Optimal functional supervised classification with separation condition

Sébastien Gadat, Sébastien Gerchinovitz, and Clément Marteau

January 8, 2018

Abstract

We consider the binary supervised classification problem with the Gaussian functional model introduced in [7]. 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 [9] and [14] we also derive a logarithmic lower bound showing that the popular k-nearest neighbors classifier is far from optimality in this specific functional setting.

1 Introduction

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 [13] and [5] 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 $(X_1, Y_1), \ldots, (X_n, Y_n)$ 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 $\mathcal{S}_n := (X_i, Y_i)_{1 \leqslant i \leqslant n}$ 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 [27, 2], the k-nearest neighbors algorithm [16, 9, 3, 14], support vector machines [32], or random forests [4], just to name a few.

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, e.g., in stochastic population dynamics [24], in signal processing [10], or in finance [23]. This binary supervised functional classification problem was tackled with nonparametric procedures such as kernel methods or the k-nearest neighbours algorithm. For example, [11] studied the nearest neighbour rule in any metric space, while [22] analyzed the performances of the k-nearest neighbours algorithm in terms of a metric covering measure. Such metric entropy arguments were also used in [8], or with kernel methods in [1].

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

$$\forall t \in [0,1], \qquad dX(t) = Yf(t)dt + (1-Y)g(t)dt + dW(t), \qquad (1)$$

where $(W(t))_{0 \leqslant t \leqslant 1}$ is a standard Brownian motion, and where Y is a Bernoulli $\mathcal{B}(1/2)$ random variable independent from $(W(t))_{0 \leqslant t \leqslant 1}$. 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.

1

The white noise model has played a key role in statistical theoretical developments; see, e.g., the seminal contributions of [17] in nonparametric estimation and of [26] 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 $(X(t))_{0 \leqslant t \leqslant 1}$. 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 [7]. Actually our setting is less general since [7] 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 [7], as indicated below.

Some notation. 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:

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

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*

$$\Phi^\star = \arg\min_{\Phi} \mathcal{R}_{f,g}(\Phi), \tag{2}$$

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*

$$\sup_{(f,g) \in \mathcal{E}} \left\{ \mathcal{R}_{f,g}(\Phi) - \inf_{\Phi} \mathcal{R}_{f,g}(\Phi) \right\} \tag{3}$$

over some set $\mathcal{E}$ of pairs of functions. In the sequel we focus on Sobolev classes $\mathcal{H}_s(R)$ (see (21)) and consider subsets $\mathcal{E} \subseteq \mathcal{H}_s(R)^2$ parametrized by a separation lower bound Δ on $\|f - g\|$.

Main contributions and outline of the paper. 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. We derive an excess risk bound that generalizes the worst-case results of [7] but also imply faster rates when the distance $\|f - g\|$ is large. This acceleration is a consequence of the nice properties of the margin (see also, e.g., [2] and [14]).

  Theorem (A). *The classifier $\widehat{\Phi}_{d_n}$ defined in (16) 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 $n \geqslant N_{R,s}$ large enough,*

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

- 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 $n \geqslant N_{R,s}$ of observations, any classifier $\widehat{\Phi}$ must satisfy (omitting again logarithmic factors and constant factors depending only on s and R):*

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

2

- 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., [31, 14]) 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 $k_n^{opt}(\widehat{d}) = \lfloor n^{4/(4+\widehat{d})} \rfloor$, the $\widehat{d}$-dimensional $k_n^{opt}(\widehat{d})$-nearest neighbors classifier Φ_{NN} suffers a logarithmic excess risk in the worst case:*

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

Most proofs are postponed to Appendix A (for the upper bounds) and to Appendix B (for the lower bounds).

Other useful notation. 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 $(X_i, Y_i)_{1\leqslant i\leqslant n}$. For notational convenience, the measure $\mathbb{P}$ will alternatively stand for $\mathbb{P} = \mathbb{P}_{X,Y} \times \mathbb{P}_{\otimes n}$ (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. $\mathbb{P}_g$).

