Bernoulli **26**(3), 2020, 1797–1831
https://doi.org/10.3150/19-BEJ1170

Optimal functional supervised classification with separation condition

SÉBASTIEN GADAT[1], SÉBASTIEN GERCHINOVITZ[2] and CLÉMENT MARTEAU[3]

[1] *Toulouse School of Economics, Institut Universitaire de France. E-mail: sebastien.gadat@math.univ-toulouse.fr*

[2] *Institut Mathématiques de Toulouse & IRT Saint-Exupéry, 3 rue Tarfaya, 31405 Toulouse, France. E-mail: sebastien.gerchinovitz@math.univ-toulouse.fr*

[3] *Univ Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5208, Institut Camille Jordan F-69622 Villeurbanne, France. E-mail: marteau@math.univ-lyon1.fr*

We consider the binary supervised classification problem with the Gaussian functional model introduced in (*Math. Methods Statist.* **22** (2013) 213–225). Taking advantage of the Gaussian structure, we design a natural plug-in classifier and derive a family of upper bounds on its worst-case excess risk over Sobolev spaces. These bounds are parametrized by a separation distance quantifying the difficulty of the problem, and are proved to be optimal (up to logarithmic factors) through matching minimax lower bounds. Using the recent works of (In *Advances in Neural Information Processing Systems* (2014) 3437–3445 Curran Associates) and (*Ann. Statist.* **44** (2016) 982–1009), we also derive a logarithmic lower bound showing that the popular k-nearest neighbors classifier is far from optimality in this specific functional setting.

Keywords: functional data; supervised classification

1. Introduction

Motivation. The binary supervised classification problem is perhaps one of the most common tasks in statistics and machine learning. Even so, this problem still fosters new theoretical and applied questions because of the large variety of the data encountered so far. We refer the reader to [16] and [7] and to the references therein for a comprehensive introduction to binary supervised classification. This problem unfolds as follows. The learner has access to n independent copies $(X_1, Y_1), \dots, (X_n, Y_n)$ of a pair (X, Y), where X lies in a measurable space $\mathcal{H}$ and $Y \in \{0, 1\}$. The goal of the learner is to predict the label Y after observing the new input X, with the help of the sample $\mathcal{S}_n := (X_i, Y_i)_{1 \le i \le n}$ to learn the unknown joint distribution $\mathbb{P}_{X,Y}$ of the pair (X, Y).

In some standard situations, X lies in the simplest possible Hilbert space: $\mathcal{H} = \mathbb{R}^d$, which corresponds to the finite-dimensional binary classification problem. This setting has been extensively studied so far. Popular classification procedures that are now theoretically well understood include the ERM method [2,29], the k-nearest neighbors algorithm [4,12,18,19], support vector machines [35], or random forests [5], just to name a few.

Functional framework. However there are situations where the inputs X_i and X are better modelled as functions; the set $\mathcal{H}$ is then infinite-dimensional. Practical examples can be found, for example, in stochastic population dynamics [25], in signal processing [13], or in finance [24]. This

1350-7265 © 2020 ISI/BS

binary supervised functional classification problem has been at the core of several investigations. We mention, among others, [15,22,32] or [3]. This problem was also tackled with nonparametric procedures such as kernel methods or the k-nearest neighbours algorithm. For example, [14] studied the nearest neighbour rule in any metric space, while [23] analyzed the performances of the k-nearest neighbours algorithm in terms of a metric covering measure. Such metric entropy arguments were also used in [11], or with kernel methods in [1]. In [32], the authors develop an infinite-dimensional adaptation of the Support Vector Machine method to handle classification of functional signals.

1.1. Our functional model

In the present work, we focus on one of the most elementary diffusion classification models: we suppose that the input $X = (X(t))_{t\in[0,1]}$ is a continuous trajectory, solution to the stochastic differential equation

$$\forall t \in [0,1], \quad dX(t) = Yf(t)\,dt + (1-Y)g(t)\,dt + dW(t), \tag{1.1}$$

where $(W(t))_{0\le t\le 1}$ is a standard Brownian motion, Y is a Bernoulli $\mathcal{B}(1/2)$ random variable independent from $(W(t))_{0\le t\le 1}$, and were $f, g \in \mathbb{L}^2([0,1])$. In particular, in the sample $\mathcal{S}_n$, trajectories X_i labeled with $Y_i = 1$ correspond to observations of the signal f, while trajectories X_i labeled with $Y_i = 0$ correspond to g.

The white noise model has played a key role in statistical theoretical developments; see, for example, the seminal contributions of [20] in nonparametric estimation and of [27] in adaptive nonparametric estimation. In our supervised classification setting, the goal is not to estimate f and g but to predict the value of Y given an observed continuous trajectory $(X(t))_{0\le t\le 1}$. Of course, we assume that both functions f and g are unknown so that the joint distribution $\mathbb{P}_{X,Y}$ of the pair $((X(t))_{t\in[0,1]}, Y)$ is unknown. Without any assumption on f and g, there is no hope to solve this problem in general. However, learning the functions f and g (and thus $\mathbb{P}_{X,Y}$) from the sample $\mathcal{S}_n$ becomes statistically feasible when f and g are smooth enough.

The functional model considered in this paper is very close to the one studied by [9]. Actually our setting is less general since [9] considered more general diffusions driven by state-dependent drift terms $t \longmapsto f(t, X(t))$ and $t \longmapsto g(t, X(t))$. We focus on a simpler model, but derive refined risk bounds (with a different approach) that generalize the worst-case bounds of [9], as indicated below.

1.2. Link with the Gaussian sequence space model

Equation (1.1) is a condensed writing of a functional classification model that may be made more explicit with the help of an Hilbert basis. Recall that the Gaussian sequence space model is a statistical setting where we observe an infinite random sequence $Z = (Z_j)_{j\ge 1}$ with $Z_j = a_j + \varepsilon_j$, for i.i.d. $\mathcal{N}(0,1)$ random variables ε_j and coefficients a_j that are typically square-summable. Next, we explain how to reduce our model (1.1) to a set of two instances of the Gaussian sequence space model.

In the following, we introduce the functional space $\mathbb{L}^2([0,1])$ as the set of square Lebesgue-integrable functions f on $[0,1]$, with $\mathbb{L}^2$-norm $\|f\| = (\int_0^1 f^2(t)\,dt)^{1/2}$ and inner product $\langle f, g\rangle = \int_0^1 f(t)g(t)\,dt$. With a slight abuse of notation, when X is a solution of (1.1) and $\varphi \in \mathbb{L}^2([0,1])$, we set:

$$\begin{aligned}\langle \varphi, X\rangle &= \int_0^1 \varphi(t)\,dX(t)\\ &= Y\int_0^1 \varphi(t)f(t)\,dt + (1-Y)\int_0^1 \varphi(t)g(t)\,dt + \int_0^1 \varphi(t)\,dW(t).\end{aligned}$$

The above almost sure equality implies that the conditional distribution of $\langle \varphi, X\rangle$ given Y is Gaussian with expectation $\langle \varphi, f\rangle \mathbb{1}_{\{Y=1\}} + \langle \varphi, g\rangle \mathbb{1}_{\{Y=0\}}$ and variance $\|\varphi\|^2$. Therefore, for any $\varphi \in \mathbb{L}^2([0,1])$, the distribution of $\langle \varphi, X\rangle$ is a mixture of two Gaussian distributions:

$$\frac{1}{2}\mathcal{N}\bigl(\langle \varphi, f\rangle, \|\varphi\|^2\bigr) + \frac{1}{2}\mathcal{N}\bigl(\langle \varphi, g\rangle, \|\varphi\|^2\bigr).$$

We now consider $(\varphi_j)_{j\in\mathbb{N}^*}$ a given orthonormal basis of $\mathbb{L}^2([0,1])$. In the sequel, the coefficients $(c_j(h))_{j\geq 1}$ of any function $h \in \mathbb{L}^2([0,1])$ w.r.t. the basis $(\varphi_j)_{j\geq 1}$ are defined as

$$c_j(h) := \langle \varphi_j, h\rangle = \int_0^1 h(s)\varphi_j(s)\,ds, \quad j \geq 1,$$

and its $\mathbb{L}^2$-projection onto $\mathrm{Span}(\varphi_j, 1 \leq j \leq d)$ is given by

$$\Pi_d(h) = \sum_{j=1}^{d} c_j(h)\varphi_j. \tag{1.2}$$

In particular, we will pay a specific attention to the coefficients of f and g involved in (1.1),

$$\forall j \geq 1, \quad \theta_j := c_j(f) \quad \text{and} \quad \mu_j := c_j(g), \tag{1.3}$$

and to their d-dimensional projections $f_d := \Pi_d(f) = \sum_{j=1}^d \theta_j\varphi_j$ and $g_d := \Pi_d(g) = \sum_{j=1}^d \mu_j\varphi_j$.

An important feature of the white noise model is that the coefficients $(\langle \varphi_j, W\rangle)_{j\geq 1}$ associated with different frequencies of the standard Brownian motion are *independent*. This is because they are jointly Gaussian, with a diagonal infinite covariance matrix:

$$\forall j \neq j', \quad \mathbb{E}\bigl[\langle \varphi_j, W\rangle\langle \varphi_{j'}, W\rangle\bigr] = \int_0^1 \varphi_j(t)\varphi_{j'}(t)\,dt = 0.$$

The above remarks imply together with (1.3) that

$$\forall j \geq 1, \quad \mathcal{L}\bigl(\langle \varphi_j, X\rangle | Y = 1\bigr) = \mathcal{N}(\theta_j, 1) \quad \text{and} \quad \mathcal{L}\bigl(\langle \varphi_j, X\rangle | Y = 0\bigr) = \mathcal{N}(\mu_j, 1), \tag{1.4}$$

and that the coefficients $(\langle \varphi_j, X\rangle)_{j\geq 1}$ are conditionally independent given Y.

Therefore, the observation of a trajectory $X = (X(t))_{t\in[0,1]}$ given by (1.1) can be equivalently rewritten as the observation of one among two instances of the Gaussian sequence space model (with coefficients θ_j of μ_j), depending on the value of Y.

Remark 1.1 (Functional Linear Discriminant Analysis). Our setting and algorithm can be seen as an infinite-dimensional variant of Linear Discriminant Analysis (LDA). In LDA the goal is to predict the label of a random vector drawn from the mixture of two multivariate Gaussian distributions with the same covariance matrix. The Bayes classifier (known in this case as Fisher's linear discriminant rule) involves a linear function of the observed vector. In our functional setting, the above paragraphs indicate that the observed trajectory X is a mixture of two infinite-dimensional Gaussian distributions with the identity matrix as covariance matrix. The Bayes classifier also involves a linear function of the observed trajectory X, and so does our classifier (see (2.4) and Section 3.1 below). In fact, our classifier may be seen as a linear discriminant analysis method coupled with a projection step on a suitably chosen finite-dimensional space.

Several earlier works already studied functional analogues of Linear or Quadractic Discriminant Analysis. For instance, [6] studied how bad Fisher's rule behaves in high dimension, but also proved risk upper bounds in infinite dimension when the two signals f and g belong to a Sobolev-like compact set and are well separated (with a possibly unknown covariance matrix). [22] proposed a functional Quadratic Discriminant Analysis method with an additional estimation of the covariance of the processes. [15] obtains universal consistency (convergence to the Bayes risk in probability) of some centroid-method classification when the number of observations goes to $+\infty$ without any rate of convergence. In this paper we prove a family of excess risk bounds interpolating those of [9] and [6] by providing a sharp description of the role played by the distance $\Delta = \|f - g\|$, as well as matching lower bounds (up to log factors).

We refer the reader to Remark 3.2 for a more detailed comparison of our results with earlier works. Finally, we would also like to mention the general survey of [36] (see in particular Section 4.2), which gives some references not only on Fisher LDA but also on methods inspired by logistic regression with functional entries and by functional principal component analysis.

1.3. Main contributions and outline of the paper

We introduce some notation and definitions in order to present our contributions below. In our setting, a *classifier* Φ is a measurable function, possibly depending on the sample $\mathcal{S}_n$, that maps each new input $X = (X(t))_{t\in[0,1]}$ to a label in $\{0, 1\}$. The *risk* associated with each classifier Φ depends on f and g and is defined by:

$$\mathcal{R}_{f,g}(\Phi) := \mathbb{P}\big(\Phi(X) \neq Y\big),$$

where the expectation is taken with respect to all sources of randomness (i.e., both the sample $\mathcal{S}_n$ and the pair (X, Y)). The goal of the learner is to construct a classifier $\widehat{\Phi}$ based on the sample $\mathcal{S}_n$ that mimics the *Bayes classifier*

$$\Phi^\star = \arg\min_{\Phi} \mathcal{R}_{f,g}(\Phi), \tag{1.5}$$

where the infimum is taken over all possible classifiers (the oracle $\Phi^\star$ is impractical since f and g and thus $\mathbb{P}_{X,Y}$ are unknown). We measure the quality of $\widehat{\Phi}$ through its *worst-case excess risk*

$$\sup_{(f,g)\in\mathcal{E}}\left\{\mathcal{R}_{f,g}(\widehat{\Phi})-\inf_{\Phi}\mathcal{R}_{f,g}(\Phi)\right\} \tag{1.6}$$

over some set $\mathcal{E}$ of pairs of functions. In the sequel, we focus on Sobolev classes $\mathcal{H}_s(R)$ (see (3.9)) and consider subsets $\mathcal{E}\subseteq\mathcal{H}_s(R)^2$ parametrized by a separation lower bound Δ on $\|f-g\|$.

In Section 2, we first state preliminary results about the margin behavior that will prove crucial in our analysis. We then make three types of contributions:

- In Section 3, we design a classifier $\widehat{\Phi}_{d_n}$ based on a thresholding rule. It can be seen as a generalization of Linear Discriminant Analysis to the Gaussian sequence space model under smoothness assumptions. We derive an excess risk bound that generalizes both the worst-case results of [9] and the fast rates of [6], Theorem 2, when the distance $\|f-g\|$ is large. In particular we show that there is a continuum between all these rates, as a function of $\|f-g\|$. The acceleration is a consequence of the nice properties of the margin (see also, e.g., [2] and [18]).

Theorem (A). *The classifier $\widehat{\Phi}_{d_n}$ defined in (3.4) with $d_n\approx n^{\frac{1}{2s+1}}$ has an excess risk roughly bounded by (omitting logarithmic factors and constant factors depending only on s and R): for $n\geq N_{R,s}$ large enough,*

$$\sup_{\substack{f,g\in\mathcal{H}_s(R)\\ \|f-g\|\geq\Delta}}\left\{\mathcal{R}_{f,g}(\widehat{\Phi}_{d_n})-\inf_{\Phi}\mathcal{R}_{f,g}(\Phi)\right\}\lesssim\begin{cases}n^{-\frac{s}{2s+1}} & \text{if } \Delta\lesssim n^{-\frac{s}{2s+1}}\\ \dfrac{1}{\Delta}n^{-\frac{2s}{2s+1}} & \text{if } \Delta\gtrsim n^{-\frac{s}{2s+1}}\end{cases}$$

- In Section 4.1, we derive a matching minimax lower bound (up to logarithmic factors) showing that the above worst-case bound cannot be improved by any classifier.

Theorem (B). *For any number $n\geq N_{R,s}$ of observations, any classifier $\widehat{\Phi}$ must satisfy (omitting again logarithmic factors and constant factors depending only on s and R):*

$$\sup_{\substack{f,g\in\mathcal{H}_s(R)\\ \|f-g\|\geq\Delta}}\left\{\mathcal{R}_{f,g}(\widehat{\Phi})-\inf_{\Phi}\mathcal{R}_{f,g}(\Phi)\right\}\gtrsim\begin{cases}n^{-\frac{s}{2s+1}} & \text{if } \Delta\lesssim n^{-\frac{s}{2s+1}}\\ \dfrac{1}{\Delta}n^{-\frac{2s}{2s+1}} & \text{if } \Delta\gtrsim n^{-\frac{s}{2s+1}}\end{cases}$$

- Finally, in Section 4.2, we show that the well-known k-nearest neighbors rule tuned in a classical and optimal way (see, e.g., [18,33]) is far from optimality in our specific functional setting.

