Statistica Sinica Preprint No: SS-2022-0258	
Title	Fast Convergence on Perfect Classification for Functional Data
Manuscript ID	SS-2022-0258
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0258
Complete List of Authors	Tomoya Wakayama and Masaaki Imaizumi
Corresponding Authors	Masaaki Imaizumi
E-mails	imaizumi@g.ecc.u-tokyo.ac.jp

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

Fast Convergence on Perfect Classification for Functional Data

Tomoya Wakayama[†], Masaaki Imaizumi[†,‡]

[†]*The University of Tokyo,* [‡]*RIKEN AIP*

Abstract: We investigate perfect classification on functional data using finite samples. Perfect classification for functional data is easier to achieve than for finite-dimensional data, because a sufficient condition for the existence of a perfect classifier, called the Delaigle–Hall condition, is available only for functional data. However, a large sample size is required to achieve perfect classification, even when the Delaigle–Hall condition holds, because the minimax convergence rate of the errors with functional data has a logarithm order in the sample size. We resolve this complication by proving that the Delaigle–Hall condition also achieves fast convergence of the misclassification error in a sample size under the bounded entropy condition on functional data. We study a reproducing kernel Hilbert space-based classifier under the Delaigle–Hall condition, and show that the convergence rate of its misclassification error has an exponential order in the sample size. Technically, our proof is based on (i) connecting the Delaigle–Hall condition and a margin of classifiers, and (ii) handling metric entropy of functional data. The results of our experiments support our findings, and show that other classifiers for functional data have a similar property.

Key words and phrases: Convergence Rate, Functional Data, Perfect Classification, Reproducing Kernel Hilbert Space

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

1. Introduction

The classification problem is one of the most general and significant problems in functional data analysis. The goal of classification is to predict labels or categories from functional data given in the form of (possibly) infinite-dimensional random curves. Because of its versatility, classification has many applications in the field of science (Varughese *et al.*, 2015), engineering (Gannaz, 2014; Li *et al.*, 2013; Florindo *et al.*, 2011), medicine (Chang *et al.*, 2014; Islam, 2020; Dai *et al.*, 2017), and others. Several methods have been developed to solve this problem, including distance-based (Alonso *et al.*, 2012; Ferraty and Vieu, 2003), k-nearest neighbor (Biau *et al.*, 2005; Cérou and Guyader, 2006), partially least square (Preda *et al.*, 2007; Preda and Saporta, 2005), orthonormal basis (Delaigle and Hall, 2013, 2012), Bayesian (Wang and Qu, 2014; Yang *et al.*, 2014), and logistic regression methods (Araki *et al.*, 2009). For a survey, see Cuevas (2014).

Perfect classification for functional data was studied by Delaigle and Hall (2012), and has the advantage of being able to use infinite-dimensional data. This notion refers to the convergence of the misclassification error to zero under an optimal classifier, which is also referred to as realizability (Shalev-Shwartz and Ben-David, 2014). Seminal works (Delaigle and Hall, 2012, 2013) show that under certain conditions on the mean and covariance functions of func-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

tional data (hereafter, the *Delaigle–Hall condition*), there exists a classifier that achieves perfect classification asymptotically. This result does not usually hold with finite-dimensional data, which is why it focuses on infinite-dimensional vectors, called functional data. Berrendero *et al.* (2018) describe a relation between this notion and a reproducing kernel, and various methods have been shown to have a connection to a perfect classifier (Cérou and Guyader, 2006; Dai *et al.*, 2017; Cuesta-Aboertos and Dutta, 2016; Hanneke *et al.*, 2021).

One difficulty is the need for a large sample size to achieve perfect classification, suggested by a convergence analysis of the misclassification error in the sample size. Nonparametric methods for functional classification are known to have a very slow convergence rate, owing to the infinite dimensionality of functional data. Let $R(f)$ be a misclassification error under a classifier f. Meister (2016) proves that any classifier $\widetilde{f}_n$ consisting of n observations has the following relationship with some data-generating process:

$$R(\widetilde{f}_n) - \inf_f R(f) \geq c(\log n)^{-\alpha}.$$

Here, $c > 0$ is a universal constant, and $\alpha > 0$ is a parameter that depends on the data-generating process. This result shows that the misclassification error of functional data cannot avoid errors that decay only on the logarithmic order in the general setting. Because logarithmic decay is slower than every decay with a polynomial order, the convergence of this unavoidable error is very slow. As

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

such, even if a perfect classifier exists, it can be difficult to benefit from it.

This study resolves the aforementioned possibility by showing that the Delaigle–Hall condition also makes the convergence of the excess misclassification error sufficiently fast. To achieve our goal, we consider a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ and study a classifier $\widehat{f}_n \in \mathcal{H}$ from n observations using empirical loss minimization. Furthermore, we consider a family of functional data that satisfies a bounded entropy condition, implying the continuity and the boundedness of a norm of such data. Then, we show that $\widehat{f}_n$ obtains the following convergence under the Delaigle–Hall condition:

$$E\left(R(\widehat{f}_n) - \inf_{f\in\mathcal{H}} R(f)\right) \leq 2\exp(-\beta n),$$

with some parameter $\beta > 0$. This exponential convergence in n is faster than all polynomial convergence, ensuring perfect classification.

Note that the classifier $\widehat{f}_n$ is constructed as a linear sum of given kernel functions. Functional data analysis using an RKHS is widely used in both linear and nonlinear regression problems (Preda, 2007; Lian, 2007; Cai and Yuan, 2012; Cui *et al.*, 2020; Tian *et al.*, 2020), but is not widely used for the classification problem, with the exception of Rincón and Ruiz-Medina (2012). Note that our approach differs from that of Berrendero *et al.* (2018), who considers functional data as RKHSs, because we construct the classifier using RKHSs.

As a technical contribution, our theoretical results follow two ideas. First,

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

we introduce a *hard-margin condition*, which describes the ease of classification problems, and connect it to the Delaigle–Hall condition. In a general setting, a hard-margin condition is suitable for a perfectly classifiable setting, such as Koltchinskii and Beznosova (2005). We newly develop a new hard-margin condition for functional data, and then prove that the Delaigle–Hall condition implies the hard-margin condition by using covariance structures of functional data. Second, we develop a metric entropy analysis on a classifier for functional data. To analyze the speed of convergence of the empirical risk minimization classifier, we study the excess empirical risk on a space of classifiers. However, because classifiers for functional data are more complicated than those of ordinary cases, we cannot use the traditional theoretical results. We derive a new entropy bound for this purpose, which enables us to develop the theory. To the best of our knowledge, both technical points are new theoretical results.