We write $\mathbb{L}^2([0,1])$ for the set of square Lebesgue-integrable functions f on $[0,1]$, with $\mathbb{L}^2$-norm $\|f\| = \left(\int_0^1 f^2(t)dt \right)^{1/2}$ 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) and $\varphi \in \mathbb{L}^2([0,1])$, we set $\langle \varphi, X \rangle = \int_0^1 \varphi(t)dX(t)$.

Finally we write $\mathrm{Ber}(p)$ or simply $\mathcal{B}(p)$ 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 (2). This optimal classifier is known as *the Bayes classifier* of the classification problem (see, e.g., [16, 13]).

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., [19]). 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

$$q_f(X) := \frac{d\mathbb{P}_f}{d\mathbb{P}_0}(X) = \exp\left(\int_0^1 f(s)dX_s - \frac{1}{2}\|f\|^2 \right) . \tag{4}$$

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

$$q_g(X) := \frac{d\mathbb{P}_g}{d\mathbb{P}_0}(X) = \exp\left(\int_0^1 g(s)dX_s - \frac{1}{2}\|g\|^2 \right) . \tag{5}$$

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) is given by

$$
\begin{aligned}
\eta(X) &:= \mathbb{E}[Y|X] = \mathbb{P}(Y = 1|X) \\
&= \frac{\frac{d\mathbb{P}_f}{d\mathbb{P}_0}(X)}{\frac{d\mathbb{P}_f}{d\mathbb{P}_0}(X) + \frac{d\mathbb{P}_g}{d\mathbb{P}_0}(X)} \\
&= \frac{\exp\left(\int_0^1 (f(s) - g(s))dX_s - \frac{1}{2}\|f\|^2 + \frac{1}{2}\|g\|^2 \right)}{1 + \exp\left(\int_0^1 (f(s) - g(s))dX_s - \frac{1}{2}\|f\|^2 + \frac{1}{2}\|g\|^2 \right)} .
\end{aligned}
\tag{6}
$$

As an example, if we assume that we observe $dX(t) = f(t)dt$ with $X(0) = 0$, then $\eta(X) = \frac{\exp\left(\frac{1}{2}\|f-g\|^2\right)}{1+\exp\left(\frac{1}{2}\|f-g\|^2\right)}$, 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).

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

$$\Phi^\star(X) := \mathbb{1}_{\left\{\eta(X) \geqslant \frac{1}{2}\right\}} = \mathbb{1}_{\left\{\int_0^1 (f(s)-g(s))dX_s \geqslant \frac{1}{2}\|f\|^2 - \frac{1}{2}\|g\|^2\right\}} . \tag{7}$$

It is well known that the Bayes classifier $\Phi^\star$ corresponds to the optimal classifier of the considered binary classification problem (see, e.g., [13]) in the sense that it satisfies (2). In particular, for any other classifier Φ, the excess risk of classification is given by

$$\mathcal{R}_{f,g}(\Phi) - \mathcal{R}_{f,g}(\Phi^\star) = \mathbb{E}\left[|2\eta(X) - 1| \mathbb{1}_{\Phi(X) \neq \Phi^\star(X)}\right] . \tag{8}$$

In our statistical setting, the functions f and g are unknown so that it is impossible to compute the oracle Bayes classifier (7). However, we can construct an approximation of it using the sample $(X_i, Y_i)_{1 \leqslant i \leqslant n}$. 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 will be a key ingredient of our analysis.