Theorem (D). *For any threshold (dimension) $\widehat{d}\in\mathbb{N}^*$ based on a sample-splitting policy, and for the optimal choice $k_n^{\mathrm{opt}}(\widehat{d})=\lfloor n^{4/(4+\widehat{d})}\rfloor$, the $\widehat{d}$-dimensional $k_n^{\mathrm{opt}}(\widehat{d})$-nearest neigh-*

bors classifier Φ_{NN} suffers a logarithmic excess risk in the worst case:

$$\sup_{f,g\in\mathcal{H}_s(r)}\Big\{\mathcal{R}_{f,g}(\Phi_{NN})-\inf_{\Phi}\mathcal{R}_{f,g}(\Phi)\Big\}\gtrsim\log(n)^{-2s}.$$

Most proofs are postponed to Appendix 4.2.3 (for the upper bounds) and to the supplementary material [17] (for the lower bounds).

Other useful notation (*main body and supplementary material*). We denote the joint distribution of the pair $((X(t))_{t\in[0,1]},Y)$ by $\mathbb{P}_{X,Y}$, and write $\mathbb{P}_{\otimes^n}=P_{X,Y}^{\otimes n}$ for the joint distribution of the sample $(X_i,Y_i)_{1\leq i\leq n}$. For notational convenience, the measure $\mathbb{P}$ will alternatively stand for $\mathbb{P}=\mathbb{P}_{X,Y}\otimes\mathbb{P}_{\otimes^n}$ (we integrate over both the sample $\mathcal{S}_n$ and the pair (X,Y)) or for any other measure made clear by the context. The distribution of $(X(t))_{t\in[0,1]}$ will be denoted by $\mathbb{P}_X$, while the distribution of $(X(t))_{t\in[0,1]}$ *conditionally* on the event $\{Y=1\}$ (resp. $\{Y=0\}$) will be written as $\mathbb{P}_f$ (resp. $\mathbb{P}_g$).

Finally, we write $\mathcal{B}(p)$ for the Bernoulli distribution of parameter $p\in[0,1]$, as well as $\mathcal{B}(n,p)$ for the binomial distribution with parameters $n\in\mathbb{N}^*$ and $p\in[0,1]$. We also set $x\wedge y=\min\{x,y\}$ for all $x,y\in\mathbb{R}$.

2. Preliminary results

2.1. Bayes classifier

We start by deriving an explicit expression for the optimal classifier $\Phi^\star$ introduced in (1.5). This optimal classifier is known as *the Bayes classifier* of the classification problem (see, e.g., [16, 19]).

Let $\mathbb{P}_0$ denote the Wiener measure on the set of continuous functions on $[0,1]$. It is easy to check that the law of $X|Y$ is absolutely continuous with respect to $\mathbb{P}_0$ (see, e.g., [21]). Indeed, for any continuous trajectory X, the Girsanov formula implies that the density of $\mathbb{P}_f$ (i.e., of $X|\{Y=1\}$) with respect to the reference measure $\mathbb{P}_0$ is given by

$$q_f(X):=\frac{d\mathbb{P}_f}{d\mathbb{P}_0}(X)=\exp\left(\int_0^1 f(s)\,dX(s)-\frac{1}{2}\|f\|^2\right). \tag{2.1}$$

Similarly, the density of $\mathbb{P}_g$ (i.e., of $X|\{Y=0\}$) with respect to $\mathbb{P}_0$ is

$$q_g(X):=\frac{d\mathbb{P}_g}{d\mathbb{P}_0}(X)=\exp\left(\int_0^1 g(s)\,dX(s)-\frac{1}{2}\|g\|^2\right). \tag{2.2}$$

In the sequel, we refer to q_f and q_g as the likelihood ratios of the models $\mathbb{P}_f$ and $\mathbb{P}_g$ versus $\mathbb{P}_0$. Now, using the Bayes formula, we can easily see that the regression function η associated with

(1.1) is given by

$$\begin{aligned}
\eta(X) &:= \mathbb{E}[Y|X] = \mathbb{P}(Y=1|X) \\
&= \frac{\frac{d\mathbb{P}_f}{d\mathbb{P}_0}(X)}{\frac{d\mathbb{P}_f}{d\mathbb{P}_0}(X) + \frac{d\mathbb{P}_g}{d\mathbb{P}_0}(X)} \\
&= \frac{\exp(\int_0^1 (f(s)-g(s))\,dX(s) - \frac{1}{2}\|f\|^2 + \frac{1}{2}\|g\|^2)}{1+\exp(\int_0^1 (f(s)-g(s))\,dX(s) - \frac{1}{2}\|f\|^2 + \frac{1}{2}\|g\|^2)}. \qquad (2.3)
\end{aligned}$$

As an example, if we assume that we observe $dX(t) = f(t)\,dt$ with $X(0)=0$, then $\eta(X) = \frac{\exp(\frac{1}{2}\|f-g\|^2)}{1+\exp(\frac{1}{2}\|f-g\|^2)}$, which is larger than or equal to $1/2$ and gets closer to 1 when $\|f-g\|$ increases. Roughly speaking, this means in that example that the distribution $\mathbb{P}_f$ is more likely than the distribution $\mathbb{P}_g$, which is consistent with the definition of the model given by (1.1).

The Bayes classifier $\Phi^\star$ of the classification problem is then given by

$$\Phi^\star(X) := \mathbb{1}_{\{\eta(X)\geq \frac{1}{2}\}} = \mathbb{1}_{\{\int_0^1 (f(s)-g(s))\,dX(s) \geq \frac{1}{2}\|f\|^2 - \frac{1}{2}\|g\|^2\}}. \qquad (2.4)$$

It is well known that the Bayes classifier $\Phi^\star$ corresponds to the optimal classifier of the considered binary classification problem (see, e.g., [16]) in the sense that it satisfies (1.5). In particular, for any other classifier Φ, the excess risk of classification is given by

$$\mathcal{R}_{f,g}(\Phi) - \mathcal{R}_{f,g}(\Phi^\star) = \mathbb{E}\big[|2\eta(X)-1|\mathbb{1}_{\{\Phi(X)\neq\Phi^\star(X)\}}\big]. \qquad (2.5)$$

We refer for instance to [4] for the proof of (2.5). In our statistical setting, the functions f and g are unknown so that it is impossible to compute the oracle Bayes classifier (2.4). However, we can construct an approximation of it using the sample $(X_i, Y_i)_{1\leq i\leq n}$. In Section 3, we design a plug-in estimator combined with a projection step, and analyze its excess risk under a smoothness assumption on f and g. The next result on the margin (Proposition 1 below) will be a key ingredient of our analysis.

Remark 2.1. In this paper, we concentrate our attention on the model (1.1) in the specific situation where $Y \sim \mathcal{B}(1/2)$. However we might have investigated the case where $Y \sim \mathcal{B}(p)$ for some $p \in]0,1[$. In such a situation, the regression function has the following expression:

$$\eta(X) = \frac{p\exp(\int_0^1 (f(s)-g(s))\,dX(s) - \frac{1}{2}\|f\|^2 + \frac{1}{2}\|g\|^2)}{p\exp(\int_0^1 (f(s)-g(s))\,dX(s) - \frac{1}{2}\|f\|^2 + \frac{1}{2}\|g\|^2) + 1 - p}.$$

All the results displayed below as, e.g., margin, rates of convergence and so on, can be generalized to this setting in a similar way. This generalization requires a huge amount of technical notation and does not change the spirit of the presented results, unless p is allowed to converge toward 0 or 1 as $n \to +\infty$. This last setting is far beyond the scope of this paper.

2.2. Control of the margin in the functional model

As was shown in earlier works on binary supervised classification (see, e.g., [29] or [2]), the probability mass of the region where the regression function η is close to $1/2$ plays an important role in the convergence rates. The behaviour of the function η is classically described by a so-called *margin assumption*: there exist $\alpha \geq 0$ and $\varepsilon_0, C > 0$ such that, for all $0 < \varepsilon \leq \varepsilon_0$,

$$\mathbb{P}_X\left(\left|\eta(X) - \frac{1}{2}\right| \leq \varepsilon\right) \leq C\varepsilon^{\alpha}. \tag{2.6}$$

We will show in Proposition 1 which parameters $\alpha, \varepsilon_0, C > 0$ are associated with Model (1.1). The role of (2.6) is easy to understand: classifying a trajectory X for which $\eta(X)$ is close to $1/2$ is necessarily a challenging problem because the events $\{Y = 1\}$ and $\{Y = 0\}$ are almost equally likely. This not only makes the optimal (Bayes) classifier $\mathbb{1}_{\{\eta(X)\geq 1/2\}}$ error-prone, but it also makes the task of mimicking the Bayes classifier difficult. Indeed, any slightly bad approximation of η when $\eta(X) \simeq 1/2$ can easily lead to a prediction different from $\mathbb{1}_{\{\eta(X)\geq 1/2\}}$. A large value of the margin parameter α indicates that most trajectories X are such that $\eta(X)$ is far from $1/2$: this makes in a sense the classification problem easier.

Our first contribution, detailed in Proposition 1 below, entails that the margin parameter associated with Model (1.1) crucially depends on the distance between the functions f and g of interest. The proof is postponed to Appendix A.2.

Proposition 1. *Let X be distributed according to Model* (1.1), *and set* $\Delta := \|f - g\|$. *Then, for all* $0 < \varepsilon \leq 1/8$, *we have*

$$\mathbb{P}_X\left(\left|\eta(X) - \frac{1}{2}\right| \leq \varepsilon\right) \leq 1 \wedge \frac{10\varepsilon}{\Delta}.$$

In particular, if the distance $\|f - g\|$ is bounded from below by a positive constant, then (2.6) is satisfied with a margin parameter $\alpha = 1$. If, instead, $\|f - g\|$ is allowed to be arbitrarily small, then nothing can be guaranteed about the margin parameter (except the obvious value $\alpha = 0$ that always works).

3. Upper bounds on the excess risk

In this section, we construct a classifier with nearly optimal excess risk. We detail its construction in Section 3.1 and analyze its approximation and estimation errors in Sections 3.2 and 3.3. Our main result, Theorem 3.1, is stated in Section 3.4. Nearly matching lower bounds will be provided in Section 4.

3.1. A classifier in a finite-dimensional setting

Our classifier—defined in Section 3.1.2 below—involves a projection step with coefficients θ_j and μ_j introduced in Section 1.2 that are estimated in Section 3.1.1.

3.1.1. *Estimation of* $(\theta_j)_{1\le j\le d}$ *and* $(\mu_j)_{1\le j\le d}$

In order to estimate the θ_j and μ_j, we split the sample $(X_i)_{1\le i\le n}$ into two subsamples $(X_i^0)_{1\le i\le N_0}$ and $(X_i^1)_{1\le i\le N_1}$ corresponding to either $Y_i = 0$ or $Y_i = 1$, where

$$N_0 := \sum_{i=1}^{n} \mathbb{1}_{\{Y_i=0\}} \quad \text{and} \quad N_1 := \sum_{i=1}^{n} \mathbb{1}_{\{Y_i=1\}}. \tag{3.1}$$

The sizes N_0 and N_1 are random variables; they satisfy $N_0 + N_1 = n$ and both have a binomial distribution $\mathcal{B}(n, 1/2)$. In particular, the two subsamples have (with high probability) approximately the same sizes.

Note from (1.1) that the two subsamples $(X_i^0)_{1\le i\le N_0}$ and $(X_i^1)_{1\le i\le N_1}$ correspond to observations of the functions g and f respectively. Following our comments from Section 1.2, it is natural to define the random coefficients $(X_{i,j}^0)_{\substack{1\le i\le N_0\\ 1\le j\le d}}$ and $(X_{i,j}^1)_{\substack{1\le i\le N_1\\ 1\le j\le d}}$ by

$$\begin{cases} X_{i,j}^0 := \langle \varphi_j, X_i^0\rangle, & i = 1,\dots,N_0,\\ X_{i,j}^1 := \langle \varphi_j, X_i^1\rangle, & i = 1,\dots,N_1, \end{cases} \qquad j \in \{1,\dots,d\}, \tag{3.2}$$

where the dimension $d \in \mathbb{N}^*$ will be determined later (as a function of n).

We provide a more formal definition of the above quantities in Appendix A.1. As can be seen from (A.2) and Remark A.1 therein, conditionally on $(Y_1,\dots,Y_n)$, the random coefficients $X_{i,j}^0$, $1 \le i \le N_0$, are i.i.d. $\mathcal{N}(\mu_j, 1)$, while the coefficients $X_{i,j}^1$, $1 \le i \le N_1$, are i.i.d. $\mathcal{N}(\theta_j, 1)$. It is therefore natural to estimate the coefficients μ_j and θ_j for every $j \in \{1,\dots,d\}$ by

$$\widehat{\mu}_j = \mathbb{1}_{\{N_0>0\}} \frac{1}{N_0} \sum_{i=1}^{N_0} X_{i,j}^0 \quad \text{and} \quad \widehat{\theta}_j = \mathbb{1}_{\{N_1>0\}} \frac{1}{N_1} \sum_{i=1}^{N_1} X_{i,j}^1. \tag{3.3}$$

Note that we arbitrarily impose the value 0 for $\widehat{\mu}_j$ when $N_0 = 0$ or for $\widehat{\theta}_j$ when $N_1 = 0$. This convention has a negligible impact, since with high probability N_0 and N_1 are both positive.

3.1.2. *A simple classifier*

We now build a simple classifier using the estimators $\widehat{\mu}_j$ and $\widehat{\theta}_j$ defined in (3.3). After observing a new trajectory $X = (X(t))_{0\le t\le 1}$, we construct the vector $\mathbf{X}_d \in \mathbb{R}^d$ defined by

$$\mathbf{X}_d := \big(\langle\varphi_1, X\rangle, \dots, \langle\varphi_d, X\rangle\big).$$

Then, we assign the label 1 to the trajectory X if $\mathbf{X}_d$ is closer to $\widehat{\theta} := (\widehat{\theta}_1,\dots,\widehat{\theta}_d)$ than to $\widehat{\mu} := (\widehat{\mu}_1,\dots,\widehat{\mu}_d)$, and the label 0 otherwise. More formally, our classifier $\widehat{\Phi}_d$ is defined for all trajectories X by

$$\widehat{\Phi}_d(X) = \begin{cases} 1 & \text{if } \|\mathbf{X}_d - \widehat{\theta}\|_d \le \|\mathbf{X}_d - \widehat{\mu}\|_d,\\ 0 & \text{if } \|\mathbf{X}_d - \widehat{\theta}\|_d > \|\mathbf{X}_d - \widehat{\mu}\|_d, \end{cases} \tag{3.4}$$

where $\|x\|_d = \sqrt{\sum_{j=1}^d x_j^2}$ denotes the Euclidean norm in $\mathbb{R}^d$; we also write $\langle x, y\rangle_d = \sum_{j=1}^d x_j y_j$ for the associated inner product.