Note the bounded entropy condition on the functional data for our result. First, the condition requires a kind of continuity of the functional data, for example, Lipschitz continuity. Second, it requires that a norm of the functional data is bounded almost surely, which excludes, for example, Gaussian processes. These restrictions are necessary for our proof with an entropy condition. To clarify this point, we provide several examples of stochastic processes that satisfy the entropy condition.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

We conduct numerical experiments to confirm our theoretical findings. Our results show that the convergence speed of the misclassification error by the RKHS method varies depending on whether or not the Delaigle–Hall and hard-margin conditions are satisfied. We also test several additional classification methods for functional data under the conditions. The results show that the RKHS method and nonparametric classification methods, such as the Gaussian process method, give similar effects on their convergence rates, but that linear methods, such as a linear discriminant analysis, do not cause such an effect.

The remainder of the paper is organized as follows. Section 2 introduces our setting and method. Section 3 explains the perfect classification and its convergence result. In Section 4, we confirm our theoretical result using experiments. Section 5 concludes the paper. The online Supplementary Material contains proofs and additional examples.

1.1 Notation

For $r \in \mathbb{R}$, $\mathrm{sign}(r)$ is a sign function that is 1 if $r > 0$, -1 if $r < 0$, and 0 if $r = 0$. In addition, $\lceil r \rceil$ denotes the largest integer which is no more than r. For $r, r' \in \mathbb{R}$, $r \vee r' = \max\{r, r'\}$. For a function $f : \Omega \to \mathbb{R}$ on a set Ω, $\|f\|_{L^\infty} = \sup_{x\in\Omega} |f(x)|$ denotes a sup-norm, and $\|f\|_n^2 = n^{-1} \sum_{i=1}^n f(X_i)^2$ is an empirical norm with observations $X_1, ..., X_n$. For an event $\mathcal{E}$, $\mathbf{1}\{\mathcal{E}\}$ is an

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

indicator function that is one if $\mathcal{E}$ holds, and zero otherwise. For two sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$, $a_n \gtrsim b_n$ denotes that there exists a constant $c > 0$ such that $a_n \geq cb_n$, for all $n \geq \overline{n}$, with some finite $\overline{n} \in \mathbb{N}$; $a_n \lesssim b_n$ denotes its opposite, and $a_n \asymp b_n$ means that both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold. For a variable z, let C_z be some positive and finite constant that depends only on z. For a space Ω with a distance d and $\delta > 0$, let $\mathcal{N}(\delta, \Omega, d)$ be the covering number of $\mathcal{H}$, that is, the minimal number of balls that cover Ω with the radius δ in terms of d.

2. Preliminary

2.1 Problem Setting

We consider a functional classification problem. Let $\mathcal{X}$ be a subset of an L^2-space on an index set $\mathcal{T} \subset \mathbb{R}^d$, with some $d \geq 1$, and consider its inner product $\langle x, x' \rangle = \int_{\mathcal{T}} x(t)x'(t)dt$, with $x, x' \in \mathcal{X}$, and its induced L^2-norm $\|\cdot\|$. Let $\mathcal{B}(\mathcal{X})$ be an associated Borel σ-field of $\mathcal{X}$. Suppose we have observations $(X_1, Y_1), ..., (X_n, Y_n)$ that are n independent copies of a random object (X, Y) from a joint distribution P, where X is an $\mathcal{X}$-valued random function and Y is a $\{-1, 1\}$-valued discrete random label. We write $w = P(Y = -1) \in (0, 1)$. Let Π on $\mathcal{B}(\mathcal{X})$ be a marginal measure of X. For each label, we define conditional measures on $\mathcal{B}(\mathcal{X})$ for functional data as $P_+ = \Pi(\cdot \mid Y = 1)$ and $P_- = \Pi(\cdot \mid Y = -1)$. By the definitions, we obtain $P_+(\mathcal{X}) = P_-(\mathcal{X}) = 1$ and

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

$\Pi = wP_- + (1-w)P_+$.

The goal of this problem is to construct a classifier that outputs a label from a functional input in $\mathcal{X}$. For a given function $f : \mathcal{X} \to \mathbb{R}$, a corresponding binary classifier is defined as $\text{sign} \circ f$. We define the misclassification error of f as

$$R(f) = P\{(X, Y) : Y \neq \text{sign}(f(X))\},$$

which is also referred as a generalization error of the classification problem.

We discuss the existence of a minimizer of $R(f)$ that is an optimal function for the Bayes classifier. To this end, we develop a density function of P_+ and P_-. Unlike the finite-dimensional data case, it is not trivial to define the densities, because function spaces do not have the useful Lebesgue measure. Instead, we use Π as a base measure, which is absolute continuous to P_+ and P_-, and define the following densities by the Radon–Nikodym derivative $p_+ = dP_+/d\Pi$ and $p_- = dP_-/d\Pi$. The following result shows that we can guarantee the function as a minimizer with p_+ and p_-. The proof is deferred to the Supplementary Material.

Lemma 1. *We define a function* $f_0 : \mathcal{X} \to \mathbb{R}$ *as* $f_0(x) = (1-w)p_+(x) - wp_-(x)$. *Then,* f_0 *minimizes* $R(f)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

2.2 Methodology: RKHS Classifier

We provide a setting for an RKHS for functional data. Let $\mathcal{H}$ be a Hilbert space on $\mathcal{X}$. In addition, let $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ be an inner product of $\mathcal{H}$, and let $\| \cdot \|_{\mathcal{H}}$ be an induced norm of $\mathcal{H}$. A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is referred to as a reproducing kernel for $\mathcal{H}$ if it satisfies (i) for every $x \in \mathcal{X}$, $K(\cdot, x) \in \mathcal{H}$ holds, and (ii) for every $x \in \mathcal{X}$ and $f \in \mathcal{H}$, $f(x) = \langle f, K(\cdot, x) \rangle_{\mathcal{H}}$ holds. It is well known that a reproducing kernel K is symmetric, nonnegative definite and uniquely determined by an RKHS $\mathcal{H}$. Furthermore, a set of linear form $\{\sum_{i=1}^{n} c_i K(x_i, \cdot) : c_i \in \mathbb{R}, x_i \in \mathcal{X}\}$ is dense in $\mathcal{H}$; see Berlinet and Thomas-Agnan (2011).

An important property of an RKHS is, for any $x, x' \in \mathcal{X}$ and $f \in \mathcal{H}$, there exists a constant $c_{\mathcal{H}}$, such that

$$|f(x)| \leq c_{\mathcal{H}} \|f\|_{\mathcal{H}}, \text{ and } |f(x) - f(x')| \leq \|f\|_{\mathcal{H}} \|x - x'\| \tag{2.1}$$

holds. For the proof, see Proposition 4.30 in Steinwart and Christmann (2008). Hereafter, we set $c_{\mathcal{H}} = 1$, without loss of generality. In addition, we impose the condition that $\mathcal{H}$ is dense in a set of continuous functions $C(\mathcal{X})$. This property is referred to as *universality*, and is satisfied by many common RKHSs (see Definition 4.52 and Corollary 4.55 in Steinwart and Christmann (2008)).