2.2 Control of the margin in the functional model

As was shown in earlier works on binary supervised classification (see, e.g., [27] 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 $\alpha \geqslant 0$ and $\varepsilon_0, C > 0$ such that, for all $0 < \varepsilon \leqslant \varepsilon_0$,

$$\mathbb{P}_X\left(\left|\eta(X) - \frac{1}{2}\right| \leqslant \varepsilon\right) \leqslant C\varepsilon^\alpha . \tag{9}$$

We will show in Proposition 1 which parameters $\alpha, \varepsilon_0, C > 0$ are associated with Model (1). The role of (9) 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 $\mathbb{1}_{\eta(X) \geqslant 1/2}$ 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 $\mathbb{1}_{\eta(X) \geqslant 1/2}$. 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) crucially depends on the distance between the functions f and g of interest. The proof is postponed to Appendix A.1.

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

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

In particular, if the distance $\|f - g\|$ is bounded from below by a positive constant, then (9) 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 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 3.1.3 below—involves a projection step with coefficients θ_j and μ_j introduced in Section 3.1.1 and estimated in Section 3.1.2.

3.1.1 $\mathbb{L}^2$ orthonormal basis $(\varphi_j)_{j \geqslant 1}$ and white noise model

Orthonormal basis. Let $(\varphi_j)_{j \in \mathbb{N}^*}$ be any orthonormal basis of $\mathbb{L}^2([0,1])$, and $d \in \mathbb{N}^*$ be some dimension that will be chosen as a function of the size n of the sample (for projection purposes). In the sequel, the coefficients $(c_j(h))_{j \geqslant 1}$ of any function $h \in \mathbb{L}^2([0,1])$ w.r.t. the basis $(\varphi_j)_{j \geqslant 1}$ are defined by

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

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

$$\Pi_d(h) = \sum_{j=1}^d c_j(h)\varphi_j. \tag{10}$$

In particular, we will pay a specific attention to the coefficients of f and g involved in (1),

$$\forall j \geqslant 1, \qquad \theta_j := c_j(f) \qquad \text{and} \qquad \mu_j := c_j(g), \tag{11}$$

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

White noise model. We now make a few comments on the white noise model considered in (1) (see also [17] for further details and its link with the infinite Gaussian sequence model). First note that, for all $\varphi \in \mathbb{L}^2([0,1])$, almost surely,

$$\int_0^1 \varphi(t)dX(t) = Y \int_0^1 f(t)\varphi(t)dt + (1-Y) \int_0^1 g(t)\varphi(t)dt + \int_0^1 \varphi(t)dW(t).$$

Recall that, with a slight abuse of notation, we write $\langle \varphi, X \rangle := \int_0^1 \varphi(t)dX(t)$. 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, the distribution of $\langle \varphi, X \rangle$ is a mixture of two Gaussian distributions:

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

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

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

The above remarks imply together with (11) that

$$\forall j \geqslant 1, \qquad \mathcal{L}\left(\langle \varphi_j, X \rangle | Y = 1\right) = \mathcal{N}(\theta_j, 1) \qquad \text{and} \qquad \mathcal{L}\left(\langle \varphi_j, X \rangle | Y = 0\right) = \mathcal{N}(\mu_j, 1), \tag{12}$$

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

3.1.2 Estimation of $(\theta_j)_{1 \leqslant j \leqslant d}$ and $(\mu_j)_{1 \leqslant j \leqslant d}$

In order to estimate the θ_j and μ_j, we split the sample $(X_i, Y_i)_{1 \leqslant i \leqslant n}$ into two subsamples corresponding to either $Y_i = 0$ or $Y_i = 1$. More formally, we define the two subsamples $(X_i^0)_{1 \leqslant i \leqslant N_0}$ and $(X_i^1)_{1 \leqslant i \leqslant N_1}$ (one of which can be empty) by

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

where

$$N_0 := \sum_{i=1}^{n} \mathbb{1}_{\{Y_i=0\}} \quad \text{and} \quad N_1 := \sum_{i=1}^{n} \mathbb{1}_{\{Y_i=1\}}, \tag{13}$$

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

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

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, both subsamples have (with high probability) approximately the same sizes.