Plug-in classifier. It shall be noticed that $\widehat{\Phi}_d$ is a plug-in classifier in a truncated space. Starting from (2.3) for the regression function η, we consider the 'truncated' regression function η_d by replacing f and g with their projections $\Pi_d(f) = \sum_{j=1}^d \theta_j \varphi_j$ and $\Pi_d(g) = \sum_{j=1}^d \mu_j \varphi_j$, that is,

$$\begin{aligned}\eta_d(X) &:= \frac{\exp(\int_0^1 (\Pi_d(f)(s) - \Pi_d(g)(s))\, dX(s) - \frac{1}{2}\|\Pi_d(f)\|^2 + \frac{1}{2}\|\Pi_d(g)\|^2)}{1 + \exp(\int_0^1 (\Pi_d(f)(s) - \Pi_d(g)(s))\, dX(s) - \frac{1}{2}\|\Pi_d(f)\|^2 + \frac{1}{2}\|\Pi_d(g)\|^2)} \\ &= \frac{\exp(\langle \boldsymbol{\theta}_d - \boldsymbol{\mu}_d, \mathbf{X}_d\rangle_d - \frac{1}{2}\|\boldsymbol{\theta}_d\|_d^2 + \frac{1}{2}\|\boldsymbol{\mu}_d\|_d^2)}{1 + \exp(\langle \boldsymbol{\theta}_d - \boldsymbol{\mu}_d, \mathbf{X}_d\rangle_d - \frac{1}{2}\|\boldsymbol{\theta}_d\|_d^2 + \frac{1}{2}\|\boldsymbol{\mu}_d\|_d^2)},\end{aligned} \tag{3.5}$$

where $\boldsymbol{\theta}_d := (\theta_j)_{1\le j\le d}$, and $\boldsymbol{\mu}_d := (\mu_j)_{1\le j\le d}$. We also define the associated oracle classifier

$$\Phi_d^\star(X) := \mathbb{1}_{\{\eta_d(X)\ge \frac{1}{2}\}} = \mathbb{1}_{\{\int_0^1 (\Pi_d(f)(s) - \Pi_d(g)(s))\, dX(s) \ge \frac{1}{2}\|\Pi_d(f)\|^2 - \frac{1}{2}\|\Pi_d(g)\|^2\}}. \tag{3.6}$$

As shown in Remark 3.1, η_d and $\Phi_d^\star$ correspond to the regression function and the Bayes classifier where the learner has only access to the projected input $\mathbf{X}_d \in \mathbb{R}^d$, rather than the whole trajectory X.

Now, following classical arguments, we are ready to reinterpret $\widehat{\Phi}_d$ as a plug-in classifier. Note that

$$\begin{aligned}\|\mathbf{X}_d - \widehat{\theta}\|_d \le \|\mathbf{X}_d - \widehat{\mu}\|_d &\Longleftrightarrow \frac{\exp(\frac{1}{2}\{\|\mathbf{X}_d - \widehat{\mu}\|_d^2 - \|\mathbf{X}_d - \widehat{\theta}\|_d^2\})}{1 + \exp(\frac{1}{2}\{\|\mathbf{X}_d - \widehat{\mu}\|_d^2 - \|\mathbf{X}_d - \widehat{\theta}\|_d^2\})} \ge \frac{1}{2} \\ &\Longleftrightarrow \widehat{\eta}_d(X) \ge \frac{1}{2},\end{aligned}$$

where the estimated regression function $\widehat{\eta}_d$ is defined by

$$\widehat{\eta}_d(X) := \frac{\exp(\langle \widehat{\theta} - \widehat{\mu}, \mathbf{X}_d\rangle_d - \frac{1}{2}\|\widehat{\theta}\|_d^2 + \frac{1}{2}\|\widehat{\mu}\|_d^2)}{1 + \exp(\langle \widehat{\theta} - \widehat{\mu}, \mathbf{X}_d\rangle_d - \frac{1}{2}\|\widehat{\theta}\|_d^2 + \frac{1}{2}\|\widehat{\mu}\|_d^2)}. \tag{3.7}$$

In other words, we can write $\widehat{\Phi}_d(X) = \mathbb{1}_{\{\widehat{\eta}_d(X)\ge 1/2\}}$ with $\widehat{\eta}_d$ the truncated estimation introduced in (3.5).

Proof strategy. In the next sections, we upper bound the excess risk of $\widehat{\Phi}_d$. We use the following classical decomposition (all quantities below are defined in Section 1 and Equations (3.4) and (3.6)):

$$\mathcal{R}_{f,g}(\widehat{\Phi}_d) - \mathcal{R}_{f,g}(\Phi^\star) = \underbrace{\mathcal{R}_{f,g}(\widehat{\Phi}_d) - \mathcal{R}_{f,g}(\Phi_d^\star)}_{\text{estimation error}} + \underbrace{\mathcal{R}_{f,g}(\Phi_d^\star) - \mathcal{R}_{f,g}(\Phi^\star)}_{\text{approximation error}}.$$

The first term of the right-hand side (estimation error) measures how close $\widehat{\Phi}_d$ is to the oracle $\Phi_d^\star$ in the truncated space; we analyse it in Section 3.3 below. The second term (approximation error) quantifies the statistical loss induced by the d-dimensional projection; we study it in Section 3.2.

3.2. Approximation error

We first upper bound the approximation error $\mathcal{R}_{f,g}(\Phi_d^\star) - \mathcal{R}_{f,g}(\Phi^\star)$, where the two oracle classifiers $\Phi_d^\star$ and $\Phi^\star$ are defined by (3.6) and (2.4) respectively. Comparing the definitions of η and η_d in (2.3) and (3.5), we can expect that, for d large enough, $\Pi_d(f) \approx f$ and $\Pi_d(g) \approx g$, so that $\eta_d(X) \approx \eta(X)$ and therefore $\mathcal{R}_{f,g}(\Phi_d^\star) \approx \mathcal{R}_{f,g}(\Phi^\star)$.

Lemma 1 below quantifies this approximation. The proof is postponed to Appendix A.3. We recall that, for notational convenience, we write $f_d = \Pi_d(f)$ and $g_d = \Pi_d(g)$.

Lemma 1. *Let X be distributed according to Model* (1.1), *and recall that* $\Delta := \|f - g\|$. *Let* $0 < \varepsilon \leq 1/8$ *and* $d \in \mathbb{N}^\star$ *such that*

$$\max\left(\|f - f_d\|^2, \|g - g_d\|^2\right) \leq \frac{\varepsilon^2}{512 \ln(1/\varepsilon^2)}. \tag{3.8}$$

Then, the two oracle classifiers $\Phi_d^\star$ *and* $\Phi^\star$ *defined by* (3.6) *and* (2.4) *satisfy*

$$\mathcal{R}_{f,g}\left(\Phi_d^\star\right) - \mathcal{R}_{f,g}\left(\Phi^\star\right) \leq 12\varepsilon^2 + 2\varepsilon\left(1 \wedge \frac{10\varepsilon}{\Delta}\right).$$

We stress that the distance Δ between f and g has a strong influence on the approximation error. In particular, if Δ is bounded from below independently from n, then the approximation error is at most of the order of ε^2, while it can only be controlled by ε if $\Delta \lesssim \varepsilon$. This key role of Δ is a consequence of the margin behavior analyzed in Proposition 1 (Section 2.2) and will also appear in the estimation error.

A smoothness assumption. When $d \in \mathbb{N}^*$ is fixed, we can minimize the bound of Lemma 1 in ε. Unsurprisingly the resulting bound involves the distances $\|f - f_d\|$ and $\|g - g_d\|$ of f and g to their projections f_d and g_d. In the sequel, we assume that the functions f and g belong to $\mathbb{L}^2([0,1])$ and are smooth in the sense that their (Fourier) coefficients w.r.t. the basis $(\varphi_j)_{j\geq 1}$ decay sufficiently fast. More precisely, we assume that, for some parameters $s, R > 0$, the functions f and g belong to the set

$$\mathcal{H}_s(R) := \left\{ h \in \mathbb{L}^2\left([0,1]\right) : \sum_{j=1}^{+\infty} c_j(h)^2 j^{2s} \leq R^2 \right\}. \tag{3.9}$$

The set $\mathcal{H}_s(R)$ corresponds to a class of smooth functions with smoothness parameter s: when $s = 0$, we simply obtain the $\mathbb{L}^2([0,1])$-ball of radius R. For larger s, for example $s = 1$, we obtain a smaller Sobolev space of functions such that $f' \in \mathbb{L}^2([0,1])$ with $\|f'\|_2 \leq R$.

Under the above assumption on the tail of the spectrum of f and g, the loss of accuracy induced by the projection step is easy to quantify. Indeed, for all $f \in \mathcal{H}_s(R)$ we have

$$\|f - f_d\|^2 = \sum_{j=d+1}^{+\infty} c_j(f)^2 \leq d^{-2s} \sum_{j=d+1}^{+\infty} c_j(f)^2 j^{2s} \leq R^2 d^{-2s},$$

so that, omitting logarithmic factors, ε can be chosen of the order of Rd^{-s} in the statement of Lemma 1.

3.3. Estimation error

We now upper bound the estimation error $\mathcal{R}_{f,g}(\widehat{\Phi}_d) - \mathcal{R}_{f,g}(\Phi_d^\star)$ of our classifier $\widehat{\Phi}_d$. To that end, we first reinterpret η_d and $\Phi_d^\star$; this will be useful to rewrite the estimation error as an excess risk (as in (2.5)) in the truncated space. The next remark follows from direct calculations.

Remark 3.1. Denote by $\mathbf{X}_d := (\langle \varphi_j, X\rangle)_{1\leq j\leq d}$, $\boldsymbol{\theta}_d = (\theta_j)_{1\leq j\leq d}$, and $\boldsymbol{\mu}_d = (\mu_j)_{1\leq j\leq d}$ the versions of X, θ, and μ in the truncated space. Then,

$$\eta_d(X) = \frac{\frac{1}{2} q_{f_d}(X)}{\frac{1}{2} q_{f_d}(X) + \frac{1}{2} q_{g_d}(X)} \quad \text{and} \quad q_{f_d}(X) = e^{\frac{1}{2}\|\mathbf{X}_d\|^2} e^{-\frac{1}{2}\|\mathbf{X}_d - \boldsymbol{\theta}_d\|^2}.$$

Since the conditional distribution of $\mathbf{X}_d$ is $\mathcal{N}(\boldsymbol{\theta}_d, I_d)$ given $Y = 1$ and $\mathcal{N}(\boldsymbol{\mu}_d, I_d)$ given $Y = 0$, this entails that $\eta_d(X) = \mathbb{P}(Y = 1|\mathbf{X}_d)$ almost surely.

In other words, η_d is the regression function of the restricted classification problem where the learner has only access to the projected trajectory $\mathbf{X}_d \in \mathbb{R}^d$, instead of the whole trajectory X. The function $\Phi_d^* = \mathbb{1}_{\{\eta_d \geq 1/2\}}$ is the associated Bayes classifier.

We are now ready to compare the risk of our classifier $\widehat{\Phi}_d$ to that of the d-dimensional oracle $\Phi_d^\star$. The proof of the next lemma is postponed to Appendix A.4. (The value of 4608 could most probably be improved.) We recall that $f_d = \Pi_d(f)$ and $g_d = \Pi_d(g)$.

Lemma 2. *We consider Model* (1.1). *Let $d \in \mathbb{N}^*$ and set $\Delta_d := \|f_d - g_d\|$. Let $0 < \varepsilon \leq 1/8$ and $n \geq 27$ such that*

$$\left(\Delta_d + 2\sqrt{\frac{d\log(n)}{n}}\right)\sqrt{\frac{d\log(n)}{n}} \leq \frac{\varepsilon}{48}. \tag{3.10}$$

Then, the classifiers $\widehat{\Phi}_d$ and $\Phi_d^\star$ defined by (3.4) *and* (3.6) *satisfy*

$$\mathcal{R}_{f,g}(\widehat{\Phi}_d) - \mathcal{R}_{f,g}(\Phi_d^\star) \leq 2\varepsilon\left(1 \wedge \frac{10\varepsilon}{\Delta_d}\right) + 6\exp\left(-\frac{n\varepsilon^2}{4608 d\log n}\right) + \frac{13}{n}.$$

In the same vein as for the approximation error, the estimation error bound above strongly depends on the distance between the two functions f_d and g_d of interest. This is again a consequence of the margin behavior analyzed in Proposition 1 (Section 2.2).

More precisely, when ε is chosen at least of the order of $\sqrt{d/n}\log(n)$ (in order to kill the exponential term), the estimation error bound above is roughly of the order of $\min\{\varepsilon, \varepsilon^2/\Delta_d\}$. In particular, if Δ_d is bounded from below, then the estimation error is at most of the order of $\varepsilon^2 \approx d\log^2(n)/n$. On the other hand, if no lower bound is available for Δ_d, then the only estimation error bound we get is a slower rate of the order of $\varepsilon \approx \sqrt{d/n}\log(n)$.

3.4. Convergence rate under a smoothness assumption

We now state the main result of this paper. We upper bound the excess risk $\mathcal{R}_{f,g}(\widehat{\Phi}_d) - \mathcal{R}_{f,g}(\Phi^\star)$ of our classifier when f and g belong to subsets of the Sobolev ball $\mathcal{H}_s(R)$ defined in (3.9). These subsets are parametrized by a separation distance Δ: a larger value of Δ makes the classification problem easier, as reflected by the non-increasing bound below.

Theorem 3.1. *There exist an absolute constant $c > 0$ and a constant $N_{s,R} \geq 86$ depending only on s and R such that the following holds true. For all $s, R > 0$ and all $n \geq N_{s,R}$, the classifier $\widehat{\Phi}_{d_n}$ defined by (3.4) with $d_n = \lfloor (R^2 n)^{\frac{1}{2s+1}} \rfloor$ satisfies*

$$\sup_{\substack{f,g\in\mathcal{H}_s(R)\\ \|f-g\|\geq\Delta}} \left\{ \mathcal{R}_{f,g}(\widehat{\Phi}_{d_n}) - \inf_{\Phi} \mathcal{R}_{f,g}(\Phi) \right\}$$

$$\leq \begin{cases} cR^{\frac{1}{2s+1}} n^{-\frac{s}{2s+1}} \log(n) & \text{if } \Delta < R^{\frac{1}{2s+1}} n^{-\frac{s}{2s+1}} \log(n) \\ \dfrac{c}{\Delta} R^{\frac{2}{2s+1}} n^{-\frac{2s}{2s+1}} \log^2(n) & \text{if } \Delta \geq R^{\frac{1}{2s+1}} n^{-\frac{s}{2s+1}} \log(n) \end{cases}$$

The proof is postponed to Appendix A.5 and combines Lemmas 1 and 2 from the previous sections.

Note that the two bounds of the right-hand side coincide when $\Delta = R^{1/(2s+1)} n^{-s/(2s+1)} \log(n)$. Therefore, there is a continuous transition from a slow rate (when Δ is small) to a fast rate (when Δ is large). This leads to the following remark.

Remark 3.2 (Novelty of the bound).

- Taking $\Delta = 0$, we recover the worst-case bound of [9], Corollary 4.4, (where $u = 1/s$) up to logarithmic factors. As shown by Theorem 4.1 below, this slow rate is unimprovable for a small distance $\|f - g\|$.
- In the much easier regime when $\|f - g\|$ is bounded from below, we recover the faster rate of [6], Theorem 2. This improved rate is a consequence of the margin behavior (see, e.g., [2,18]), but not of the choice of d_n that is oblivious to $\|f - g\|$. Our bound shows there is a continuum between these slow and fast rates, as a function of $\|f - g\|$.
- Continuous transitions from slow rates to faster rates were already derived in the past. For instance, for any supervised classification problem where the margin $|2\eta(X) - 1| \geq h$ is almost surely bounded from below, [30], Corollary 3, showed that the excess risk w.r.t. a class of VC-dimension V varies continuously from $\sqrt{V/n}$ to $V/(nh)$ (omitting log factors) as a function of the margin parameter h. In a completely different setting, [31], Theorem 5,

analyzed the minimax excess risk for nonparametric regression with well-specified and misspecified models. They showed a continuous transition from slow to faster rates when the distance of the regression function to the statistical model decreases to zero.

- To conclude the discussion, we mention other related works on binary classification with either high-dimensional or functional data (see also Remark 1.1 in the introduction). For instance, [28] studied the minimax rate of classification when X is drawn from a mixture of two high-dimensional Gaussian distributions $\mathcal{N}(\mu_0, \sigma^2 I_p)$ and $\mathcal{N}(\mu_1, \sigma^2 I_p)$ when $p \gg n$ and the vector $\mu_1 - \mu_0$ is sparse, with nearly matching lower bounds. Later [10] generalized these results to an unknown general covariance matrix Σ with bounded spectrum. (Earlier works include, e.g., that of [34].) In the infinite-dimensional setting we can also mention the work of [15], who proved a somewhat weak consistency result of the proposed classifier without any convergence rate in terms of n or Δ (among others). Finally, [3] showed universal consistency of the kNN functional classifier with the help of some Besicovich condition (see our Proposition 3 and our supplementary material [17]) without any rate of convergence, while they derived sub-optimal rate $n^{-1/6}$ with a plug-in rule in the Gaussian model. In our contribution, we provide detailed upper bounds with nearly matching lower bounds, thus characterizing the minimax rate of convergence as a function of the sample size n, the smoothness s, and the separation distance Δ.