For binary classification problems, we define a classifier $\widehat{f}_n \in \mathcal{H}$. Let $\ell : \mathbb{R} \to \mathbb{R}$ be a loss function, such that $\ell \geq \mathbf{1}_{(-\infty, 0]}$ is decreasing, bounded by one,

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

convex, and 1-Lipschitz continuous. We consider the following optimization problem:

$$\widehat{f}_n = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i f(X_i)) + \lambda \|f\|_{\mathcal{H}}^2 , \tag{2.2}$$

where $\lambda > 0$ is a regularization coefficient. This minimization problem can be solved in several ways, depending on the loss function. We further assume that ℓ is twice differentiable, its first derivative ℓ' is negative, increasing, and bounded below by -1, and its second derivative ℓ'' is bounded above by 1. The logit loss $\ell(t) = \log(1 + \exp(-t))$ is a common choice of a loss function that satisfies this requirement.

3. Main Result on Convergence of Misclassification Error

Our aim is to study the excess risk $R(\widehat{f}_n) - \inf_{f \in \mathcal{H}} R(f)$ under a perfect classifiable setting. To this end, we require an assumption for perfect classification.

3.1 Delaigle–Hall Condition for Perfect Classification

The Delaigle–Hall condition, established by Delaigle and Hall (2012, 2013), is a condition for functional data to be asymptotically perfect classifiable.

We provide some notation. Consider covariance functions of $X \sim \Pi$ as $C(t, t') = \operatorname{cov}(X(t), X(t'))$, which is assumed to be strictly positive definite and uniformly bounded. Further, random functions X_+ and X_- are drawn from

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

P_+ and P_-, respectively, with a corresponding bounded covariance function $C_\ell(t,t') = \mathrm{cov}(X_\ell(t), X_\ell(t'))$, for $\ell \in \{-,+\}$. Their spectral decompositions (e.g., Theorem 4.6.5 in Hsing and Eubank (2015)) are written as

$$C(t,t') = \sum_{j=1}^{\infty} \theta_j \phi_j(t)\phi_j(t') \text{ and } C_\ell(u,v) = \sum_{j=1}^{\infty} \theta_{\ell j}\phi_{\ell j}(u)\phi_{\ell j}(v),$$

where $\{\theta_{\ell j}, \phi_{\ell j}\}_{j=1}^{\infty}$, and $\{\theta_j, \phi_j\}_{j=1}^{\infty}$ are pairs of nonzero eigenvalues and eigenfunctions of C_ℓ and C, for $\ell \in \{-,+\}$. We assume that they are sorted as $\theta_1 \geq \theta_2 \geq \ldots$ and $\theta_{\ell,1} \geq \theta_{\ell,2} \geq \ldots$. We also introduce coefficients of mean functions for each label. Let X_+ and X_- be random functions generated from P_+ and P_-, respectively. Then we define its mean as

$$\mu_+ := E_{P_+}[X_+] = \sum_{j=1}^{\infty} \mu_{+,j}\phi_j, \text{ and } \mu_- := E_{P_-}[X_-] = \sum_{j=1}^{\infty} \mu_{-,j}\phi_j,$$

with coefficients $\mu_{+,j}$ and $\mu_{-,j}$, for $j \in \mathbb{N}$, by the generalized Fourier decomposition. Using the basis $\{\phi_j\}_{j=1}^{\infty}$, we express the difference $\mu_+ - \mu_- = \sum_{j=1}^{\infty} \mu_j \phi_j$ by coefficients μ_j, for $j \in \mathbb{N}$.

We now introduce a condition for perfect classification. The following condition is developed in section 4.2 of Delaigle and Hall (2012):

Definition 1 (Delaigle–Hall condition (Delaigle and Hall, 2012))**.** The joint measure P satisfies the Delaigle–Hall conditions if the following holds for $\ell \in \{+,-\}$:

$$\lim_{M\to\infty} \frac{(\sum_{j=1}^{M} \theta_j^{-1}\mu_j^2)^2}{\sum_{j=1}^{\infty} \theta_{\ell j}(\sum_{i=1}^{M} \theta_i^{-1}\mu_i \int \phi_i(u)\phi_{\ell j}(u)du)^2} = \infty. \tag{3.1}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

We can simplify condition (3.1) under a specific setting: if C_+ and C_- have common eigenfunctions, that is, $\phi_j = \phi_{+j} = \phi_{-j}$, then the condition (3.1) is rewritten more intuitively as

$$\sum_{j=1}^{\infty} \theta_j^{-1} \mu_j^2 = \infty.$$

This condition indicates that the covariance of functional data is too concise compared with the mean difference. If the functional data are nearly independent for each input, the Delaigle–Hall condition is more likely to be satisfied, because θ_j decays faster as j increases. The Delaigle–Hall condition implies the following result.

Proposition 1 (Theorem 1 in Delaigle and Hall (2012))**.** *If the Delaigle–Hall condition is satisfied and $X \mid Y$ is Gaussian, then there is a perfect classification; that is,* $\inf_f R(f) = 0$ *holds.*

A similar result holds without the Gaussianity of $X|Y$ (see Theorem 2 in Delaigle and Hall (2012)).

This Delaigle–Hall condition gives a sufficient condition for Gaussian measures on infinite-dimensional spaces to be mutually singular, based on the classical Hájek–Feldman theorem Da Prato and Zabczyk (2014). Because this type of singularity appears more easily than on finite-dimensional spaces, this shows one advantage of using infinite-dimensional functional data. Next, we provide an

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

example of functional data distributions that satisfy the Delaigle–Hall condition.

Example 1 (Decaying Coefficients)**.** We give a specific example of θ_j and μ_j, and consider when the Delaigle–Hall condition is satisfied. Suppose that $\theta_j \asymp j^{-\alpha}$ with $\alpha > 0$ and $\mu_j \asymp j^{-\beta}$ with $\beta > 0$ hold. Then, the Delaigle–Hall condition is satisfied with $2\beta - \alpha \leq 1$. Because α describes the complexity of the covariance $C(t, t')$ and β indicates the smoothness of μ, the Delaigle–Hall condition is more likely to be satisfied when the functional data are less smooth and the covariance decays quickly.

3.2 Conditions

Here, we discuss several assumptions needed for the fast convergence. Recall that $\mathcal{N}(\varepsilon, \mathcal{X}, d)$ denotes a covering number of $\mathcal{X}$, which is common in empirical process theory and statistical learning theory (for an introduction, see Van Der Vaart and Wellner (1996)). We consider the following condition.

Assumption 1 (Covering Bound)**.** *There exists constants $\bar{\varepsilon} > 0$, $\gamma > 0$, and $V > 0$ such that for every $\varepsilon \in (0, \bar{\varepsilon})$, the following holds:*

$$\log \mathcal{N}(\varepsilon, \mathcal{X}, d) \leq V \varepsilon^{-\gamma}.$$

Meister (2016) uses this type of assumption for convergence analysis of functional data. Here, γ represents the complexity of the functional data space $\mathcal{X}$,

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

and controls the smoothness of $\mathcal{X}$ and a dimension of inputs for X.