Note from (1) that the two subsamples $(X_i^0)_{1 \leqslant i \leqslant N_0}$ and $(X_i^1)_{1 \leqslant i \leqslant N_1}$ correspond to observations of the functions g and f respectively. Following our comments from Section 3.1.1, it is natural to define the random coefficients $(X_{i,j}^0)_{\substack{1 \leqslant i \leqslant N_0 \\ 1 \leqslant j \leqslant d}}$ and $(X_{i,j}^1)_{\substack{1 \leqslant i \leqslant N_1 \\ 1 \leqslant j \leqslant d}}$ by

$$\begin{cases} X_{i,j}^0 := \langle \varphi_j, X_i^0 \rangle = \mu_j + \varepsilon_{i,j}^0, & i = 1 \ldots N_0, \\ X_{i,j}^1 := \langle \varphi_j, X_i^1 \rangle = \theta_j + \varepsilon_{i,j}^1, & i = 1 \ldots N_1, \end{cases} \quad j \in \{1, \ldots, d\}, \tag{14}$$

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

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

By independence of the random variables $Y_1, W_1, \ldots, Y_n, W_n$ used to generate the sample $(X_i, Y_i)_{1 \leqslant i \leqslant n}$ according to (1), and by the comments made in Section 3.1.1, we have the following conditional independence property for the $\varepsilon_{i,j}^k$.

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

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

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

For every $j \in \{1, \ldots, d\}$, we use the coefficients defined in (14) to estimate the coefficients μ_j and θ_j by

$$\widehat{\mu}_j = \mathbb{1}_{\{N_0>0\}} \frac{1}{N_0} \sum_{i=1}^{N_0} X_{i,j}^0 \quad \text{and} \quad \widehat{\theta}_j = \mathbb{1}_{\{N_1>0\}} \frac{1}{N_1} \sum_{i=1}^{N_1} X_{i,j}^1. \tag{15}$$

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.

3.1.3 A simple classifier

We now build a simple classifier using the estimators $\widehat{\mu}_j$ and $\widehat{\theta}_j$ defined in (15). After observing a new trajectory $X = (X(t))_{0 \leqslant t \leqslant 1}$, we construct the vector $\mathbf{X}_d \in \mathbb{R}^d$ defined by

$$\mathbf{X}_d := \left(\langle \varphi_1, X \rangle, \ldots, \langle \varphi_d, X \rangle \right).$$

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

$$\widehat{\Phi}_d(X) = \begin{cases} 1 & \text{if} \quad \|\mathbf{X}_d - \widehat{\theta}\|_d \leqslant \|\mathbf{X}_d - \widehat{\mu}\|_d \\ 0 & \text{if} \quad \|\mathbf{X}_d - \widehat{\theta}\|_d > \|\mathbf{X}_d - \widehat{\mu}\|_d \end{cases}, \tag{16}$$

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.

6

Reinterpretation as a plug-in classifier We now explain why $\widehat{\Phi}_d$ can be reinterpreted as a plug-in classifier in a truncated space. Recall the expression (6) for the regression function η. It is thus natural to 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 $\Pi_d(g) = \sum_{j=1}^d \mu_j \varphi_j$, i.e.,

$$
\begin{aligned}
\eta_d(X) &:= \frac{\exp\left(\int_0^1 (\Pi_d(f)(s) - \Pi_d(g)(s))dX_s - \frac{1}{2}\|\Pi_d(f)\|^2 + \frac{1}{2}\|\Pi_d(g)\|^2\right)}{1 + \exp\left(\int_0^1 (\Pi_d(f)(s) - \Pi_d(g)(s))dX_s - \frac{1}{2}\|\Pi_d(f)\|^2 + \frac{1}{2}\|\Pi_d(g)\|^2\right)} \\
&= \frac{\exp\left(\langle \boldsymbol{\theta}_d - \boldsymbol{\mu}_d, \mathbf{X}_d \rangle_d - \frac{1}{2}\|\boldsymbol{\theta}_d\|_d^2 + \frac{1}{2}\|\boldsymbol{\mu}_d\|_d^2\right)}{1 + \exp\left(\langle \boldsymbol{\theta}_d - \boldsymbol{\mu}_d, \mathbf{X}_d \rangle_d - \frac{1}{2}\|\boldsymbol{\theta}_d\|_d^2 + \frac{1}{2}\|\boldsymbol{\mu}_d\|_d^2\right)},
\end{aligned}
\tag{17}
$$