Note also that, though the choice of the parameter d_n does not depend on Δ, it still depends on the (possibly unknown) smoothness parameter s. Though designing an adaptive classifier is beyond the scope of this paper, it might be addressed via the Lepski method (see, e.g., [27]) after adapting it to the classification setting.

4. Lower bounds on the excess risk

In this section, we derive two types of excess risk lower bounds.

The first one decays polynomially with n and applies to *any* classifier. This minimax lower bound indicates that, up to logarithmic factors, the excess risk of Theorem 3.1 cannot be improved in the worst case. This result is derived via standard nonparametric statistical tools (e.g., Fano's inequality) and is stated in Section 4.1.

Our second lower bound is of a different nature: it decays logarithmically with n and only applies to the nonparametric k-nearest neighbors algorithm evaluated on projected trajectories $\mathbf{X}_{i,d} \in \mathbb{R}^d$ and $\mathbf{X}_d \in \mathbb{R}^d$, where $\mathbf{X}_{i,d} = (\langle \varphi_j, X \rangle)_{1 \le j \le d}$ for all $i \in \{1, \dots, n\}$. We allow d to be chosen adaptively via a sample-splitting strategy, and we consider k tuned (optimally) as a function of d. Our logarithmic lower bound indicates that this popular algorithm is not fit for our particular model; see Section 4.2 below.

4.1. A general minimax lower bound

We provide a lower bound showing that the excess risk bound of Theorem 3.1 is minimax optimal up to logarithmic factors. The proof is postponed to the supplementary material [17].

Theorem 4.1. *Consider the statistical model* (1.1) *and the set* $\mathcal{H}_s(R)$ *defined in* (3.9), *where* $s, R > 0$ *and where* $(\varphi_j)_{j\leq 1}$ *is any Hilbert basis of* $\mathbb{L}^2([0,1])$. *Then, every classifier* $\widehat{\Phi}$ *satisfies, for any number* $n \geq \max\{R^{1/s}, (32\log(2)+2)^{2s+1}/(3R^2/4)\}$ *of i.i.d. observations* $(X_i, Y_i)_{1\leq i\leq n}$ *from* (1.1) *and all* $\Delta \in (0, R/2]$,

$$\sup_{\substack{f,g\in\mathcal{H}_s(R)\\ \|f-g\|\geq\Delta}} \left\{\mathcal{R}_{f,g}(\widehat{\Phi}) - \inf_{\Phi}\mathcal{R}_{f,g}(\Phi)\right\}$$

$$\geq \begin{cases} ce^{-2R^{2/(2s+1)}} R^{1/(2s+1)} n^{-s/(2s+1)} & \text{if } \Delta < R^{1/(2s+1)} n^{-s/(2s+1)} \\ \dfrac{ce^{-2\Delta^2}}{\Delta} R^{2/(2s+1)} n^{-2s/(2s+1)} & \text{if } \Delta \geq R^{1/(2s+1)} n^{-s/(2s+1)} \end{cases}$$

for some absolute constant $c > 0$.

We note two minor differences between the upper and lower bounds: Theorem 3.1 involves extra logarithmic factors, while Theorem 4.1 involves an extra term of $e^{-2\Delta^2}$. Fortunately both terms have a minor influence (note that $e^{-2\Delta^2} \geq e^{-8R^2}$ since $f, g \in \mathcal{H}_s(R)$). We leave the question of identifying the exact rate for future work.[1]

If we omit logarithmic factors and constant factors depending only on s and R, Theorems 3.1 and 4.1 together imply that, for $n \geq N_{s,R}$ large enough:

- when $\Delta \lesssim R^{1/(2s+1)} n^{-s/(2s+1)}$, the optimal worst-case excess risk is of the order of $n^{-s/(2s+1)}$;
- when $\Delta \gtrsim R^{1/(2s+1)} n^{-s/(2s+1)}$, the optimal worst-case excess risk is of the order of $n^{-2s/(2s+1)}/\Delta$.

4.2. Lower bound for the *k*-NN classifier

In this section, we focus on the k-nearest neighbor (kNN) classifier. This classification rule has been intensively studied over the past fifty years. In particular, this method provides interesting theoretical and practical properties. It is quite easy to handle and implement. Indeed, given a sample $\mathcal{S} = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$, a number of neighbors k, a norm $\|.\|$ and a new incoming observation, the kNN classifier is defined as

$$\Phi_{n,k}(X) = \mathbb{1}_{\{\frac{1}{k}\sum_{j=1}^{k} Y_{(j)}(X) > 1/2\}}, \tag{4.1}$$

where the $Y_{(j)}$ correspond to the label of the $X_{(j)}$ re-arranged according to the ordering

$$\|X_{(1)} - X\| \leq \cdots \leq \|X_{(n)} - X\|.$$

We refer the reader, see, for example, to [16,19] or [4] for more details.

[1] Possible solutions include: slightly improving Proposition 1 via a tighter Gaussian concentration bound (to gain a factor of nearly $e^{-\Delta^2/8}$), and optimizing the constant appearing in the exponential term of Lemma 4 of [17].

We are interested below in the performances of the classifier $\Phi_{n,k}$ in this functional setting. For this purpose, we will use the recent contribution of [12] that provides a lower bound of the misclassification rate of the kNN classifier in a very general framework. This lower bound is expressed as the measure of an uncertain set around $\eta \simeq 1/2$. We emphasize that we want to understand if a truncation strategy associated to a non parametric supervised classification approach is suitable for this kind of problem.

4.2.1. *Finite-dimensional case*

Smoothness parameter β. We shall consider first a finite d-dimensional case for our Gaussian translation model. In that case, Remark 3.1 in Section 3.3 reveals that the truncation approach problem we are studying is, without loss of generality, equivalent to a supervised classification in $\mathbb{R}^d$ where conditionally on the event $\{Y=0\}$ (resp. $\{Y=1\}$), $\mathbf{X}_d$ is a standard Gaussian variable (resp. a Gaussian random variable with mean m and variance 1). If γ_d refers to the Gaussian density:

$$\forall x \in \mathbb{R}^d \quad \gamma_d(x) := (2\pi)^{-d/2} e^{-\|x\|^2/2},$$

then in that case, the Bayes classifier in $\mathbb{R}^d$ is:

$$\Phi_d^\star(x) = \mathbb{1}_{\{\eta_d(x) \geq 1/2\}} \quad \text{with} \quad \eta_d(x) = \frac{\gamma_d(x)}{\gamma_d(x) + \gamma_d(x-m)} \quad \forall x \in \mathbb{R}^d,$$

In the following, to simplify the notations, we will drop the subscript d in all these terms and will write γ, η instead of γ_d, η_d. Following [12], the rate of convergence of the kNN depends on a smoothness parameter β involved in the next inequality:

$$\forall x \in \mathbb{R}^d \quad \left|\eta\big(B(x,r)\big) - \eta(x)\right| \leq L\mu\big(B(x,r)\big)^\beta \tag{4.2}$$

where $\eta(B(x,r))$ refers to the mean value of η on $B(x,r)$ w.r.t. the distribution of the design $\mathbf{X}$ given by $\mu = \frac{1}{2}\gamma(\cdot) + \frac{1}{2}\gamma(\cdot - m)$. Therefore, our first task is to determine the value of β in our Gaussian translation model. We begin with a simple proposition that entails that the value of β corresponding to our Gaussian translation model in $\mathbb{R}^d$ is $2/d$. The proof of Proposition 2 is postponed to the supplement [17].

Proposition 2. *Assume that $\|x\| \leq R$ for some $R \in \mathbb{R}^+$. Then an explicit constant L_R exists such that*

$$\forall r \leq \frac{1}{R} \quad \left|\eta\big(B(x,r)\big) - \eta(x)\right| \leq L_R \mu\big(B(x,r)\big)^{2/d}.$$

An important point given in the previous proposition is that when we are considering design points x such that $\|x\| \leq R/2$ and $\|m\| \leq R/2$, we then have

$$\forall r \leq 1 \; \forall x \in B(0, R/2) \quad \left|\eta\big(B(x,r)\big) - \eta(x)\right| \leq 60\pi e d R^2 e^{R^2/d} \mu\big(B(x,r)\big)^{2/d},$$

so that the constant L_R involved in the statement of Proposition 2 can be chosen as:

$$L_R = 60\pi e d R^2 e^{R^2/d} \tag{4.3}$$

According to inequality (4.2) and thanks to Proposition 2, the smoothness of the Gaussian translation model is given by:

$$\beta_d = 2/d.$$

Now, we slightly modify the approach of [12] to obtain a lower bound on the excess risk that involves the margin of the classification problem. As pointed above, in the Gaussian translation model, when the two classes are well separated (meaning that the center of the two classes are separated with a distance independent on n), the margin parameter is equal to 1 (see Theorem 1).

Optimal calibration of the kNN. Before giving our first result on the rate of convergence of the kNN classifier, we remind first some important facts regarding the choice of the number of neighbors k for the kNN classifier. The ability of the kNN to produce a universally consistent classification rule highly depends on the choice of the bandwidth parameter k_n. In particular, this bandwidth parameter must satisfy $k_n \longrightarrow +\infty$ and $k_n/n \longrightarrow 0$ as $n \longrightarrow +\infty$ to produce an asymptotically vanishing variance and bias (see, e.g., [16] for details). However, to obtain an optimal rate of convergence, k_n has to be chosen to produce a nice trade-off between the bias and the variance of the excess risk. It is shown in [12] that, when the marginal law of X is compactly supported, the optimal calibration k_n^{opt} is:

$$\frac{1}{\sqrt{k_n^{\text{opt}}(d)}} = c\left(\frac{k_n^{\text{opt}}(d)}{n}\right)^{\beta_d} \quad \Leftrightarrow \quad k_n^{\text{opt}}(d) \sim n^{\frac{4}{4+d}} \tag{4.4}$$

where c refers to any non negative constant and $\beta_d = 2/d$ refers to the smoothness parameter of the model involved in Inequality (4.2). On the other hand, it is shown in [18] that (almost) optimal rates of convergence can be obtained in the non-compact case, choosing for instance,

$$k_n \sim n^{\frac{2}{2+d+\tau}}, \tag{4.5}$$

for some positive τ. The following results provides a lower bound on the convergence rate with a number of neighbor k contained in a range of values.

Proposition 3. *For any $k, d \in \mathbb{N}$, we denote by $\Phi_{k,n,d}$ the kNN classifier based on the training sample $((\mathbf{X}_{i,d}, Y_i))_{i=1,\dots,n}$. There exists a constant C_1 such that for any $d \in \mathbb{N}$*

$$\mathcal{R}_{f,g}(\Phi_{k,n,d}) - \mathcal{R}(\Phi_d^\star) \geq \frac{C_1}{k_n}$$

when $k \in \mathcal{K}_n$ where

$$\mathcal{K}_n = \mathcal{K}_n(d) = \left\{\ell \in \mathbb{N} \text{ s.t. } \frac{1}{\sqrt{\ell}} \geq d\left(\frac{\ell}{n}\right)^{2/d} \text{ and } \ell \leq n\right\}.$$

The proof of this result is given in supplementary material [17].

Remark 4.1. Proposition 3 is an important intermediary result to understand the behaviour of kNN with functional data. We briefly comment on this result below.

- The set $\mathcal{K}_n$ contains all the integers from 1 to an integer equivalent to $n^{4/(4+d)}d^{-2d/(4+d)}$. In particular, the "optimal" standard calibration of k_n given by Equation (4.4) is included in the set $\mathcal{K}_n$ and Proposition 3 applies in particular for such a calibration.
- Proposition 3 entails that tuning the kNN classifier in an "optimal way" cannot produce faster rates of convergence than $n^{-4/(d+4)}$, even with some additional informations on the considered model (here the Gaussian distribution of the conditional distributions):

  $$\mathcal{R}_{f,g}(\Phi_{k_n^{\text{opt}}(d),n,d}) - \mathcal{R}(\Phi_d^\star) \geq C_1 n^{-\frac{4}{d+4}}.$$

  These performances have to be compared to those obtained with our procedure that explicitly exploits the additional knowledge of Gaussian conditional distributions (see, e.g., Lemma 2).
- The last important point is that the lower bound in the statement of Proposition 3 appears to be seriously damaged when d increases. This is a classical feature of the curse of dimensionality. For us, it invalidates any approach that will jointly associate a truncation strategy with a kNN plug-in classifier: we will be led to choose d large with n to avoid too much loss of information but in the same time this will harms the statistical misclassification.

4.2.2. *Lower bound of the misclassification rate with truncated strategies*

As pointed by Proposition 3, the global behavior of the kNN classifier heavily depends on the choice of the dimension d. In the same time, the size of d is important to obtain a truncated Bayes classifier $\Phi_d^\star$ close to the Bayes classifier $\Phi^\star$. To assess the performance of kNN, we consider a sample splitting strategy $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$ where $(\mathcal{S}_1, \mathcal{S}_2)$ is a partition of $\mathcal{S}$ where the size of both partitions is proportional to n. Then, $\mathcal{S}_1$ is used to choose a dimension $\widehat{d}$, then we apply an optimal kNN classifier method based on the samples of $\mathcal{S}_2$ on the truncated spaces with $\Pi_{\widehat{d}}$ with $k_n^{\text{opt}}(\widehat{d})$ chosen as in Equation (4.4). It is important to note that the sample splitting strategy produces a choice $\widehat{d}$ independent on the samples in $\mathcal{S}_2$.

Theorem 4.2 below shows that for any sample splitting strategy, every choice of $\widehat{d}$ will lead to bad performances of classification on model (1.1). The proof is postponed to supplementary material [17].

Theorem 4.2. *Consider the partition $(\mathcal{S}_1, \mathcal{S}_2)$ of the sample $\mathcal{S}$. Whatever the selection rule $\widehat{d}$ based on $\mathcal{S}_1$ for the dimension d, any kNN classifier $\Phi_{k_n^{\text{opt}},n,\widehat{d}}$ based on the sample $\mathcal{S}_2$ has an excess risk with a logarithmic decreasing order. More precisely*

$$\inf_{\widehat{d}\in\mathbb{N}} \sup_{f,g\in\mathcal{H}_s(r)} \max_{k\in\mathcal{K}_n(\widehat{d})} \mathcal{R}_{f,g}(\Phi_{k,n,\widehat{d}}) - R_n(\Phi^\star) \gtrsim \log(n)^{-2s},$$

where the infimum over $\widehat{d}$ is taken over any tuning strategy for d based on the first sample.

The main conclusion of this section and of Theorem 4.2 is that the kNN rule based on a truncation strategy does not lead to satisfying rates of convergence, regardless the choice of the dimension $\widehat{d}$ is. We stress that this result is only valid for a specific range for k. Although this appears to cover classical tuning approach regarding the existing literature as, for example, those mentioned in (4.4) and (4.5), obtaining a global lower bound (i.e., for any choice of k) remains an open (and difficult) problem. Even though we suspect that such a logarithmic lower bound also holds for some more general procedures (without sample splitting and with a more general possible choice of k_n), we do not have any proof of such a result.

4.2.3. *Further comments*

It should be kept in mind that the misclassification excess risk of the suitably truncated LDA classifier $\widehat{\Phi}_{d_n}$ proposed in Equation (3.4) decreases at a polynomial rate, which is an important encouragement for its use. We stress that the difference in terms of performances between our classifier and the kNN classifier finds its origin in the finite dimensional case (Section 4.2.1). Indeed, comparing Lemma 1 and Lemma 2 with Proposition 3 and the choice of k given in Equation (4.4) indicates that the dependance w.r.t. the dimension d is completely different from one method to another. This can be explained by the fact that, for any fixed d, our procedure takes advantage of the Gaussian behavior of the data and is essentially parametric. On the other hand, the kNN classifier is a non-parametric method: it is hence outperformed in Gaussian situations. However, it appears to be more robust w.r.t. any misspecification of the model, which is an important feature of non-parametric classifiers.