Assumption 1 restricts the form of functional data in $\mathcal{X}$ in the following ways. First, it requires a kind of continuity or differentiability for the functional data. The degree of smoothness is adjusted by the decay rate γ in Assumption 1. Second, it also requires that a norm of the functional data be bounded. This constraint excludes Gaussian processes, and so we use a truncated version of Gaussian processes instead. The following examples satisfy Assumption 1. Additional examples are deferred to the Supplementary Material.

Example 2 (Smooth Path)**.** For $\alpha \in \mathbb{N}$, suppose that $\mathcal{X}$ is a set of functions f on $[0,1]^p$ that have partial derivatives up to an order $\alpha - 1$ that are uniformly bounded by some constant, and the highest partial derivatives are Lipschitz continuous. In this case, a setting $\gamma = p/\alpha$ satisfies Assumption 1 with $d = \|\cdot\|_{L^\infty}$ (Theorem 2.7.1 in Van Der Vaart and Wellner (1996)). □

Example 3 (Nonsmooth Path)**.** For $\alpha' \in (0,1]$, $\mathcal{X}$ is a set of α'-Hölder-continuous functions on $[0,1]^p$, which is a set of functions $f : [0,1]^p \to \mathbb{R}$, such that

$$|f(x) - f(x')| \leq C\|x - x'\|^{\alpha}$$

holds for every $x, x' \in [0,1]^p$, with some constant $C > 0$. In this case, a setting $\gamma = p/\alpha$ satisfies Assumption 1 with $d = \|\cdot\|_{L^\infty}$ (Theorem 2.7.1 in Van Der Vaart and Wellner (1996)). Note that this set includes nondifferentiable

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

functions. □

Example 4 (Unbounded Path with Finite Peaks)**.** We consider a family of functions f on $[0,1]$, such that

$$f(x) = g(x) + \sum_{j=1}^{J} \psi(x; a_j, t_j),$$

where g is a function from the Sobolev space with order $\alpha \in N$ (a space of α-times weakly differentiable functions in terms of $\|\cdot\|_{L^2}$), $\psi(x; a_j, t_j) = a_j/|(x - t_j)|^{1/3}$ is an unbounded peak function with scale parameters $a_j \in [0,1]$ and fixed locations $t_j \in [0,1]$, and J is a number of peaks. We can show that a set of such functions satisfies Assumption 1 with $\gamma = 1/\alpha$, and with $d = \|\cdot\|_{L^2}$ and sufficiently large $V > 0$; see Proposition **??** in the Supplementary Material. □

Next, we give the second condition for the distribution Π of X. For $x \in \mathcal{X}$ and $\delta > 0$, define $B(x; \delta)$ as an x-centered open ball with radius δ in terms of $\|\cdot\|$. Then, we impose the following assumption.

Assumption 2 (Positive Small Ball Probability)**.** *For any x in a support of Π and $\delta > 0$, $\Pi(B(x;\delta)) > 0$ holds.*

This assumption is satisfied for a general class of distributions, even in the functional data setting. We provide several examples in the Supplementary Material.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

3.3 Convergence Result

Now, we provide our main result on the convergence speed of the generalization error. We provide an outline of the proof and details of the constants outline in the next section, and the full proof in the Supplementary Material.

Theorem 1. *Let $\mathcal{H}$ be an RKHS on $\mathcal{X}$ with a universal kernel. Let $\widehat{f}_n \in \mathcal{H}$ be a classifier, minimizing the empirical loss, as defined in* (2.2)*. Suppose that the Delaigle–Hall condition and Assumptions 1 and 2 hold. Then, there exist positive and finite constants $C_{V,\gamma}$ and $C_{V,\Pi,\mathcal{H}}$, such that the following inequality holds for any $\lambda \in [\underline{\lambda}, \overline{\lambda}]$, with $\underline{\lambda} = \max\{(\log n)^{-1/\gamma}, C_{V,\gamma} \log\log n/n\}$ and $\overline{\lambda} = C_{V,\Pi,\mathcal{H}}$, and any $n \in \mathbb{N}$, as $\overline{\lambda} \geq \underline{\lambda}$:*

$$E\left[R(\widehat{f}_n) - \inf_{f \in \mathcal{H}} R(f)\right] \leq 2\exp(-\beta n),$$

where $\beta > 0$ is a parameter that depends on $\mathcal{H}, \Pi$, and V.

This result shows that very fast convergence of the generalization error is obtained under the Delaigle–Hall condition and a sufficient sample size. In other words, because this convergence is exponential in n, the error decays faster than all polynomial convergence in n, in contrast to the logarithmic convergence of Meister (2016), that is, $R(\widetilde{f}_n) - \inf_f R(f) \geq C(\log n)^{-1/\gamma}$ holds. This suggests that adding the Delaigle–Hall condition reduces the complexity of the functional classification problem more than expected.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

The following two technical points are important. First, the minimum required n is determined by γ, which reflects the complexity of the functional data in Assumption 1. When the functional data are more complex, that is, γ is large, the required sample size increases. Second, β depends on various parameters and is complicated to describe. Rigorous values are provided in the full proof.

Remark 1 (Role of the RKHS)**.** We use RKHSs for classifiers for the following reasons. First, the pointwise bound (2.1) in RKHSs is important for the error analysis. Second, an RKHS is closely related to the Delaigle–Hall condition (3.1), because the condition is regarded as measuring the difference between the means of the distributions in terms of an RKHS norm. This relation makes our error analysis simple.

Remark 2 (Selection of $\mathcal{H}$)**.** We discuss the effect of the choice of the RKHS. The exponential convergence in n, which is the main claim of Theorem 1, holds for all RKHSs, as long as the requirements are satisfied.

3.4 Proof Overview

The proof of Theorem 1 comprises three steps: (i) rewrite the Delaigle–Hall condition as a hard-margin condition; (ii) decompose the misclassification error; and (iii) study each of the components. Hereafter, we set L as a Lipschitz constant of f^* and assume that $\|f^*\|_{\mathcal{H}} \geq 1$, without loss of generality.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

Step (i): Rewrite the Delaigle–Hall condition as a hard-margin condition: We first introduce the hard-margin condition, which is a general condition in many classification problems.

Definition 2 (Hard-Margin Condition)**.** A margin of Π with $f : \mathcal{X} \to \mathbb{R}$ is defined as

$$\delta(f, \Pi) = \sup \{\delta : \Pi(\{x : |f(x)| < \delta\}) = 0\}.$$

We say Π satisfies the hard-margin condition with given f if $\delta(f, \Pi) > 0$ holds.

This condition requires that a discrepancy between the sets $\{x : f(x) > \delta\}$ and $\{x : f(x) < -\delta\}$ is large, almost surely. In other words, the margin with f is contained in a Π-null set. A margin is useful notion for handling the difficulty of classification problems. This condition is related to a common condition for other classification problems, namely, the strong noise condition (Koltchinskii and Beznosova, 2005; Audibert and Tsybakov, 2007).

