{X}_n, P_{g_n}|\mathbf{X}_n) \leq \left\{ 1 - \prod_{j=1}^{n} \left(1 - \frac{1}{2} H^2 \big(f_\varepsilon(\cdot - g_0(X_j)), f_\varepsilon(\cdot - g_n(X_j)) \big) \right)^2 \right\}^{1/2},$$

almost surely, where $H(f_1, f_2)$ denotes the Hellinger distance between two densities f_1 and f_2. Applying the expectation to both sides, Jensen's inequality and some information theoretic arguments yield that

$$E\,\mathrm{TV}(P_{g_0}|\mathbf{X}_n, P_{g_n}|\mathbf{X}_n) \leq \left\{ 1 - \left(1 - \frac{1}{8} E \big(g_n(X_1) - g_0(X_1) \big)^2 \int |f_\varepsilon'(x)|^2 / f_\varepsilon(x)\,\mathrm{d}x \right)^{2n} \right\}^{1/2}.$$

Since the restrictions of g_n and g_0 to the domain $\mathcal{Y}$ coincide on $\mathcal{Y} \setminus B_{\mathcal{Y}}(w_n, h_n)$ we deduce that

$$E \big(g_n(X_1) - g_0(X_1) \big)^2 \leq d^2 h_n^{2\beta} \|\vartheta\|_\infty^2 P_{X,n}\big(B_{\mathcal{Y}}(w_n, h_n)\big) \leq d^2 h_n^{2\beta} m_n^{-1} \|\vartheta\|_\infty^2,$$

so that (5.4) is bounded away from zero whenever $h_n^{2\beta} m_n^{-1} = \mathcal{O}(1/n)$. Under the selection of h_n and δ_n as in part (a) this condition is satisfied. That completes the proof. $\qquad \square$

Proof of Theorem 4.2. The inequality (2.3) yields that the set $\mathcal{Y}$ contains infinitely many elements. Fix three different $y_1, y_2, y_3 \in \mathcal{Y}$. We introduce the sets $\mathcal{Y}_j$, $j = 1, 2, 3$, with

$$\mathcal{Y}_j := \big\{ y \in \mathcal{Y} \colon \rho(y, y_j) \leq \rho(y, y_k), \forall k \in \{1, 2, 3\} \big\},$$

whose union includes $\mathcal{Y}$ as a subset. Note that

$$\tfrac{1}{3} \mathcal{N}_{\mathcal{X}}(\delta, \mathcal{Y}, \rho) \leq \mathcal{N}_{\mathcal{X}}(\delta, \mathcal{Y}_j, \rho) \leq \mathcal{N}_{\mathcal{X}}(\delta, \mathcal{Y}, \rho), \tag{5.5}$$

holds true for all $\delta > 0$ for at least one $j = 1, 2, 3$. Select $j = 1, 2, 3$ such that $\mathcal{Y}' := \mathcal{Y}_j$ satisfies the above inequality; and put z_{-1}, z_0 equal to the other y_k, $k \neq j$. Note that

$$\rho(y, z_l) \geq M := \min\big\{ \rho(y_r, y_s) \colon r \neq s \big\} / 2,$$

holds for all $y \in \mathcal{Y}'$ and $l = 0, -1$. Clearly, we have $\rho(z_0, z_{-1}) \geq M$ as well. The inequalities (2.1), (2.2), (2.3) and (5.5) yield the existence of $z_1, \ldots, z_{d_n} \in \mathcal{Y}'$ with some even number $d_n \geq \lfloor \exp(c_{x,0} \delta_n^{-\gamma})/3 \rfloor - 1$ such that $\rho(z_j, z_k) > \delta_n$ for any sequence $(\delta_n)_n \downarrow 0$. Therefore, the balls $B_{\mathcal{Y}}(z_j, \delta_n/4)$, $j = 1, \ldots, d_n$ are pairwise disjoint. By R_n we denote the discrete probability measure which fulfills

$$R_n(\{z_0\}) = R_n(\{z_{-1}\}) = 2\kappa M_0^{-\beta}/C,$$

$$R_n(\{z_j\}) = \big(1 - 4\kappa M_0^{-\beta}/C\big)/d_n, \qquad j = 1, \ldots, d_n,$$

and $M_0 := \min\{C^{-1/\beta}, M\}$. We define the functions

$$f_\theta(y) := 1 + \theta_0 \frac{1}{2} C M_0^\beta \left(1_{\{z_{-1}\}}(y) - 1_{\{z_0\}}(y)\right) + \frac{1}{2} C \delta_n^\beta \sum_{j=1}^{d_n/2} \theta_j \left(1_{\{z_{2j-1}\}}(y) - 1_{\{z_{2j}\}}(y)\right),$$

where $\theta := (\theta_0, \ldots, \theta_{d_n/2})$ denotes some binary vector. Therein we stipulate that

$$\kappa \in \left(0, M_0^\beta C/8\right), \tag{5.6}$$

which does not depend on n. For n sufficiently large, f_θ is bounded by $1/2$ from below and by $3/2$ from above – uniformly with respect to the vector θ. Furthermore, the functions f_θ integrate to one with respect to the probability measure R_n so that the functions f_θ are probability densities. The probability measure generated by f_θ is denoted by P_θ.