Appendix: Proof of the upper bounds

The goal of this section is to prove the polynomial upper bound of Theorem 3.1 together with the intermediate results of Proposition 1 and Lemmas 1 and 2. We will pay a specific attention to the acceleration (in terms of the number n of samples) obtained when the functions f and g appearing in (1.1) are well separated.

A.1. Preliminary comments: More formal definition of the two subsamples

For the sake of rigor, we provide a more formal definition of the quantities introduced in Section 3.1.1. The conditional independence property stated in Remark A.1 below will also be useful to control the estimation error in the risk bounds.

Recall that we split the sample $(X_i)_{1\le i\le n}$ into two subsamples $(X_i^0)_{1\le i\le N_0}$ and $(X_i^1)_{1\le i\le N_1}$ corresponding to either $Y_i=0$ or $Y_i=1$. More formally, if $\tau_1^0, \tau_2^0, \dots, \tau_{N_0}^0$ refers to a numbering of the points in class 0 and $\tau_1^1, \tau_2^1, \dots, \tau_{N_1}^1$ to a numbering of the points in class 1, we define the two subsamples $(X_i^0)_{1\le i\le N_0}$ and $(X_i^1)_{1\le i\le N_1}$ (one of which can be empty) by

$$\begin{cases} X_i^0 := X_{\tau_i^0}, & 1\le i\le N_0 \\ X_i^1 := X_{\tau_i^1}, & 1\le i\le N_1 \end{cases}$$

where

$$N_0 := \sum_{i=1}^{n} \mathbb{1}_{\{Y_i=0\}} \quad \text{and} \quad N_1 := \sum_{i=1}^{n} \mathbb{1}_{\{Y_i=1\}}. \tag{A.1}$$

Formally, τ_i^k is the index $t \in \{1, \dots, n\}$ such that $Y_t = k$ for the i-th time, that is, for all $k \in \{0, 1\}$ and $i \in \{1, \dots, N_k\}$,

$$\tau_i^k := \min\left\{ t \in \{1, \dots, n\} : \sum_{t'=1}^{t} \mathbb{1}_{\{Y_{t'}=k\}} \geq i \right\}.$$

The random coefficients $(X_{i,j}^0)_{\substack{1\leq i\leq N_0\\ 1\leq j\leq d}}$ and $(X_{i,j}^1)_{\substack{1\leq i\leq N_1\\ 1\leq j\leq d}}$ are then given by

$$\begin{cases} X_{i,j}^0 := \langle \varphi_j, X_i^0 \rangle = \mu_j + \varepsilon_{i,j}^0, & i = 1, \dots, N_0, \\ X_{i,j}^1 := \langle \varphi_j, X_i^1 \rangle = \theta_j + \varepsilon_{i,j}^1, & i = 1, \dots, N_1, \end{cases} \quad j \in \{1, \dots, d\}, \tag{A.2}$$

for some suitable dimension $d \in \mathbb{N}^*$. The above equalities involving μ_j or θ_j are obtained by Section 1.2 and by setting

$$\varepsilon_{i,j}^k = \int_0^1 \varphi_j(t)\, dW_{\tau_i^k}(t),$$

where the W_i denote i.i.d copies of the standard Brownian motion $(W(t))_{0\leq t\leq 1}$ associated with the X_i. By independence of the random variables $Y_1, W_1, \dots, Y_n, W_n$ used to generate the sample $(X_i, Y_i)_{1\leq i\leq n}$ according to (1.1), and by the comments made in Section 1.2, we have the following conditional independence property for the $\varepsilon_{i,j}^k$.

Remark A.1. Conditionally on $(Y_1, \dots, Y_n)$, the nd random variables (or any $(Y_1, \dots, Y_n)$-measurable permutation of them)

$$\varepsilon_{1,1}^0, \dots, \varepsilon_{1,d}^0, \varepsilon_{2,1}^0, \dots, \varepsilon_{2,d}^0, \dots, \varepsilon_{N_0,1}^0, \dots, \varepsilon_{N_0,d}^0,$$
$$\varepsilon_{1,1}^1, \dots, \varepsilon_{1,d}^1, \varepsilon_{2,1}^1, \dots, \varepsilon_{2,d}^1, \dots, \varepsilon_{N_1,1}^1, \dots, \varepsilon_{N_1,d}^1$$

are i.i.d. $\mathcal{N}(0, 1)$. As a consequence, on the event $\{N_0 > 0\} \cap \{N_1 > 0\}$, the random variables $N_k^{-1/2} \sum_{i=1}^{N_k} \varepsilon_{i,j}^k$, $1 \leq j \leq d$, $k \in \{0, 1\}$, are i.i.d. $\mathcal{N}(0, 1)$ conditionally on $(Y_1, \dots, Y_n)$.

A.2. Proof of Proposition 1 (control of the margin)

We start by proving Proposition 1, that is, we analyze the margin behavior in Model (1.1). This result is a key ingredient to derive our excess risk upper bounds.

Proof of Proposition 1. We use the Girsanov Equations (2.1) and (2.2) that define the likelihood ratio q_f and q_g. We therefore deduce that

$$\begin{aligned}
&\mathbb{P}_X\left(\left|\eta(X)-\frac{1}{2}\right|\le\varepsilon\right)\\
&\quad=\mathbb{P}_X\left(\frac{|q_f(X)-q_g(X)|}{2(q_f(X)+q_g(X))}\le\varepsilon\right)\\
&\quad=\mathbb{P}_X\left(\left\{\frac{|q_f(X)-q_g(X)|}{2(q_f(X)+q_g(X))}\le\varepsilon\right\}\cap\left\{q_f(X)\le q_g(X)\right\}\right)\\
&\qquad+\mathbb{P}_X\left(\left\{\frac{|q_f(X)-q_g(X)|}{2(q_f(X)+q_g(X))}\le\varepsilon\right\}\cap\left\{q_f(X)>q_g(X)\right\}\right)\\
&\quad\le\mathbb{P}_X\left(\left\{\frac{|q_f(X)-q_g(X)|}{4q_g(X)}\le\varepsilon\right\}\cap\left\{q_f(X)\le q_g(X)\right\}\right)\\
&\qquad+\mathbb{P}_X\left(\left\{\frac{|q_f(X)-q_g(X)|}{4q_f(X)}\le\varepsilon\right\}\cap\left\{q_f(X)>q_g(X)\right\}\right)\\
&\quad\le\mathbb{P}_X\left(\left|\frac{q_f(X)}{q_g(X)}-1\right|\le4\varepsilon\right)+\mathbb{P}_X\left(\left|\frac{q_g(X)}{q_f(X)}-1\right|\le4\varepsilon\right).
\end{aligned}\tag{A.3}$$

The two terms of the last line are handled similarly, and we only deal with the first one. We note that

$$\frac{q_f(X)}{q_g(X)}=\exp\left(\int_0^1(f-g)(s)\,dX(s)-\frac{1}{2}\left[\|f\|^2-\|g\|^2\right]\right).$$

Using the fact that $Y\sim\mathcal{B}(1/2)$ and conditioning by $Y=1$ and $Y=0$, we can see that

$$\begin{aligned}
&\mathbb{P}_X\left(\left|\frac{q_f(X)}{q_g(X)}-1\right|\le4\varepsilon\right)\\
&\quad=\mathbb{P}\left(\left|e^{\int_0^1(f-g)(s)f(s)\,ds+\int_0^1(f-g)(s)\,dW(s)-\frac{1}{2}[\|f\|^2-\|g\|^2]}-1\right|\le4\varepsilon\right)\mathbb{P}(Y=1)\\
&\qquad+\mathbb{P}\left(\left|e^{\int_0^1(f-g)(s)g(s)\,ds+\int_0^1(f-g)(s)\,dW(s)-\frac{1}{2}[\|f\|^2-\|g\|^2]}-1\right|\le4\varepsilon\right)\mathbb{P}(Y=0)\\
&\quad=\frac{1}{2}\mathbb{P}\left(\left|e^{\frac{1}{2}\|f-g\|^2+\int_0^1(f-g)(s)\,dW(s)}-1\right|\le4\varepsilon\right)\\
&\qquad+\frac{1}{2}\mathbb{P}\left(\left|e^{-\frac{1}{2}\|f-g\|^2+\int_0^1(f-g)(s)\,dW(s)}-1\right|\le4\varepsilon\right)\\
&\quad=\frac{1}{2}\mathbb{P}\left(\left|e^{\frac{1}{2}\Delta^2+\Delta\xi}-1\right|\le4\varepsilon\right)+\frac{1}{2}\mathbb{P}\left(\left|e^{-\frac{1}{2}\Delta^2+\Delta\xi}-1\right|\le4\varepsilon\right),
\end{aligned}\tag{A.4}$$

where $\Delta:=\|f-g\|$ and $\xi\sim\mathcal{N}(0,1)$ because $\int_0^1[f(s)-g(s)]\,dW(s)\sim\mathcal{N}(0,\Delta^2)$.

Using the inequalities $\ln(1+4\varepsilon) \le 4\varepsilon$ and $\ln(1-4\varepsilon) \ge -8\varepsilon$ when $\varepsilon \le 1/8$, the above probability can be upper bounded as

$$\begin{aligned}\mathbb{P}_X\left(\left|\frac{q_f(X)}{q_g(X)} - 1\right| \le 4\varepsilon\right) &\le \frac{1}{2}\mathbb{P}\left(-\frac{8\varepsilon}{\Delta} - \frac{\Delta}{2} \le \xi \le \frac{4\varepsilon}{\Delta} - \frac{\Delta}{2}\right) \\ &\quad + \frac{1}{2}\mathbb{P}\left(-\frac{8\varepsilon}{\Delta} + \frac{\Delta}{2} \le \xi \le \frac{4\varepsilon}{\Delta} + \frac{\Delta}{2}\right), \\ &\le \frac{5\varepsilon}{\Delta},\end{aligned}$$

where the last inequality follows from $\mathbb{P}(a \le \xi \le b) \le (b-a)/\sqrt{2\pi}$ and $12/\sqrt{2\pi} \le 5$. Inverting the roles of f and g, we get by symmetry of the problem that the second term of (A.3) is also upper bounded by $5\varepsilon/\Delta$. This concludes the proof. □

Remark A.2. Following the same proof strategy, it is easy to check that the same result holds in the truncated space, that is, replacing η with η_d and Δ with $\Delta_d := \|\Pi_d(f-g)\|$. Namely, for all $d \in \mathbb{N}^*$ and all $0 < \varepsilon \le 1/8$,

$$\mathbb{P}_X\left(\left|\eta_d(X) - \frac{1}{2}\right| \le \varepsilon\right) \le 1 \wedge \frac{10\varepsilon}{\Delta_d}.$$

In particular, Equation (A.4) holds with $q_{f_d}(X)/q_{g_d}(X)$ on the left-hand side and with Δ_d on the right-hand side because $\int_0^1 \Pi_d(f-g)(s)\,dW(s) \sim \mathcal{N}(0, \Delta_d^2)$.

A.3. Proof of Lemma 1 (control of the approximation error)

One key ingredient of the proof is to control the excess risk $\mathcal{R}_{f,g}(\Phi_d^\star) - \mathcal{R}_{f,g}(\Phi^\star)$ in terms of the closeness of f_d and g_d to f and g respectively. To do so, we set

$$\delta_d := \|f - f_d\| = \sqrt{\|f\|^2 - \|f_d\|^2} \quad \text{and} \quad \widetilde{\delta}_d := \|g - g_d\| = \sqrt{\|g\|^2 - \|g_d\|^2}.$$

Proof of Lemma 1. We start with the well-known formula on the excess risk of any classifier (see, e.g., [19]):

$$\mathcal{R}_{f,g}(\Phi_d^\star) - \mathcal{R}_{f,g}(\Phi^\star) = \mathbb{E}\big[\big|2\eta(X) - 1\big|\mathbb{1}_{\{\Phi_d^\star(X) \ne \Phi^\star(X)\}}\big].$$

Then, following a classical control of the excess risk (see, e.g., [18]),

$$\begin{aligned}&\mathcal{R}_{f,g}(\Phi_d^\star) - \mathcal{R}_{f,g}(\Phi^\star) \\ &\quad = \mathbb{E}\big[\big|2\eta(X) - 1\big|\mathbb{1}_{\{\Phi_d^\star(X) \ne \Phi^\star(X)\}}[\mathbb{1}_{\{|\eta(X)-1/2| \le \varepsilon\}} + \mathbb{1}_{\{|\eta(X)-1/2| > \varepsilon\}}]\big] \\ &\quad \le \underbrace{2\varepsilon\mathbb{P}\big(\big|\eta(X) - 1/2\big| \le \varepsilon\big)}_{:=T_1} + \underbrace{\mathbb{P}\big(\big\{\Phi_d^\star(X) \ne \Phi^\star(X)\big\} \cap \big\{\big|\eta(X) - 1/2\big| > \varepsilon\big\}\big)}_{:=T_2}. \qquad \text{(A.5)}\end{aligned}$$

Note that, up to the quantity 2ε, the term T_1 corresponds to the margin behavior discussed in Section 2.2 above. By Proposition 1 (note that $0 < \varepsilon \le 1/8$), we have

$$T_1 := 2\varepsilon \mathbb{P}\big(\big|\eta(X) - 1/2\big| \le \varepsilon\big) \le 2\varepsilon\left(1 \wedge \frac{10\varepsilon}{\Delta}\right).$$

To control the second term T_2, we note (classically) that $\Phi^\star(X) = \mathbb{1}_{\{\eta(X)\ge 1/2\}}$ and $\Phi_d^\star(X) = \mathbb{1}_{\{\eta_d(X)\ge 1/2\}}$ together imply that

$$T_2 := \mathbb{P}\big(\big\{\Phi_d^\star(X) \ne \Phi^\star(X)\big\} \cap \big\{\big|\eta(X) - 1/2\big| > \varepsilon\big\}\big) \le \mathbb{P}\big(\big|\eta_d(X) - \eta(X)\big| > \varepsilon\big).$$

Using $Y \sim \mathcal{B}(1/2)$ and the conditional distribution of $X|Y$, we have

$$T_2 \le \frac{1}{2}\underbrace{\mathbb{P}_f\big(\big|\eta_d(X) - \eta(X)\big| > \varepsilon\big)}_{:=T_{2,1}} + \frac{1}{2}\underbrace{\mathbb{P}_g\big(\big|\eta_d(X) - \eta(X)\big| > \varepsilon\big)}_{:=T_{2,2}}.$$

For the sake of brevity, we only study $T_{2,1}$ (the second term $T_{2,2}$ can be upper bounded similarly by symmetry of the problem and by inverting the roles of f and g). Recall from (2.1) that q_f denotes the likelihood ratio of the model $\mathbb{P}_f$. Next, we decompose $\eta - \eta_d$ using the four (a.s. positive) likelihood ratios q_f, q_g, q_{f_d}, and q_{g_d}:

$$\eta - \eta_d = \frac{q_f}{q_f + q_g} - \frac{q_{f_d}}{q_{f_d} + q_{g_d}} = \frac{q_f - q_{f_d}}{q_f + q_g} + q_{f_d}\left(\frac{1}{q_f + q_g} - \frac{1}{q_{f_d} + q_{g_d}}\right).$$