To show the connection between the Delaigle–Hall condition and the hard-margin condition, we introduce f^* as follows. We define the sum of the orthogonal basis $\psi_M = \sum_{j=1}^{M} \theta_j^{-1}(\mu_{+,j} - \mu_{-,j})\phi_j$ and f_M^* as

$$f_M^*(x) = \left(\langle x - \mu_+, \psi_M \rangle\right)^2 - \left(\langle x - \mu_-, \psi_M \rangle\right)^2.$$

Furthermore, we define $f^* = \lim_{M \to \infty} f_M^*$. This function measures whether the input x is closer to μ_+ or μ_-, with the weight ψ_∞, and the sign of $f^*(x)$

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

works as a classifier. Then, the following result shows the equivalence of the Delaigle–Hall and hard-margin conditions:

Proposition 2 (Delaigle–Hall implies hard-margin)**.** *If the Delaigle–Hall conditions holds, then* $\delta(f^*, \Pi) > 0$ *holds.*

Proposition 2 shows that the Delaigle–Hall condition leads to the margin of Π being positive. This is similar to Theorem 5 in Berrendero *et al.* (2018), which states that the Delaigle–Hall condition is equivalent to the discrepancy between the supports of two measures P_+ and P_- under the Gaussian homoscedastic model. Proposition 2 also shows that f^* is an effective classifier with a sufficiently large margin under the Delaigle–Hall condition. The proof is based on an idea in Delaigle and Hall (2012), who apply a property that distances between functional data become infinitely large under the weighting by θ_j from the covariance.

Step (ii): Generalization Error Decomposition: In preparation, we convert the perfect classifier f^* into a controllable form. To this end, we define $\widetilde{f}_M(x) := f_M(x)/|f_M(x)|$ and $\widetilde{f}^*(x) = \lim_{M\to\infty} \widetilde{f}_M(x)$. Because the risk depends only on the sign of f^*, we have $R(\widetilde{f}^*) = R(f^*)$. In the following, we study the classification error based on $\widetilde{f}^*$, rather than f^*.

The first step is to rewrite the generalization error as an integral that involves probabilities associated with the signs of $\widetilde{f}^*$ and $\widehat{f}$. The standard calculation

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

yields the following transformation, by the Bochner integral:

$$E[R(\widehat{f}_n) - R(\widetilde{f}^*)] \leq \int |\eta(x)| \Pr(\widehat{f}_n(x)\widetilde{f}^*(x) \leq 0)\, d\Pi(x),$$

where $\eta(x) = E[Y|X = x]$. Next, for each x, we decompose the probability term $\Pr(\widehat{f}_n(x)\widetilde{f}^*(x) \leq 0)$. For x such that $\widetilde{f}^*(x) > 0$ holds, the misclassification error is rewritten as

$$\Pr(\widehat{f}_n(x)\widetilde{f}^*(x) \leq 0) = \Pr(\widehat{f}_n(x) \leq 0) \leq \underbrace{\Pr(\widehat{f}_n(x) \leq 0\,,\, \|\widehat{f}_n\|_{\mathcal{H}} \leq U)}_{=T_1} + \underbrace{\Pr(\|\widehat{f}_n\|_{\mathcal{H}} > U)}_{=T_2},$$

with a threshold value $U > 0$, specified in the full proof. We divide the event by the value of $\|\widehat{f}_n\|_{\mathcal{H}}$ associated with U, and then study each probability term separately.

Step (iii): Bound the Probability Terms: We bound T_1, using the hard-margin condition. Define $L_n(f) := n^{-1}\sum_{i=1}^n \ell(Y_i f(X_i)) + \lambda\|f\|_{\mathcal{H}}^2$. We show that $\widehat{f}_n$ cannot be a minimizer of $L_n(f)$ as (2.2) when T_1 is large under the hard-margin condition. Then, using the contradiction, we prove that T_1 converges exponentially in n. This part mainly follows the same proof in Koltchinskii and Beznosova (2005).

We bound T_2 using the empirical process technique. This part is specific to functional data, and hence some theory, such as Koltchinskii and Beznosova (2005) does not work. First, we show that bounding the excess loss $L_n(\widehat{f}_n) - L_n(\widetilde{f}^*)$ is sufficient to achieve the goal. To show the convergence of the excess

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

loss, we develop a covering number bound for $\mathcal{H}$ (Lemma **??** in the Supplementary Material), and develop the following bound with probability at least $1 - \exp(-t)$:

$$|L_n(f) - L(f)| \leq Rc_{V,\gamma}(\log n)^{-1/\gamma} + \sqrt{2t/n},$$

for any $f \in \mathcal{H}$ such that $\|f\|_{\mathcal{H}} \leq \|f^\dagger\|_{\mathcal{H}}$ holds, and any $t > 0$ (Lemma **??** in the Supplementary Material). Here, $c_{V,\gamma}$ is a constant depending on V and γ, which are specified in the full proof. As a result, we achieve our goal with a sufficiently large n.

4. Experiments

In this section, we conduct numerical experiments to support our theoretical result; that is, we analyze the change in the convergence rate of various classification methods for functional data under the Delaigle–Hall and hard-margin conditions.

4.1 Experimental Setting

For the functional classification problem, we consider the following settings. We generate functional data from two groups, with labels $\{-1, 1\}$. For each group, we generate n functions on $\mathcal{T} = [0, 1]$, with a northogonal basis $\phi_0(t) = 1$ and $\phi_j(t) = \sqrt{2}\sin(\pi j t), \forall j \geq 1$. Here, n is set from 1 to 3000. For a label $+1$,

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

we generate functional data $X_{i+}(t) = \sum_{j=0}^{50}(\theta_j^{1/2} Z_{j+} + \mu_{j+})\phi_j(t)$ with random variables Z_{j+} and coefficients θ_j, μ_{j+}, for $j = 0, 1, ..., 50$ and $i = 1, ..., n$. Similarly, for a label -1, we generate $X_{i-}(t) = \sum_{j=0}^{50}(\theta_j^{1/2} Z_{j-} + \mu_{j-})\phi_j(t)$ with random variables Z_{j-} and coefficients μ_{j-}.