We write θ' for the corresponding vector $\theta' := (1 - \theta_0, \ldots, 1 - \theta_{d_n/2})$. Then

$$\mathrm{TV}(P_\theta, P_{\theta'}) = \frac{1}{2} \int \left|f_\theta(y) - f_{\theta'}(y)\right| \mathrm{d}R_n(y) \geq \frac{1}{4} \left(C M_0^\beta R_n(\{z_0\}) + C M_0^\beta R_n(\{z_{-1}\})\right) = \kappa,$$

so that $(P_\theta, P_{\theta'}) \in \mathcal{P}_\kappa$. Furthermore, we have

$$\frac{\mathrm{d}P_\theta}{\mathrm{d}(P_\theta + P_{\theta'})}(y) = \frac{f_\theta(y)}{f_\theta(y) + f_{\theta'}(y)},$$

so that

$$\sup_{\theta'' = \theta, \theta'} \left|\frac{\mathrm{d}P_{\theta''}}{\mathrm{d}(P_\theta + P_{\theta'})}(y) - \frac{\mathrm{d}P_{\theta''}}{\mathrm{d}(P_\theta + P_{\theta'})}(x)\right| \leq \max\left\{\left|f_\theta(x) - f_\theta(y)\right|, \left|f_{\theta'}(x) - f_{\theta'}(y)\right|\right\}.$$

For n sufficiently large (precisely, for $\delta_n < 2^{-1/\beta} M_0$), we can verify that

$$\left|f_\theta(x) - f_\theta(y)\right| \leq C \rho(x, y)^\beta,$$

for all $x, y \in \mathcal{Y}_0 = \{z_{-1}, z_0, \ldots, z_{d_n}\}$ and all $\theta \in \{0, 1\}^{d_n/2+1}$. This provides that $(P_\theta, P_{\theta'}) \in \mathcal{P}_{C,\beta,\kappa}$ for all $\theta \in \{0, 1\}^{d_n/2+1}$ for n large enough.

The underlying statistical experiment is less informative than the model, in which exactly n i.i.d. training data are drawn for each group, that is, we observe the samples $X_1, \ldots, X_n$ from P_X and $Y_1, \ldots, Y_n$ from P_Y. Therefore, as we are proving a lower bound, we may switch to the latter statistical model. As in the proof of Theorem 3.2(a), we apply Assouad's lemma (see, e.g., Tsybakov [23]) and the Bretagnolle–Huber inequality (see Bretagnolle and Huber [6]), which yield that

$$\mathcal{E}_n(\varphi_n) \geq \frac{1}{4} \sum_{b=0}^{1} \sum_{l=0}^{d_n/2} R_n(\{z_{2l}\}) \cdot C \left(M_0^\beta 1_{\{0\}}(l) + \delta_n^\beta 1_{(0,\infty)}(l)\right)$$

$$\times \left(1 - E_\theta \left\{1 - \exp\left(-n\mathcal{K}(f_{\theta,l,1}, f_{\theta,l,0}) - n\mathcal{K}(f_{\theta',l,0}, f_{\theta',l,1})\right)\right\}^{1/2}\right),$$

where $\mathcal{K}$ denotes the Kullback–Leibler distance between some densities. For $l \geq 1$, we deduce that

$$
\begin{aligned}
\mathcal{K}(f_{\theta,l,1}, f_{\theta,l,0}) &= \int \left\{ \log \frac{f_{\theta,l,1}(x)}{f_{\theta,l,0}(x)} \right\} f_{\theta,l,1}(x) \, \mathrm{d}R_n(x) \\
&\leq \int \frac{f_{\theta,l,1}(x)}{f_{\theta,l,0}(x)} \left| f_{\theta,l,1}(x) - f_{\theta,l,0}(x) \right| \mathrm{d}R_n(x) \qquad (5.7) \\
&\leq 3C\delta_n^\beta R_n(\{z_{2l}\}) \leq 3C\big(1 - 4\kappa M_0^{-\beta}/C\big)\delta_n^\beta d_n^{-1},
\end{aligned}
$$

almost surely. The same upper bound can be established for $\mathcal{K}(f_{\theta',l,0}, f_{\theta',l,1})$ analogously. Now we specify

$$
\delta_n = (c_\delta \log n)^{-1/\gamma},
$$

for some constant $c_\delta > 0$. Choosing c_δ sufficiently large, (5.7) yields that

$$
\begin{aligned}
\liminf_{n\to\infty} \delta_n^{-\beta} \mathcal{E}_n(\varphi_n) &\geq \liminf_{n\to\infty} \frac{1}{4} C \sum_{b=0}^{1} \sum_{l=1}^{d_n/2} R_n(\{z_{2l}\}) \\
&\geq \frac{1}{4}\big(1 - 4\kappa M_0^{-\beta}/C\big) \geq \frac{1}{8},
\end{aligned}
$$

when using (5.6) in the last step. The selection of δ_n completes the proof. $\qquad\square$