In order to upper bound $T_{2,1}$, we use the triangle inequality three times in the decomposition above, we note that

$$\left|\frac{1}{q_f + q_g} - \frac{1}{q_{f_d} + q_{g_d}}\right| = \left|\frac{q_{f_d} - q_f + q_{g_d} - q_g}{(q_f + q_g)(q_{f_d} + q_{g_d})}\right| \le \frac{|q_{f_d} - q_f|}{q_f q_{f_d}} + \frac{|q_{g_d} - q_g|}{q_g q_{f_d}},$$

and we use the inclusion $\{Z_1 + Z_2 + Z_3 > \varepsilon\} \subseteq \{Z_1 > \varepsilon/2\} \cup \{Z_2 > \varepsilon/4\} \cup \{Z_3 > \varepsilon/4\}$ valid or any random variables Z_1, Z_2, Z_3. We get:

$$\begin{aligned}
&T_{2,1} \\
&\quad \le \mathbb{P}_f\left(\big|q_{f_d}(X) - q_f(X)\big| > \frac{\varepsilon}{2}\big|q_f(X) + q_g(X)\big|\right) \\
&\qquad + \mathbb{P}_f\left(q_{f_d}(X)\big|q_f(X) - q_{f_d}(X)\big| > \frac{\varepsilon}{4} q_f(X) q_{f_d}(X)\right) \\
&\qquad + \mathbb{P}_f\left(q_{f_d}(X)\big|q_g(X) - q_{g_d}(X)\big| > \frac{\varepsilon}{4} q_g(X) q_{f_d}(X)\right) \\
&\quad \le \mathbb{P}_f\left(\big|q_{f_d}(X) - q_f(X)\big| > \frac{\varepsilon}{2} q_f(X)\right) + \mathbb{P}_f\left(\big|q_{f_d}(X) - q_f(X)\big| > \frac{\varepsilon}{4} q_f(X)\right)
\end{aligned}$$

$$
\begin{aligned}
&+\mathbb{P}_f\Big(\big|q_{g_d}(X)-q_g(X)\big|>\frac{\varepsilon}{4}q_g(X)\Big)\\
&\leq 2\mathbb{P}_f\Big(\big|q_{f_d}(X)-q_f(X)\big|>\frac{\varepsilon}{4}q_f(X)\Big)+\mathbb{P}_f\Big(\big|q_{g_d}(X)-q_g(X)\big|>\frac{\varepsilon}{4}q_g(X)\Big)
\end{aligned}
$$

Taking the logarithm, we can see that:

$$
\begin{aligned}
\mathbb{P}_f\left(\left|\frac{q_{f_d}(X)}{q_f(X)}-1\right|>\frac{\varepsilon}{4}\right)&=\mathbb{P}_f\left(\log\left(\frac{q_{f_d}}{q_f}\right)<\log(1-\varepsilon/4)\right)\\
&\quad+\mathbb{P}_f\left(\log\left(\frac{q_{f_d}}{q_f}\right)>\log(1+\varepsilon/4)\right)
\end{aligned}
$$

Using the inequalities $\log(1+\varepsilon/4)\geq\varepsilon/8$ and $\log(1-\varepsilon/4)\leq-\varepsilon/4$ (that hold at least for all $0<\varepsilon\leq 1$) we obtain:

$$
\begin{aligned}
T_{2,1}\leq{}&\underbrace{2\,\mathbb{P}_f\left(\log\frac{q_{f_d}(X)}{q_f(X)}>\varepsilon/8\right)}_{:=S_1}+\underbrace{2\,\mathbb{P}_f\left(\log\frac{q_{f_d}(X)}{q_f(X)}<-\frac{\varepsilon}{4}\right)}_{:=S_2}\\
&+\underbrace{\mathbb{P}_f\left(\log\frac{q_{g_d}(X)}{q_g(X)}>\frac{\varepsilon}{8}\right)}_{:=S_3}+\underbrace{\mathbb{P}_f\left(\log\frac{q_{g_d}(X)}{q_g(X)}<-\frac{\varepsilon}{4}\right)}_{:=S_4}.
\end{aligned}
\tag{A.6}
$$

The Girsanov formula makes it possible to write $\log\frac{q_{f_d}(X)}{q_f(X)}=\int_0^1(f_d-f)(s)\,dX(s)-\frac{1}{2}[\|f_d\|^2-\|f\|^2]$. We study S_1 and remark that under $\mathbb{P}_f$, $dX(s)=f(s)\,ds+dW(s)$ for all $s\in[0,1]$ so that

$$
\begin{aligned}
S_1&=\mathbb{P}_f\left(\log\frac{q_{f_d}(X)}{q_f(X)}>\frac{\varepsilon}{8}\right)\\
&=\mathbb{P}\left(\langle f_d-f,f\rangle+\int_0^1(f_d-f)(s)\,dW(s)-\frac{1}{2}\big[\|f_d\|^2-\|f\|^2\big]>\frac{\varepsilon}{8}\right)\\
&=\mathbb{P}\left(\int_0^1(f_d-f)(s)\,dW(s)>\frac{1}{2}\big[\|f\|^2-\|f_d\|^2\big]+\frac{\varepsilon}{8}\right)\\
&\leq\mathbb{P}\left(\xi>\frac{\varepsilon}{8}\right),
\end{aligned}
$$

where $\xi\sim\mathcal{N}(0,\|f_d-f\|^2)=\mathcal{N}(0,\delta_d^2)$. But, by a classical (sub)Gaussian tail bound stated, for example, in [8], p. 22, we get

$$
S_1\leq\exp\left(-\frac{(\varepsilon/8)^2}{2\delta_d^2}\right)=\exp\left(-\frac{\varepsilon^2}{128\delta_d^2}\right).\tag{A.7}
$$

Combining the last inequality with the assumption

$$\delta_d^2 \le \frac{\varepsilon^2}{512\ln(1/\varepsilon^2)} \le \frac{\varepsilon^2}{128\ln(1/\varepsilon^2)},$$

we finally obtain $S_1 \le \varepsilon^2$.

The second term S_2 introduced in (A.6) can be dealt similarly, except that we can no longer neglect the positive term $(\|f\|^2 - \|f_d\|^2)/2 = \delta_d^2/2$: considering again $\xi \sim \mathcal{N}(0, \delta_d^2)$, we have

$$S_2 = \mathbb{P}_f\left(\log\frac{q_{f_d}(X)}{q_f(X)} < -\frac{\varepsilon}{4}\right) = \mathbb{P}\left(\xi < \frac{\delta_d^2}{2} - \frac{\varepsilon}{4}\right) \le \mathbb{P}\left(\xi < -\frac{\varepsilon}{8}\right) \le \varepsilon^2,$$

where the last inequality follows from the same Gaussian concentration argument as in (A.7), and where the inequality before last is because $\delta_d^2/2 \le \varepsilon/8$. Indeed, by the assumptions of Lemma 1,

$$\varepsilon \ge \sqrt{512\ln(8^2)}\max\{\delta_d, \widetilde{\delta}_d\} \ge 24\max\{\delta_d^2, \widetilde{\delta}_d^2\}, \tag{A.8}$$

where the second inequality follows from $\max\{\delta_d, \widetilde{\delta}_d\} \le \varepsilon/24 \le 1$ (as a result of the first inequality and $\varepsilon \le 1$). Therefore, $\delta_d^2/2 \le \varepsilon/48 \le \varepsilon/8$ as claimed above.

We now focus on S_3: noting that $\log\frac{q_{g_d}(X)}{q_g(X)} = \int_0^1 (g_d - g)(s)\,dX(s) - \frac{1}{2}[\|g_d\|^2 - \|g\|^2]$, we get

$$\begin{aligned}
S_3 &= \mathbb{P}_f\left(\log\frac{q_{g_d}(X)}{q_g(X)} > \frac{\varepsilon}{8}\right)\\
&= \mathbb{P}\left(\langle g_d - g, f\rangle + \int_0^1 (g_d - g)(s)\,dW(s) - \frac{1}{2}\big[\|g_d\|^2 - \|g\|^2\big] > \frac{\varepsilon}{8}\right)\\
&= \mathbb{P}\left(\langle g_d - g, f - f_d\rangle + \int_0^1 (g_d - g)(s)\,dW(s) - \frac{1}{2}\big[\|g_d\|^2 - \|g\|^2\big] > \frac{\varepsilon}{8}\right),
\end{aligned}$$

where we used the fact that $g_d - g$ and f_d are orthogonal. Recall now that $\delta_d = \|f - f_d\|$ and $\widetilde{\delta}_d = \|g - g_d\| = \sqrt{\|g\|^2 - \|g_d\|^2}$. If $\widetilde{\xi} \sim \mathcal{N}(0, \widetilde{\delta}_d^2)$, the last equality entails

$$S_3 \le \mathbb{P}\left(\widetilde{\xi} > \frac{\varepsilon}{8} - \frac{\widetilde{\delta}_d^2}{2} - \widetilde{\delta}_d\delta_d\right) \le \mathbb{P}\left(\widetilde{\xi} > \frac{\varepsilon}{8} - \frac{\widetilde{\delta}_d^2}{2} - \frac{\widetilde{\delta}_d^2}{2} - \frac{\delta_d^2}{2}\right) \le \mathbb{P}\left(\widetilde{\xi} > \frac{\varepsilon}{16}\right) \le \varepsilon^2,$$

where the second inequality follows from $\widetilde{\delta}_d\delta_d \le (\widetilde{\delta}_d + \delta_d^2)/2$, where the third inequality is because $\max\{\delta_d^2, \widetilde{\delta}_d^2\} \le \varepsilon/24$ (by (A.8)), and where the last inequality follows from the same Gaussian tail bound as the one used in (A.7) and from the assumption $\widetilde{\delta}_d^2 \le \varepsilon^2/(512\ln(1/\varepsilon^2))$.

A similar analysis shows that the last term S_4 introduced in (A.6) also satisfies $S_4 \le \varepsilon^2$. Putting everything together, we finally get

$$T_{2,1} \le 6\varepsilon^2.$$

By symmetry of the problem and by inverting the roles of f and g, we can also see that $T_{2,2} \leq 6\varepsilon^2$. Summing the bounds on T_1, $T_{2,1}$, and $T_{2,2}$ concludes the proof. □

A.4. Proof of Lemma 2 (control of the estimation error)

Though we now focus on the estimation error, most of the proof follows similar arguments as for Lemma 1 above: comparison of two regression functions, and Gaussian-type concentration inequalities.

Proof of Lemma 2. Recall from Remark 3.1 (Section 3.3) that η_d and $\Phi_d^* = \mathbb{1}_{\{\eta_d \geq 1/2\}}$ correspond to the regression function and the Bayes classifier of the classification problem when the learner has only access to the projected input $\mathbf{X}_d := (\langle \varphi_j, X\rangle)_{1 \leq j \leq d}$. Since $\widehat{\Phi}_d(X)$ only depends on X through $\mathbf{X}_d$, its excess risk can be rewritten as

$$\mathcal{R}_{f,g}(\widehat{\Phi}_d) - \mathcal{R}_{f,g}(\Phi_d^\star) = \mathbb{E}\big[\big|2\eta_d(X) - 1\big| \mathbb{1}_{\{\widehat{\Phi}_d(X) \neq \Phi_d^\star(X)\}}\big],$$

where the expectation is with respect to both the sample $(X_i, Y_i)_{1 \leq i \leq n}$ and the new input X. Now, for all $\varepsilon > 0$, using a bound similar to (A.5), we get

$$\mathcal{R}_{f,g}(\widehat{\Phi}_d) - \mathcal{R}_{f,g}(\Phi_d^\star) \leq 2\varepsilon \mathbb{P}_X\big(\big|\eta_d(X) - 1/2\big| \leq \varepsilon\big) + \mathbb{P}\big(\big|\widehat{\eta}_d(X) - \eta_d(X)\big| > \varepsilon\big),$$

where the last inequality follows from the inclusion $\{\widehat{\Phi}_d(X) \neq \Phi_d^\star(X)\} \cap \{|\eta_d(X) - 1/2| > \varepsilon\} \subseteq \{|\widehat{\eta}_d(X) - \eta_d(X)| > \varepsilon\}$ (because $\widehat{\Phi}_d(X) = \mathbb{1}_{\{\widehat{\eta}_d(X) \geq 1/2\}}$ and $\Phi_d^*(X) = \mathbb{1}_{\{\eta_d(X) \geq 1/2\}}$). We can now apply the adaptation of Proposition 1 to the truncated space (see Remark A.2) to get

$$\mathcal{R}_{f,g}(\widehat{\Phi}_d) - \mathcal{R}_{f,g}(\Phi_d^\star) \leq 2\varepsilon\left(1 \wedge \frac{10\varepsilon}{\Delta_d}\right) + \mathbb{P}\big(\big|\widehat{\eta}_d(X) - \eta_d(X)\big| > \varepsilon\big). \tag{A.9}$$

Using $Y \sim \mathcal{B}(1/2)$ and the conditional distribution of X given Y, we have:

$$\mathbb{P}\big(\big|\widehat{\eta}_d(X) - \eta_d(X)\big| > \varepsilon\big) = \frac{1}{2}\underbrace{\mathbb{P}_f\big(\big|\widehat{\eta}_d(X) - \eta_d(X)\big| > \varepsilon\big)}_{:=T_1} + \frac{1}{2}\underbrace{\mathbb{P}_g\big(\big|\widehat{\eta}_d(X) - \eta_d(X)\big| > \varepsilon\big)}_{:=T_2},$$

where, with a slight abuse of notation, the first probability $\mathbb{P}_f(\cdot)$ is with respect to both the sample $(X_i, Y_i)_{1 \leq i \leq n}$ drawn i.i.d. from (1.1) and a new independent input X drawn from $\mathbb{P}_f$; and similarly for the second probability $\mathbb{P}_g(\cdot)$.

We now focus on T_1 until the end of the proof. (The control of T_2 is exactly similar, by symmetry of the model and by inverting the roles of f and g.) Denote by $\gamma_d(x) = (2\pi)^{-d/2} e^{-\|x\|^2/2}$ the density of the standard Gaussian distribution on $\mathbb{R}^d$. By Remark 3.1 (Section 3.3), we have, setting $\boldsymbol{\theta}_d := (\theta_1, \dots, \theta_d)$ and $\boldsymbol{\mu}_d := (\mu_1, \dots, \mu_d)$,

$$\eta_d(X) = \frac{F_d(X)}{F_d(X) + G_d(X)}, \quad \text{where } F_d(x) = \gamma_d(x - \boldsymbol{\theta}_d) \text{ and } G_d(x) = \gamma_d(x - \boldsymbol{\mu}_d).$$

Similarly, by (3.7), the estimated regression function $\widehat{\eta}_d$ can be rewritten as

$$\widehat{\eta}_d(X) = \frac{\widehat{F}_d(X)}{\widehat{F}_d(X) + \widehat{G}_d(X)}, \quad \text{where } \widehat{F}_d(x) = \gamma_d(x - \widehat{\theta}) \quad \text{and} \quad \widehat{G}_d(x) = \gamma_d(x - \widehat{\mu}).$$

Using simple algebra, we get

$$\begin{aligned} T_1 &:= \mathbb{P}_f\big(\big|\widehat{\eta}_d(X) - \eta_d(X)\big| > \varepsilon\big) \\ &= \mathbb{P}_f\left(\left|\frac{\widehat{F}_d(X)}{\widehat{F}_d(X) + \widehat{G}_d(X)} - \frac{F_d(X)}{F_d(X) + G_d(X)}\right| > \varepsilon\right) \\ &\leq \mathbb{P}_f\left(\left|\frac{\widehat{F}_d(X) - F_d(X)}{F_d(X) + G_d(X)} + \widehat{F}_d(X)\left(\frac{1}{\widehat{F}_d(X) + \widehat{G}_d(X)} - \frac{1}{F_d(X) + G_d(X)}\right)\right| > \varepsilon\right) \\ &\leq \mathbb{P}_f\left(\left|\frac{\widehat{F}_d(X) - F_d(X)}{F_d(X) + G_d(X)}\right| > \frac{\varepsilon}{3}\right) \\ &\quad + \mathbb{P}_f\left(\left|\widehat{F}_d(X)\left(\frac{1}{\widehat{F}_d(X) + \widehat{G}_d(X)} - \frac{1}{F_d(X) + G_d(X)}\right)\right| > \frac{2\varepsilon}{3}\right) \\ &=: \mathbb{P}(A_1) + \mathbb{P}(A_2). \end{aligned} \tag{A.10}$$