We consider the following two scenarios. The values of the random variables and coefficients are determined separately. In *Scenario 1*, to consider perfect classifiable data using the Delaigle–Hall condition, we set $\theta_j = j^{-2}$, $\mu_{j-} = 0$, and $\mu_{j+} = j^{-\gamma}$, and draw Z_{j+}, Z_{j-} from a standard normal Gaussian distribution. Here, γ handles the complexity of the mean of functional data, and thus determines whether the data-generating process satisfies/violates the Delaigle–Hall condition. If $\gamma \leq 3/2$, the Delaigle–Hall condition is satisfied, and is violated otherwise. In *Scenario 2*, we examine perfect classification according to the hard-margin condition. We set $\theta_j = j^{-2}$, $\mu_{j-} = 0$, and $\mu_{j+} = \mathbf{1}\{j = 0\}\mu$, and let Z_{j+}, Z_{j-} be from a uniform distribution on $[-1/2, 1/2]$. Here, μ is a key parameter in terms of satisfying or violating the hard-margin condition. If $\mu \geq 1$ holds, the hard-margin conditions are satisfied, because the domains of P_+ and P_- do not overlap. Otherwise, the hard-margin condition is violated. For each method and n, we study its misclassification rate using 1000 newly generated data sets. We repeat each simulation experiment 200 times and report its mean. The case in which the basis functions differ between

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

labels is discussed in the Supplemental Material.

4.2 RKHS Classifier and the Delaigle–Hall/hard-margin Condition

Here, we examine the misclassification rate of the RKHS method in (2.2). We set the loss function as the logit loss $\ell(u) = \log(1 + \exp(-u))$, and construct the hypothesis space $\mathcal{H}$ using the functional RKHS associated with the Gaussian kernel $k(x, x') = \exp(-\|x-x'\|^2/h)$, with functions $x = x(t)$and $x' = x'(t)$ and a hyperparameter $h > 0$. The norm in the kernel is calculated as $\sum_{j=0}^{\infty}(\xi_j - \xi_j')^2$, where $\xi_j = \langle x, \phi_j \rangle$ and $\xi_j' = \langle x', \phi_j \rangle$. By the representer theorem (Theorem 5.5 in Steinwart and Christmann (2008)), the minimization problem is rewritten as

$$\min_{\{w_j\}_{j=1}^n} \frac{1}{n}\sum_{i=1}^{n} \ell\left(Y_i \sum_{j=1}^{n} w_j k(X_i, X_j)\right) + \lambda \sum_{j=1}^{n} w_j^2,$$

with the parameters $w_1, ..., w_n$. We solve the optimization problem by using the gradient descent method. The bandwidth h and the penalized parameter λ are determined using cross-validation (CV) from $\{2^{-5}, 2^{-4}, \ldots, 2^4\}$, minimizing the misclassification rate in the newly generated test data.

In Scenario 1, we consider configurations of the mean decay parameter as $\gamma \in \{1.6, 1.7\}$ to satisfy the Delaigle–Hall condition, or $\gamma \in \{1.3, 1.4\}$ to violate the condition. In Scenario 2, we consider $\mu \in \{0.8, 0.9\}$ to satisfy the hard-margin condition, or $\mu \in \{1.1, 1.2\}$ to violate it. For each scenario, we plot *error* (logarithm of misclassification error) against $\log n$ in Figure 1.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

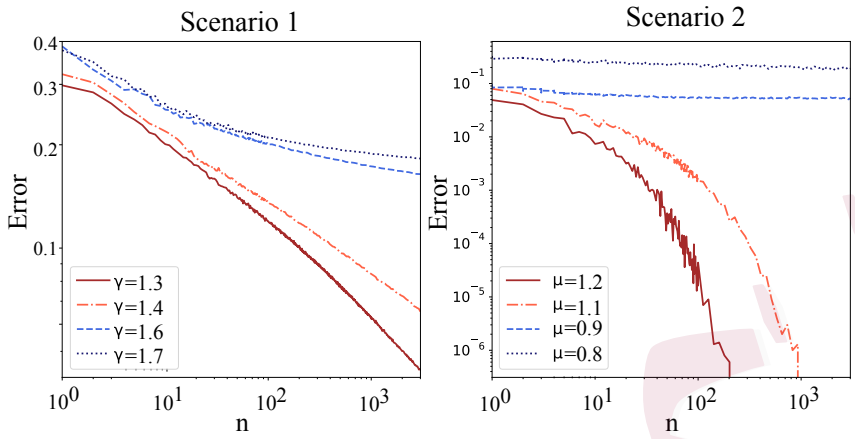

Figure 1: Error (logarithm of misclassification error rate) by the RKHS against $\log n$. Left: Scenario 1 for the Delaigle–Hall condition, with $\gamma \in \{1.3\,(\text{solid}), 1.4\,(\text{dashes}), 1.6\,(\text{dots}), 1.7\,(\text{dotdash})\}$. Right: Scenario 2 for the hard-margin condition, with $\mu \in \{1.2\,(\text{solid}), 1.1\,(\text{dashes}), 0.9\,(\text{dot}), 0.8\,(\text{dotdash})\}$.

Our results reveal the following findings. First, in Scenario 1 for the Delaigle–Hall condition, the error curves show slight differences in shape and slope. That is, the curves are convex when $\gamma = 1.6$ or 1.7 (the Delaigle–Hall condition is not satisfied), which appears to result in slow convergence. Second, in Scenario 2 for the hard-margin condition, the error curves show fast convergence only when $\mu = 1.2$ and 1.1 (the hard-margin condition is satisfied). These results show that the conditions affect the decay speed and errors, weakly for the Delaigle–Hall

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

condition, and drastically for the hard-margin condition.

Here, we investigate the effect of bandwidth selection on the results. Specifically, we consider Scenario 1, and set the bandwidth to 10, 50, and 100, repeat each simulation 200 times, and calculate the average classification error. The results are shown in Figure 2. As the bandwidth increases, the decay of the errors becomes more gradual. When the bandwidth is large, the perfect classification does not hold, because the expressive power of the kernel is reduced. Therefore, regardless of the value of γ, it becomes more difficult for the exponential decay of the errors to hold.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

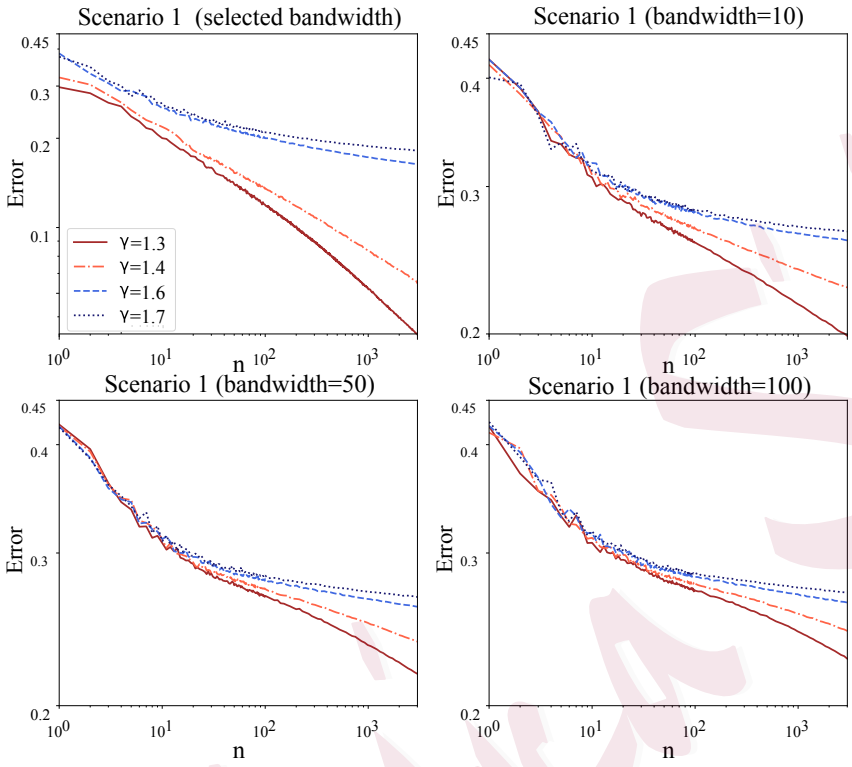

Figure 2: Error (logarithm of misclassification error rate) by the RKHS against $\log n$. Upper left: Scenario 1, with the bandwidth selected using CV. Other three: Scenario 1 for bandwidth $\lambda = 10, 50, 100$.