Acknowledgements

References

[1] Audibert, J.-Y. and Tsybakov, A.B. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.* **35** 608–633. MR2336861

[2] Benko, M., Härdle, W. and Kneip, A. (2009). Common functional principal components. *Ann. Statist.* **37** 1–34. MR2488343

[3] Berlinet, A., Elamine, A. and Mas, A. (2011). Local linear regression for functional data. *Ann. Inst. Statist. Math.* **63** 1047–1075. MR2822967

[4] Biau, G., Bunea, F. and Wegkamp, M.H. (2005). Functional classification in Hilbert spaces. *IEEE Trans. Inform. Theory* **51** 2163–2172. MR2235289

[5] Biau, G., Cérou, F. and Guyader, A. (2010). Rates of convergence of the functional k-nearest neighbor estimate. *IEEE Trans. Inform. Theory* **56** 2034–2040. MR2654492

[6] Bretagnolle, J. and Huber, C. (1979). Estimation des densités: Risque minimax. *Z. Wahrsch. Verw. Gebiete* **47** 119–137. MR0523165

[7] Butucea, C. and Tsybakov, A.B. (2007). Sharp optimality in density deconvolution with dominating bias. I. *Teor. Veroyatn. Primen.* **52** 111–128. MR2354572

[8] Carroll, R.J., Delaigle, A. and Hall, P. (2013). Unexpected properties of bandwidth choice when smoothing discrete data for constructing a functional data classifier. *Ann*. *Statist*. **41** 2739–2767. MR3161446

[9] Chen, D. and Müller, H.-G. (2012). Nonlinear manifold representations for functional data. *Ann*. *Statist*. **40** 1–29. MR3013177

[10] Delaigle, A. and Hall, P. (2012). Achieving near perfect classification for functional data. *J. R. Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **74** 267–286. MR2899863

[11] Delaigle, A. and Hall, P. (2013). Classification using censored functional data. *J. Amer. Statist*. *Assoc*. **108** 1269–1283. MR3174707

[12] Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. *Applications of Mathematics* (*New York*) **31**. New York: Springer. MR1383093

[13] Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann*. *Statist*. **19** 1257–1272. MR1126324

[14] Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis*: *Theory and Practice*. *Springer Series in Statistics*. New York: Springer. MR2229687

[15] Ferraty, F., Laksaci, A., Tadj, A. and Vieu, P. (2010). Rate of uniform consistency for nonparametric estimates with functional variables. *J. Statist*. *Plann*. *Inference* **140** 335–352. MR2558367

[16] Forzani, L., Fraiman, R. and Llop, P. (2012). Consistent nonparametric regression for functional data under the Stone–Besicovitch conditions. *IEEE Trans*. *Inform*. *Theory* **58** 6697–6708. MR2991802

[17] Gao, F. and Wellner, J.A. (2007). Entropy estimate for high-dimensional monotonic functions. *J. Multivariate Anal*. **98** 1751–1764. MR2392431

[18] Hall, P. and Horowitz, J.L. (2007). Methodology and convergence rates for functional linear regression. *Ann*. *Statist*. **35** 70–91. MR2332269

[19] Kolmogorov, A.N. and Tihomirov, V.M. (1961). ε-entropy and ε-capacity of sets in functional space. *Amer. Math*. *Soc*. *Transl*. (2) **17** 277–364. MR0124720

[20] Mas, A. (2012). Lower bound in regression for functional data by representation of small ball probabilities. *Electron. J. Stat*. **6** 1745–1778. MR2988463

[21] Meister, A. (2011). Asymptotic equivalence of functional linear regression and a white noise inverse problem. *Ann*. *Statist*. **39** 1471–1495. MR2850209

[22] Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*, 2nd ed. *Springer Series in Statistics*. New York: Springer. MR2168993

[23] Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. New York: Springer. MR2724359

[24] van de Geer, S.A. (2000). *Applications of Empirical Process Theory*. *Cambridge Series in Statistical and Probabilistic Mathematics* **6**. Cambridge: Cambridge Univ. Press. MR1739079

[25] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*. *Springer Series in Statistics*. New York: Springer. MR1385671

[26] Vapnik, V.N. (1998). *Statistical Learning Theory*. *Adaptive and Learning Systems for Signal Processing*, *Communications*, *and Control*. New York: Wiley. MR1641250

[27] Yang, Y. (1999). Minimax nonparametric classification. I. Rates of convergence. *IEEE Trans*. *Inform*. *Theory* **45** 2271–2284. MR1725115

[28] Yang, Y. (1999). Minimax nonparametric classification. II. Model selection for adaptation. *IEEE Trans*. *Inform*. *Theory* **45** 2285–2292.