Control of $\mathbb{P}(A_1)$. First note that

$$\begin{aligned} \mathbb{P}(A_1) &= \mathbb{P}_f\left(\big|\widehat{F}_d(X) - F_d(X)\big| > \frac{\varepsilon}{3}\big(F_d(X) + G_d(X)\big)\right) \\ &\leq \mathbb{P}_f\left(\big|\widehat{F}_d(X) - F_d(X)\big| > \frac{\varepsilon}{3} F_d(X)\right) \\ &= \mathbb{P}_f\left(\big|\widehat{F}_d(X)/F_d(X) - 1\big| > \frac{\varepsilon}{3}\right) \\ &= \mathbb{P}_f\left(\big|e^{\langle \mathbf{X}_d - (\widehat{\theta} + \boldsymbol{\theta}_d)/2, \widehat{\theta} - \boldsymbol{\theta}_d\rangle} - 1\big| > \frac{\varepsilon}{3}\right). \end{aligned} \tag{A.11}$$

Since we have $-\log(1-u) \geq \log(1+u) \geq u/2$ for $u \in (0,1)$, some straightforward computations yield:

$$\begin{aligned} \mathbb{P}(A_1) &\leq \mathbb{P}_f\left(\left|\left\langle \mathbf{X}_d - \frac{\widehat{\theta} + \boldsymbol{\theta}_d}{2}, \widehat{\theta} - \boldsymbol{\theta}_d\right\rangle\right| > \log\left(1 + \frac{\varepsilon}{3}\right)\right) \\ &\leq \mathbb{P}_f\left(\left|\left\langle \mathbf{X}_d - \frac{\widehat{\theta} + \boldsymbol{\theta}_d}{2}, \widehat{\theta} - \boldsymbol{\theta}_d\right\rangle\right| > \frac{\varepsilon}{6}\right) \end{aligned}$$

$$= \mathbb{P}_f\left(\left|\left\langle \mathbf{X}_d - \boldsymbol{\theta}_d + \frac{\boldsymbol{\theta}_d - \widehat{\theta}}{2}, \widehat{\theta} - \boldsymbol{\theta}_d \right\rangle\right| > \frac{\varepsilon}{6}\right)$$
$$\leq \mathbb{P}_f\left(\left|\langle \mathbf{X}_d - \boldsymbol{\theta}_d, \widehat{\theta} - \boldsymbol{\theta}_d\rangle\right| > \frac{\varepsilon}{6} - \frac{\|\widehat{\theta} - \boldsymbol{\theta}_d\|^2}{2}\right).$$

Now, note from (1.4), (A.2)–(3.3), and Remark A.1 (Section A.1) that, under $\mathbb{P}_{\otimes^n} \otimes \mathbb{P}_f$ and on the event $\{N_1 > 0\}$, the random variables $\xi_j := X_{d,j} - \theta_{d,j} = \langle \varphi_j, X\rangle - \theta_j$, $1 \leq j \leq d$, and

$$\zeta_j := \sqrt{N_1}(\widehat{\theta}_j - \theta_{d,j}) = \frac{1}{\sqrt{N_1}} \sum_{i=1}^{N_1} \varepsilon^1_{i,j}, \quad 1 \leq j \leq d,$$

are i.i.d. $\mathcal{N}(0,1)$ conditionally on $(Y_1, \dots, Y_n)$. (On the event $\{N_1 = 0\}$, we define the ζ_j so as to coincide with other independent $\mathcal{N}(0,1)$ random variables ζ'_j.) As a consequence, the random variables $\xi_1, \dots, \xi_d, \zeta_1, \dots, \zeta_d$ are i.i.d. $\mathcal{N}(0,1)$ (unconditionally).

Note also from Hoeffding's lemma (see, e.g., [8]) and $n/2 - \sqrt{n\log(n)/2} \geq n/4$ (because $n \geq 27$) that

$$\mathbb{P}\left(N_1 < \frac{n}{4}\right) \leq \mathbb{P}\left(N_1 < \frac{n}{2} - \sqrt{\frac{n\log n}{2}}\right) \leq \frac{1}{n}. \tag{A.12}$$

Therefore, writing $\zeta = (\zeta_1, \dots, \zeta_n)^T$, we deduce that

$$\mathbb{P}(A_1) \leq \mathbb{P}_f\left(\left|\langle \mathbf{X}_d - \boldsymbol{\theta}_d, \widehat{\theta} - \boldsymbol{\theta}_d\rangle\right| > \frac{\varepsilon}{6} - \frac{\|\widehat{\theta} - \boldsymbol{\theta}_d\|^2}{2}, N_1 \geq \frac{n}{4}\right) + \mathbb{P}\left(N_1 < \frac{n}{4}\right)$$
$$\leq \mathbb{P}\left(\left|\sum_{j=1}^d \xi_j \zeta_j\right| \geq \frac{\sqrt{N_1}\varepsilon}{6} - \frac{\|\zeta\|^2}{2\sqrt{N_1}}, N_1 \geq \frac{n}{4}\right) + \frac{1}{n}$$
$$\leq \mathbb{P}\left(\left|\sum_{j=1}^d \xi_j \zeta_j\right| \geq \frac{\sqrt{n}\varepsilon}{12} - \frac{\|\zeta\|^2}{\sqrt{n}}\right) + \frac{1}{n}$$
$$\leq \mathbb{P}\left(\left|\sum_{j=1}^d \xi_j \zeta_j\right| \geq \frac{\sqrt{n}\varepsilon}{24}\right) + \mathbb{P}\left(\|\zeta\|^2 > \frac{n\varepsilon}{24}\right) + \frac{1}{n}. \tag{A.13}$$

We control the first deviation probability above. First, recalling that the ξ_j and ζ_j are i.i.d. $\mathcal{N}(0,1)$, and conditioning by $(\xi_1, \dots, \xi_d)$, we get

$$\mathbb{P}\left(\left|\sum_{j=1}^d \xi_j \zeta_j\right| \geq \frac{\sqrt{n}\varepsilon}{24}\right) \leq \mathbb{E}\left[\mathbb{P}\left(\left|\sum_{j=1}^d \xi_j \zeta_j\right| \geq \frac{\sqrt{n}\varepsilon}{24} | \xi_1, \dots, \xi_d\right)\right]$$
$$\leq 2\mathbb{E}\left[\exp\left(-\frac{n\varepsilon^2}{1152\sum_{j=1}^d \xi_j^2}\right)\right],$$

where the last inequality is because, conditionally on $(\xi_1,\dots,\xi_d)$, the random variable $Z = \sum_{j=1}^d \xi_j \zeta_j$ is Gaussian with zero mean and variance $V = \sum_{j=1}^d \xi_j^2$ and thus satisfies $\mathbb{P}(|Z| > z|\xi_1,\dots,\xi_d) \le 2e^{-z^2/(2V)}$ for all $z > 0$. But, distinguishing whether $\sum_{j=1}^d \xi_j^2$ is below or above $4d\log n$, we obtain

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{j=1}^d \xi_j\zeta_j\right| \ge \frac{\sqrt{n}\varepsilon}{24}\right) &\le 2\exp\left(-\frac{n\varepsilon^2}{4608d\log n}\right) + 2\mathbb{P}\left(\sum_{j=1}^d \xi_j^2 > 4d\log n\right)\\ &\le 2\exp\left(-\frac{n\varepsilon^2}{4608d\log n}\right) + \frac{2}{n},\end{aligned}$$

where we used the concentration inequality for the χ^2 statistics of [26], Lemma 1,

$$\forall x > 0, \quad \mathbb{P}\left(\sum_{j=1}^d \xi_j^2 > d + 2\sqrt{dx} + 2x\right) \le e^{-x} \tag{A.14}$$

for $x = \log(n)$, and where we noted (since $2ab \le a^2 + b^2$ and $\log n \ge 2$ for $n \ge 27$) that

$$d + 2\sqrt{d\log n} + 2\log n \le 2d + 3\log n \le 4d\log n. \tag{A.15}$$

Plugging the above inequalities into (A.13), we finally obtain

$$\mathbb{P}(A_1) \le 2\exp\left(-\frac{n\varepsilon^2}{4608d\log n}\right) + \mathbb{P}\left(\|\zeta\|^2 > \frac{n\varepsilon}{24}\right) + \frac{3}{n}. \tag{A.16}$$

Control of $\mathbb{P}(A_2)$. We have:

$$\begin{aligned}\mathbb{P}(A_2) &:= \mathbb{P}_f\left(\left|\widehat{F}_d(X)\left(\frac{1}{\widehat{F}_d(X)+\widehat{G}_d(X)} - \frac{1}{F_d(X)+G_d(X)}\right)\right| > \frac{2\varepsilon}{3}\right)\\ &= \mathbb{P}_f\left(\left|\widehat{F}_d(X)\left(\frac{F_d(X)-\widehat{F}_d(X)+G_d(X)-\widehat{G}_d(X)}{(\widehat{F}_d(X)+\widehat{G}_d(X))(F_d(X)+G_d(X))}\right)\right| > \frac{2\varepsilon}{3}\right)\\ &\le \mathbb{P}_f\left(\left|\widehat{F}_d(X)\big(F_d(X)-\widehat{F}_d(X)\big)\right| > \frac{\varepsilon}{3}\big(F_d(X)+G_d(X)\big)\big(\widehat{F}_d(X)+\widehat{G}_d(X)\big)\right)\\ &\quad + \mathbb{P}_f\left(\left|\widehat{F}_d(X)\big(G_d(X)-\widehat{G}_d(X)\big)\right| > \frac{\varepsilon}{3}\big(F_d(X)+G_d(X)\big)\big(\widehat{F}_d(X)+\widehat{G}_d(X)\big)\right)\\ &\le \mathbb{P}_f\left(\left|\widehat{F}_d(X)\big(F_d(X)-\widehat{F}_d(X)\big)\right| > \frac{\varepsilon}{3}F_d(X)\widehat{F}_d(X)\right)\\ &\quad + \mathbb{P}_f\left(\left|\widehat{F}_d(X)\big(G_d(X)-\widehat{G}_d(X)\big)\right| > \frac{\varepsilon}{3}G_d(X)\widehat{F}_d(X)\right)\\ &\le \mathbb{P}_f\left(\left|F_d(X)-\widehat{F}_d(X)\right| > \frac{\varepsilon}{3}F_d(X)\right) + \mathbb{P}_f\left(\left|G_d(X)-\widehat{G}_d(X)\right| > \frac{\varepsilon}{3}G_d(X)\right).\end{aligned}$$

The first term has already been studied above (see (A.11) and the following inequalities) and thus satisfies the same upper bound as $\mathbb{P}(A_1)$ in (A.16). As for the second term, following the same lines as those leading to (A.13), we can see that

$$\begin{aligned}
&\mathbb{P}_f\left(\left|G_d(X)-\widehat{G}_d(X)\right|>\frac{\varepsilon}{3}G_d(X)\right)\\
&\quad\le \mathbb{P}_f\left(\left|\left\langle \mathbf{X}_d-\frac{\widehat{\mu}+\boldsymbol{\mu}_d}{2},\widehat{\mu}-\boldsymbol{\mu}_d\right\rangle\right|>\frac{\varepsilon}{6}\right)\\
&\quad\le \mathbb{P}_f\left(\left|\langle \mathbf{X}_d-\boldsymbol{\theta}_d,\widehat{\mu}-\boldsymbol{\mu}_d\rangle\right|>\frac{\varepsilon}{6}-\left|\left\langle \boldsymbol{\theta}_d-\boldsymbol{\mu}_d+\frac{\boldsymbol{\mu}_d-\widehat{\mu}}{2},\widehat{\mu}-\boldsymbol{\mu}_d\right\rangle\right|\right)\\
&\quad\le \mathbb{P}_f\left(\left|\langle \mathbf{X}_d-\boldsymbol{\theta}_d,\widehat{\mu}-\boldsymbol{\mu}_d\rangle\right|>\frac{\varepsilon}{6}-\underbrace{\left(\|\boldsymbol{\theta}_d-\boldsymbol{\mu}_d\|+\frac{\|\widehat{\mu}-\boldsymbol{\mu}_d\|}{2}\right)\|\widehat{\mu}-\boldsymbol{\mu}_d\|}_{\le 4(\Delta_d+2\sqrt{\frac{d\log(n)}{n}})\sqrt{\frac{d\log(n)}{n}}\text{ w.p. }\ge 1-2/n}\right)\\
&\quad\le \mathbb{P}_f\left(\left|\langle \mathbf{X}_d-\boldsymbol{\theta}_d,\widehat{\mu}-\boldsymbol{\mu}_d\rangle\right|>\frac{\varepsilon}{12}\right)+\frac{2}{n},
\end{aligned}$$

where we used (A.12) and (A.14)–(A.15) again, and where the last inequality holds true whenever

$$\left(\Delta_d+2\sqrt{\frac{d\log(n)}{n}}\right)\sqrt{\frac{d\log(n)}{n}}\le\frac{\varepsilon}{48}. \tag{A.17}$$

Mimicking what we did to derive (A.13), we then get

$$\begin{aligned}
\mathbb{P}_f\left(\left|G_d(X)-\widehat{G}_d(X)\right|>\frac{\varepsilon}{3}G_d(X)\right)&\le \mathbb{P}\left(\left|\sum_{j=1}^{d}\xi_j\zeta_j\right|\ge\frac{\sqrt{n}\varepsilon}{24}\right)+\frac{1}{n}+\frac{2}{n}\\
&\le 2\exp\left(-\frac{n\varepsilon^2}{4608d\log n}\right)+\frac{5}{n}.
\end{aligned}$$

Putting everything together, we can see that, provided (A.17) holds,

$$\mathbb{P}(A_2)\le 4\exp\left(-\frac{n\varepsilon^2}{4608d\log n}\right)+\mathbb{P}\left(\|\zeta\|^2>\frac{n\varepsilon}{24}\right)+\frac{8}{n}.$$

Conclusion. Combining all results above, we get, under condition (A.17),

$$T_1=\mathbb{P}(A_1)+\mathbb{P}(A_2)\le 6\exp\left(-\frac{n\varepsilon^2}{4608d\log n}\right)+2\mathbb{P}\left(\|\zeta\|^2>\frac{n\varepsilon}{24}\right)+\frac{11}{n}$$

so that (the upper bound on T_2 is identical by symmetry of the problem):

$$\mathbb{P}\left(\left|\widehat{\eta}_d(X)-\eta_d(X)\right|>\varepsilon\right)\le 6\exp\left(-\frac{n\varepsilon^2}{4608d\log n}\right)+2\mathbb{P}\left(\|\zeta\|^2>\frac{n\varepsilon}{24}\right)+\frac{11}{n}.$$

To conclude the proof, we note that, if (A.17) holds true, then $n\varepsilon/24 \geq 4d\log n \geq d + 2\sqrt{d\log n} + 2\log n$ (by (A.15)), so that $\mathbb{P}(\|\zeta\|^2 > n\varepsilon/24) \leq 1/n$ by (A.14). □

A.5. Proof of Theorem 3.1 (excess risk of $\widehat{\Phi}_{d_n}$)

In all the sequel, we fix $f, g \in \mathcal{H}_s(R)$ and show that

$$\mathcal{R}_{f,g}(\widehat{\Phi}_{d_n}) - \inf_{\Phi}\mathcal{R}_{f,g}(\Phi) \leq \begin{cases} cR^{\frac{1}{2s+1}}n^{-\frac{s}{2s+1}}\log(n) & \text{if } \Delta < R^{\frac{1}{2s+1}}n^{-\frac{s}{2s+1}}\log(n) \\ \dfrac{c}{\Delta}R^{\frac{2}{2s+1}}n^{-\frac{2s}{2s+1}}\log^2(n) & \text{if } \Delta \geq R^{\frac{1}{2s+1}}n^{-\frac{s}{2s+1}}\log(n) \end{cases} \tag{A.18}$$

where $\Delta := \|f-g\|$. This immediately entails the inequality of the theorem (i.e., the one involving the supremum) since the right-hand side of (A.18) is non-increasing in Δ.