4.3 Others Methods and the Delaigle–Hall / hard-margin Condition

In this section, we compare the misclassification errors of several common classification methods for functional data. We consider the following classifiers: (a)

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

the kernel classifier (Dai *et al.*, 2017); (b) the centroid method (Delaigle and Hall, 2012); (c) the centroid method with a partial least square (PLS) (Preda *et al.*, 2007); (d) a logistic regression with a Gaussian process (GP); and (e) a linear discriminant analysis (LDA). The hyperparameters in (b) and (c) are chosen in the same way as in Delaigle and Hall (2012). The bandwidth of the kernel in (a) and the number of components for the dimension reduction in (e) are selected using CV. The hyperparameters in (d) are optimized using Algorithm 5.1 in KI Williams (2006). We set n from 5 to 1000. The remaining settings of the data-generating process and the RKHS method are the same as those of the previous sections.

The results are shown in Figure 3: the left column shows Scenario 1 with $\gamma = 1.3,\ 1.4,\ 1.6$, and 1.7, and the right shows Scenario 2 with $\mu = 1.2,\ 1.1,\ 0.9$, and 0.8. In Scenario 1 (left column), the parameter γ does not have a significant impact on the curves, although the RKHS method shows a slight difference in shape, as in the previous section. In Scenario 2 (right column), the parameter μ has a significant impact. As μ increases and the hard-margin condition is satisfied, the nonlinear methods (RKHS, GP, centroid, and kernel classifier) achieve fast convergence. In contrast, the linear methods (PLS and LDA) do not. This finding indicates that the nonlinear methods may achieve fast convergence with the hard-margin condition.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

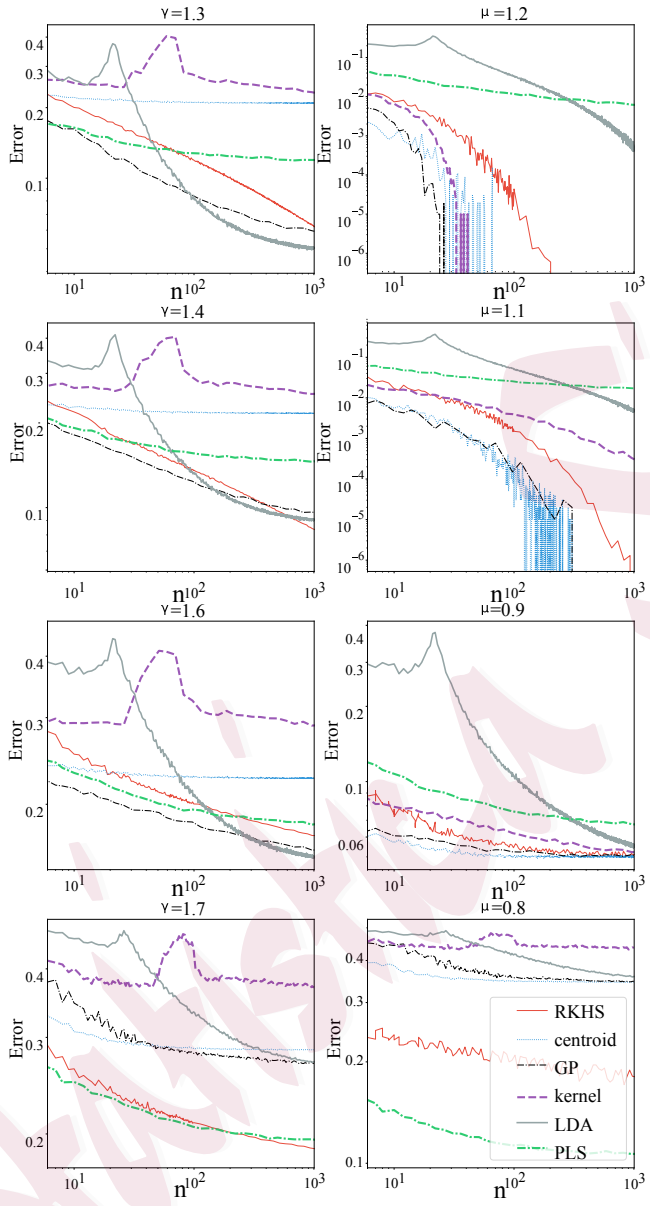

Figure 3: Error (logarithm of misclassification error) by the RKHS against $\log n$ of the RKHS method (solid), centroid method (dots), logistic regression with Gaussian process (dashdot), kernel classifier (bold dash), linear discriminant analysis (bold solid), and centroid method with partial least square (bold dashdot). Left column: Scenario 1 with $\gamma \in \{1.3, 1.4, 1.6, 1.7\}$. Right: Scenario 2 for the hard-margin condition with $\mu \in \{1.2, 1.1, 0.9, 0.8\}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

5. Conclusion

In this study, we investigate the convergence rate of the misclassification error of the classification problem for functional data, and discuss the feasibility of a small error with finite samples. The Delaigle–Hall condition guarantees the existence of a perfect classifier, which is a specific condition for functional data that cannot occur for finite-dimensional data. However, the minimax rate of the misclassification error with functional data, that is, the worst-case error, follows logarithmic convergence in the sample size. Hence, it is not clear whether we can achieve the perfect classification in practice with a realistic sample size. Our result reveals that the Delaigle–Hall condition leads not only to the existence of a perfect classifier, but also to the exponential convergence of the error. Furthermore, the Delaigle–Hall condition is helpful when estimating from finite samples. This reveals the specific advantage of treating functional data explicitly, because the Delaigle–Hall condition is specific to infinite-dimensional data.

Note that Assumption 1 on a covering number restricts the available class of functional data. This is unavoidable as long as we handle the properties of functional data in a uniform way using the notion of metric entropy. A possible way to avoid this is to use a spectral decomposition-based approach, as in Hall and Horowitz (2007), which deals directly with the randomness of functional data without entropy.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

The considered classifier is typical and also different from modern adaptive methods, such as neural networks. However, owing to this simplicity, we succeed in clarifying the theoretical properties with a perfect classification. Moreover, because analyses of adaptive methods are often extensions of analyses for simple methods, our results may serve as a basis for further research.