Recall that $\Phi = \mathbb{1}_{\{\eta\geq 1/2\}}$ is the Bayes (optimal) classifier and that $\Phi^\star_{d_n}$ is the Bayes classifier in the d_n-dimensional truncated space (see Remark 3.1 in Section 3.3). We decompose the excess risk into estimation and approximation errors and use Lemmas 2 and 1: for some values of ε_1 and ε_2 to be determined later,

$$\begin{aligned}
&\mathcal{R}_{f,g}(\widehat{\Phi}_{d_n}) - \inf_{\Phi}\mathcal{R}_{f,g}(\Phi) \\
&= \mathcal{R}_{f,g}(\widehat{\Phi}_{d_n}) - \mathcal{R}_{f,g}\big(\Phi^\star_{d_n}\big) + \mathcal{R}_{f,g}\big(\Phi^\star_{d_n}\big) - \mathcal{R}_{f,g}\big(\Phi^\star\big) \\
&\leq 2\varepsilon_1\bigg(1 \wedge \frac{10\varepsilon_1}{\Delta_{d_n}}\bigg) + 6\exp\bigg(-\frac{n\varepsilon_1^2}{4608 d_n\log n}\bigg) + \frac{13}{n} + 12\varepsilon_2^2 + 2\varepsilon_2\bigg(1 \wedge \frac{10\varepsilon_2}{\Delta}\bigg) \\
&\leq 2\varepsilon_1\bigg(1 \wedge \frac{10\varepsilon_1}{\Delta_{d_n}}\bigg) + \frac{19}{n} + 12\varepsilon_2^2 + 2\varepsilon_2\bigg(1 \wedge \frac{10\varepsilon_2}{\Delta}\bigg),
\end{aligned} \tag{A.19}$$

where $\Delta_{d_n} := \|f_{d_n} - g_{d_n}\|$, and where we assumed that $\varepsilon_1^2 \geq 4608 d_n\log^2(n)/n$ (to be checked below).

In all the sequel the value of the constant $N_{s,R}$ may change from line to line. Our first constraint on $N_{s,R}$ is that $N_{s,R} \geq 1/R^2$, so that $d_n := \lfloor (R^2 n)^{\frac{1}{2s+1}} \rfloor \geq 1$ for all $n \geq N_{s,R}$. The choice of d_n also guarantees the bias–variance tradeoff $Rd_n^{-s} \approx \sqrt{d_n/n} \approx R^{\frac{1}{2s+1}}n^{-\frac{s}{2s+1}}$. More precisely, provided $N_{s,R}$ is chosen large enough, we get for all $n \geq N_{s,R}$ that

$$\sqrt{\frac{d_n}{n}} \leq R^{\frac{1}{2s+1}}n^{-\frac{s}{2s+1}} \leq Rd_n^{-s} \leq 2R^{\frac{1}{2s+1}}n^{-\frac{s}{2s+1}}. \tag{A.20}$$

We now choose ε_1 and ε_2 so as to minimize (A.19), while meeting the assumptions of Lemmas 2 and 1.

- We choose

$$\varepsilon_1 := 48\bigg(\Delta_{d_n} + 2\sqrt{\frac{d_n\log(n)}{n}} + \sqrt{2\log(n)}\bigg)\sqrt{\frac{d_n\log(n)}{n}}.$$

This entails that $0 < \varepsilon_1 \le 1/8$ for all $n \ge N_{s,R}$ (provided $N_{s,R}$ is chosen large enough), that Assumption (3.10) of Lemma 2 holds true, and that the requirement $\varepsilon_1^2 \ge 4608 d_n \log^2(n)/n$ above is met.

- We choose

$$\varepsilon_2 := 32 R d_n^{-s} \sqrt{\log \frac{1}{32 R d_n^{-s}}}.$$

Choosing $N_{s,R}$ large enough, we can guarantee for all $n \ge N_{s,R}$ that $0 < \varepsilon_2 \le 1/8$, as well as $\log[1/(32 R d_n^{-s})] \ge 1$ so that $\varepsilon_2 \ge 32 R d_n^{-s}$ and therefore $\varepsilon_2 \ge 32 R d_n^{-s} \sqrt{\log(1/\varepsilon_2)}$, that is,

$$R^2 d_n^{-2s} \le \frac{\varepsilon_2^2}{512 \log(1/\varepsilon_2^2)}.$$

Now, note that $\|f - f_{d_n}\|^2 \le R^2 d_n^{-2s}$ for all $f \in \mathcal{H}_s(R)$ because

$$\|f - f_{d_n}\|^2 = \sum_{k=d_n+1}^{+\infty} c_k(f)^2 \le d_n^{-2s} \sum_{k=d_n+1}^{+\infty} c_k(f)^2 k^{2s} \le R^2 d_n^{-2s}.$$

Combining the above inequalities implies that Assumption (3.8) of Lemma 1 is met.

Before plugging the values of ε_1 and ε_2 into (A.19), we compare Δ_{d_n} with Δ:

$$\begin{aligned} \Delta_{d_n} &:= \|f_{d_n} - g_{d_n}\| \\ &\ge \|f - g\| - \|f - f_{d_n}\| - \|g - g_{d_n}\| \\ &\ge \Delta - 2 R d_n^{-s} \ge \frac{\Delta}{10} \end{aligned} \tag{A.21}$$

whenever $\Delta \ge (20/9) R d_n^{-s}$. By (A.20) a sufficient condition is that $\Delta \ge (40/9) R^{\frac{1}{2s+1}} n^{-\frac{s}{2s+1}}$ or even that $\Delta \ge R^{\frac{1}{2s+1}} n^{-\frac{s}{2s+1}} \log(n)$ (provided $N_{s,R} \ge e^{40/9} \approx 85.2$). This is the threshold value we use below, since it makes the right-hand side of (A.18) continuous in Δ.

Case 1: $\Delta < R^{\frac{1}{2s+1}} n^{-\frac{s}{2s+1}} \log(n)$. .

We substitute the values of ε_1 and ε_2 into (A.19) and discard the (relatively large) terms $10\varepsilon_1/\Delta_{d_n}$ and $10\varepsilon_2/\Delta$. We obtain, noting that $12\varepsilon_2^2 \le 12\varepsilon_2/8 \le 2\varepsilon_2$:

$$\begin{aligned} &\mathcal{R}_{f,g}(\widehat{\Phi}_{d_n}) - \inf_{\Phi} \mathcal{R}_{f,g}(\Phi) \\ &\le 2\varepsilon_1 + \frac{19}{n} + 12\varepsilon_2^2 + 2\varepsilon_2 \\ &\le 2\varepsilon_1 + 4\varepsilon_2 + \frac{19}{n} \end{aligned}$$

$$
\begin{aligned}
&\le 96\left(\Delta_{d_n} + 2\sqrt{\frac{d_n \log(n)}{n}} + \sqrt{2\log(n)}\right)\sqrt{\frac{d_n \log(n)}{n}} \\
&\quad + 128 R d_n^{-s}\sqrt{\log\frac{1}{32 R d_n^{-s}}} + \frac{19}{n} \\
&\le 96\big(2R + 2R^{\frac{1}{2s+1}} n^{-\frac{s}{2s+1}}\sqrt{\log(n)} + \sqrt{2\log(n)}\big) R^{\frac{1}{2s+1}} n^{-\frac{s}{2s+1}}\sqrt{\log(n)} \\
&\quad + 256 R^{\frac{1}{2s+1}} n^{-\frac{s}{2s+1}}\sqrt{\log\frac{n^{\frac{s}{2s+1}}}{32 R^{\frac{1}{2s+1}}}} + \frac{19}{n} \\
&\le c_1 R^{\frac{1}{2s+1}} n^{-\frac{s}{2s+1}}\log(n), \qquad \text{(A.22)}
\end{aligned}
$$

where the inequality before last follows from (A.20) and from $\Delta_{d_n} \le \Delta \le \|f\| + \|g\| \le 2R$ (since $f, g \in \mathcal{H}_s(R)$), and where (A.22) holds for all $n \ge N_{s,R}$ provided the absolute constant $c_1 > 0$ and the constant $N_{s,R}$ are chosen large enough.

Case 2: $\Delta \ge R^{\frac{1}{2s+1}} n^{-\frac{s}{2s+1}}\log(n)$. .

Following similar calculations, but using now the (relatively small) terms $10\varepsilon_1/\Delta_{d_n}$ and $10\varepsilon_2/\Delta$, we can see from (A.19) and then (A.21) that, for some absolute constants $c_2, c_3 > 0$,

$$
\begin{aligned}
&\mathcal{R}_{f,g}(\widehat{\Phi}_{d_n}) - \inf_{\Phi}\mathcal{R}_{f,g}(\Phi) \\
&\le \frac{20\varepsilon_1^2}{\Delta_{d_n}} + \frac{19}{n} + 12\varepsilon_2^2 + \frac{20\varepsilon_2^2}{\Delta} \\
&\le \frac{200\varepsilon_1^2}{\Delta} + \frac{20\varepsilon_2^2}{\Delta} + 12\varepsilon_2^2 + \frac{19}{n} \\
&\le c_2 R^{\frac{2}{2s+1}} n^{-\frac{2s}{2s+1}}\left(\frac{\log^2(n)}{\Delta} + \frac{\log(n)}{\Delta} + \log(n)\right) + \frac{19}{n} \\
&\le \frac{c_3 R^{\frac{2}{2s+1}} n^{-\frac{2s}{2s+1}}\log^2(n)}{\Delta}, \qquad \text{(A.23)}
\end{aligned}
$$

where the last two inequalities hold true for all $n \ge N_{s,R}$ provided $N_{s,R}$ is chosen large enough (e.g., $\log^2(n)/\Delta \ge \log(n)$ when $n \ge e^{2R} \ge e^{\Delta}$).

Conclusion. We derive (A.18) by combining (A.22) and (A.23) and by choosing $c := \max\{c_1, c_3\}$. This concludes the proof of Theorem 3.1.

Supplementary Material

Optimal functional supervised classification with separation condition (DOI: 10.3150/19-BEJ1170SUPP; .pdf). In the supplemental article [17], we provide a proof of the minimax lower bound (Theorem 4.1) and a discussion on the truncated nearest neighbor strategy (Theorem 4.2).

References

[1] Abraham, C., Biau, G. and Cadre, B. (2006). On the kernel rule for function classification. *Ann. Inst. Statist. Math.* **58** 619–633. MR2327897 https://doi.org/10.1007/s10463-006-0032-1

[2] Audibert, J.-Y. and Tsybakov, A.B. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.* **35** 608–633. MR2336861 https://doi.org/10.1214/009053606000001217

[3] Baíllo, A., Cuevas, A. and Cuesta-Albertos, J.A. (2011). Supervised classification for a family of Gaussian functional models. *Scand. J. Stat.* **38** 480–498. MR2833842 https://doi.org/10.1111/j.1467-9469.2011.00734.x

[4] Biau, G. and Devroye, L. (2015). *Lectures on the Nearest Neighbor Method. Springer Series in the Data Sciences*. Cham: Springer. MR3445317 https://doi.org/10.1007/978-3-319-25388-6

[5] Biau, G. and Scornet, E. (2016). A random forest guided tour. *TEST* **25** 197–227. MR3493512 https://doi.org/10.1007/s11749-016-0481-7

[6] Bickel, P.J. and Levina, E. (2004). Some theory of Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* **10** 989–1010. MR2108040 https://doi.org/10.3150/bj/1106314847

[7] Boucheron, S., Bousquet, O. and Lugosi, G. (2005). Theory of classification: A survey of some recent advances. *ESAIM Probab. Stat.* **9** 323–375. MR2182250 https://doi.org/10.1051/ps:2005018

[8] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities*. Oxford: Oxford Univ. Press. A nonasymptotic theory of independence, With a foreword by Michel Ledoux. MR3185193 https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

[9] Cadre, B. (2013). Supervised classification of diffusion paths. *Math. Methods Statist.* **22** 213–225. MR3107669 https://doi.org/10.3103/S1066530713030034

[10] Cai, T.T. and Zhang, L. (2019). High dimensional linear discriminant analysis: Optimality, adaptive algorithm and missing data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 675–705. MR3997097

[11] Cérou, F. and Guyader, A. (2006). Nearest neighbor classification in infinite dimension. *ESAIM Probab. Stat.* **10** 340–355. MR2247925 https://doi.org/10.1051/ps:2006014

[12] Chaudhuri, K. and Dasgupta, S. (2014). Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems* (Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence and K.Q. Weinberger, eds.) **27** 3437–3445. Curran Associates.

[13] Chonavel, T. (2002). *Statistical Signal Processing*. New-York: Springer.

[14] Cover, T.M. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory* **13** 21–27.

[15] 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 https://doi.org/10.1111/j.1467-9868.2011.01003.x

[16] Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition. Applications of Mathematics (New York)* **31**. New York: Springer. MR1383093 https://doi.org/10.1007/978-1-4612-0711-5

[17] Gadat, S., Gerchinovitz, S. and Marteau, C. (2020). Supplement to "Optimal functional supervised classification with separation condition." https://doi.org/10.3150/19-BEJ1170SUPP.

[18] Gadat, S., Klein, T. and Marteau, C. (2016). Classification in general finite dimensional spaces with the k-nearest neighbor rule. *Ann. Statist.* **44** 982–1009. MR3485951 https://doi.org/10.1214/15-AOS1395

[19] Győrfi, L. (1978). On the rate of convergence of nearest neighbor rules. *IEEE Trans. Inform. Theory* **24** 509–512. MR0501595 https://doi.org/10.1109/TIT.1978.1055898

[20] Ibragimov, I. and Khasminskii, R. (1981). *Statistical Estimation: Asymptotic Theory*. New York: Springer.

[21] Ikeda, N. and Watanabe, S. (1989). *Stochastic Differential Equations and Diffusion Processes*, 2nd ed. *North-Holland Mathematical Library* **24**. Amsterdam: North-Holland. MR1011252

[22] James, G.M. and Hastie, T.J. (2001). Functional linear discriminant analysis for irregularly sampled curves. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 533–550. MR1858401 https://doi.org/10.1111/1467-9868.00297

[23] Kulkarni, S.R. and Posner, S.E. (1995). Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Trans. Inform. Theory* **41** 1028–1039. MR1366756 https://doi.org/10.1109/18.391248

[24] Lamberton, D. and Lapeyre, B. (1996). *Introduction to Stochastic Calculus Applied to Finance*. London: CRC Press. MR1422250

[25] Lande, R., Engen, S. and Saether (2003). *Stochastic Populations Dynamics in Ecology and Conservation*. New-York: Oxford Univ. Press Inc.

[26] Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.* **28** 1302–1338. MR1805785 https://doi.org/10.1214/aos/1015957395

[27] Lepskiĭ, O.V. (1990). A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatn. Primen.* **35** 459–470. MR1091202 https://doi.org/10.1137/1135065

[28] Li, T., Yi, X., Carmanis, X. and Ravikumar, P. (2017). Minimax Gaussian classification & clustering. In *Proceedings of the* 20*th International Conference on Artificial Intelligence and Statistics*. *Proceedings of Machine Learning Research* **54** 1–9.

[29] Mammen, E. and Tsybakov, A.B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27** 1808–1829. MR1765618 https://doi.org/10.1214/aos/1017939240

[30] Massart, P. and Nédélec, É. (2006). Risk bounds for statistical learning. *Ann. Statist.* **34** 2326–2366. MR2291502 https://doi.org/10.1214/009053606000000786

[31] Rakhlin, A., Sridharan, K. and Tsybakov, A.B. (2017). Empirical entropy, minimax regret and minimax risk. *Bernoulli* **23** 789–824. MR3606751 https://doi.org/10.3150/14-BEJ679

[32] Rossi, F. and Villa, N. (2008). Recent advances in the use of SVM for functional data classification. In *Functional and Operatorial Statistics*. *Contrib. Statist.* 273–280. Heidelberg: Physica-Verlag/Springer. MR2490360 https://doi.org/10.1007/978-3-7908-2062-1_41

[33] Samworth, R.J. (2012). Optimal weighted nearest neighbour classifiers. *Ann. Statist.* **40** 2733–2763. MR3097618 https://doi.org/10.1214/12-AOS1049

[34] Shao, J., Wang, Y., Deng, X. and Wang, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *Ann. Statist.* **39** 1241–1265. MR2816353 https://doi.org/10.1214/10-AOS870

[35] Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. *Information Science and Statistics*. New York: Springer. MR2450103

[36] Wang, J.L., Chiou, J.M. and Müller, H.G. (2016). Functional data analysis. *Annu. Rev. Stat. Appl.* **3** 257–295.

Received March 2018 and revised October 2019