Supplementary Material

The Supplementary Material contains detailed proof of the main theorem and examples that satisfy conditions.

Acknowledgements

The research of Imaizumi was supported by JSPS KAKENHI (18K18114) and JST Presto (JPMJPR1852).

References

Alonso, A. M., Casado, D. and Romo, J. (2012) Supervised classification for functional data: A weighted distance approach, *Computational Statistics & Data Analysis*, **56**, 2334–2346.

Araki, Y., Konishi, S., Kawano, S. and Matsui, H. (2009) Functional logistic

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

discrimination via regularized basis expansions, *Communications in Statistics—Theory and Methods*, **38**, 2944–2957.

Audibert, J.-Y. and Tsybakov, A. B. (2007) Fast learning rates for plug-in classifiers, *The Annals of statistics*, **35**, 608–633.

Berlinet, A. and Thomas-Agnan, C. (2011) *Reproducing kernel Hilbert spaces in probability and statistics*, Springer Science & Business Media.

Berrendero, J. R., Cuevas, A. and Torrecilla, J. L. (2018) On the use of reproducing kernel hilbert spaces in functional classification, *Journal of the American Statistical Association*, **113**, 1210–1218.

Biau, G., Bunea, F. and Wegkamp, M. H. (2005) Functional classification in hilbert spaces, *IEEE Transactions on Information Theory*, **51**, 2163–2172.

Cai, T. T. and Yuan, M. (2012) Minimax and adaptive prediction for functional linear regression, *Journal of the American Statistical Association*, **107**, 1201–1216.

Cérou, F. and Guyader, A. (2006) Nearest neighbor classification in infinite dimension, *ESAIM: Probability and Statistics*, **10**, 340–355.

Chang, C., Chen, Y. and Ogden, R. T. (2014) Functional data classification: a wavelet approach, *Computational Statistics*, **29**, 1497–1513.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

Cuesta-Aboertos, J. A. and Dutta, S. (2016) On perfect classification for gaussian processes, *arXiv preprint arXiv:1602.04941*.

Cuevas, A. (2014) A partial overview of the theory of statistics with functional data, *Journal of Statistical Planning and Inference*, **147**, 1–23.

Cui, X., Lin, H. and Lian, H. (2020) Partially functional linear regression in reproducing kernel hilbert spaces, *Computational Statistics & Data Analysis*, **150**, 106978.

Da Prato, G. and Zabczyk, J. (2014) *Stochastic equations in infinite dimensions*, Cambridge university press.

Dai, X., Müller, H.-G. and Yao, F. (2017) Optimal bayes classifiers for functional data and density ratios, *Biometrika*, **104**, 545–560.

Delaigle, A. and Hall, P. (2012) Achieving near perfect classification for functional data, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **74**, 267–286.

Delaigle, A. and Hall, P. (2013) Classification using censored functional data, *Journal of the American Statistical Association*, **108**, 1269–1283.

Ferraty, F. and Vieu, P. (2003) Curves discrimination: a nonparametric functional approach, *Computational Statistics & Data Analysis*, **44**, 161–173.

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

Florindo, J. B., De Castro, M. and Bruno, O. M. (2011) Enhancing volumetric bouligand–minkowski fractal descriptors by using functional data analysis, *International Journal of Modern Physics C*, **22**, 929–952.

Gannaz, I. (2014) Classification of eeg recordings in auditory brain activity via a logistic functional linear regression model., in *International Workshop on Functional and Operatorial Statistics*, pp. p–125.

Hall, P. and Horowitz, J. L. (2007) Methodology and convergence rates for functional linear regression, *The Annals of Statistics*, **35**, 70–91.

Hanneke, S., Kontorovich, A., Sabato, S. and Weiss, R. (2021) Universal bayes consistency in metric spaces, *Annals of Statistics*, **49**, 2129–2150.

Hsing, T. and Eubank, R. (2015) *Theoretical foundations of functional data analysis, with an introduction to linear operators*, vol. 997, John Wiley & Sons.

Islam, M. N. (2020) Classification of pediatric pneumonia using chest x-rays by functional regression, *arXiv preprint arXiv:2005.03243*.

KI Williams, C. (2006) *Gaussian processes for machine learning*, Taylor & Francis Group.

Koltchinskii, V. and Beznosova, O. (2005) Exponential convergence rates in

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

classification, in *International Conference on Computational Learning Theory*, Springer, pp. 295–307.

Li, H., Xiao, G., Xia, T., Tang, Y. Y. and Li, L. (2013) Hyperspectral image classification using functional data analysis, *IEEE transactions on Cybernetics*, **44**, 1544–1555.

Lian, H. (2007) Nonlinear functional models for functional responses in reproducing kernel hilbert spaces, *Canadian Journal of Statistics*, **35**, 597–606.

Meister, A. (2016) Optimal classification and nonparametric regression for functional data, *Bernoulli*, **22**, 1729–1744.

Preda, C. (2007) Regression models for functional data by reproducing kernel hilbert spaces methods, *Journal of statistical planning and inference*, **137**, 829–840.

Preda, C. and Saporta, G. (2005) Clusterwise pls regression on a stochastic process, *Computational Statistics & Data Analysis*, **49**, 99–108.

Preda, C., Saporta, G. and Lévéder, C. (2007) Pls classification of functional data, *Computational Statistics*, **22**, 223–235.

Rincón, M. and Ruiz-Medina, M. D. (2012) Wavelet-rkhs-based functional sta-

Statistica Sinica: Preprint
doi:10.5705/ss.202022.0258

tistical classification, *Advances in Data Analysis and Classification*, **6**, 201–217.

Shalev-Shwartz, S. and Ben-David, S. (2014) *Understanding machine learning: From theory to algorithms*, Cambridge university press.

Steinwart, I. and Christmann, A. (2008) *Support vector machines*, Springer Science & Business Media.

Tian, Y., Lin, H., Lian, H. and Fan, Z. (2020) Additive functional regression in reproducing kernel hilbert spaces under smoothness condition, *Metrika*, pp. 1–14.

Van Der Vaart, A. W. and Wellner, J. A. (1996) *Weak convergence and empirical processes*, Springer.

Varughese, M. M., von Sachs, R., Stephanou, M. and Bassett, B. A. (2015) Non-parametric transient classification using adaptive wavelets, *Monthly Notices of the Royal Astronomical Society*, **453**, 2848–2861.

Wang, X. and Qu, A. (2014) Efficient classification for longitudinal data, *Computational Statistics & Data Analysis*, **78**, 119–134.

Yang, Y.-H., Chen, L.-H., Wang, C.-C. and Chen, C.-S. (2014) Bayesian fisher's discriminant for functional data, *arXiv preprint arXiv:1412.2929*.