where $\boldsymbol{\theta}_d := (\theta_j)_{1\leqslant j\leqslant d}$, and $\boldsymbol{\mu}_d := (\mu_j)_{1\leqslant j\leqslant d}$. We also define the associated oracle classifier (that still depends on the unknown functions f and g):

$$
\Phi_d^\star(X) := \mathbb{1}_{\{\eta_d(X) \geqslant \frac{1}{2}\}} = \mathbb{1}_{\{\int_0^1 (\Pi_d(f)(s) - \Pi_d(g)(s))dX_s \geqslant \frac{1}{2}\|\Pi_d(f)\|^2 - \frac{1}{2}\|\Pi_d(g)\|^2\}}.
\tag{18}
$$

As shown in Remark 2 (see Section 3.3 below), η_d and $\Phi_d^\star$ correspond to the regression function and the Bayes classifier of the restricted classification problem where the learner has only access to the projected input $\mathbf{X}_d \in \mathbb{R}^d$, rather than the whole trajectory X.

We are now ready to reinterpret $\widehat{\Phi}_d$ as a plug-in classifier. Note that

$$
\|\mathbf{X}_d - \widehat{\theta}\|_d \leqslant \|\mathbf{X}_d - \widehat{\mu}\|_d, \quad \Longleftrightarrow \quad \frac{\exp\left(\frac{1}{2}\{\|\mathbf{X}_d - \widehat{\mu}\|_d^2 - \|\mathbf{X}_d - \widehat{\theta}\|_d^2\}\right)}{1 + \exp\left(\frac{1}{2}\{\|\mathbf{X}_d - \widehat{\mu}\|_d^2 - \|\mathbf{X}_d - \widehat{\theta}\|_d^2\}\right)} \geqslant \frac{1}{2},
$$

$$
\Longleftrightarrow \quad \widehat{\eta}_d(X) \geqslant \frac{1}{2},
$$

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

$$
\widehat{\eta}_d(X) := \frac{\exp\left(\langle \widehat{\theta} - \widehat{\mu}, \mathbf{X}_d \rangle_d - \frac{1}{2}\|\widehat{\theta}\|_d^2 + \frac{1}{2}\|\widehat{\mu}\|_d^2\right)}{1 + \exp\left(\langle \widehat{\theta} - \widehat{\mu}, \mathbf{X}_d \rangle_d - \frac{1}{2}\|\widehat{\theta}\|_d^2 + \frac{1}{2}\|\widehat{\mu}\|_d^2\right)}.
\tag{19}
$$

In other words, our classifier $\widehat{\Phi}_d$ can be rewritten as $\widehat{\Phi}_d(X) = \mathbb{1}_{\{\widehat{\eta}_d(X) \geqslant 1/2\}}$ where $\widehat{\eta}_d$ is an estimator of the 'truncated' regression function η_d introduced in (17).

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, (16), and (18)):

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

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 trunctated 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 (18) and (7) respectively. Comparing the definitions of η and η_d in (6) and (17), 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.2. We recall that, for notational convenience, we write $f_d = \Pi_d(f)$ and $g_d = \Pi_d(g)$.

7

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

$$\max\left(\|f - f_d\|^2, \|g - g_d\|^2\right) \leqslant \frac{\varepsilon^2}{512 \ln(1/\varepsilon^2)} \, . \tag{20}$$

Then, the two oracle classifiers $\Phi_d^\star$ and $\Phi^\star$ defined by (18) and (7) satisfy

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

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 are smooth in that their (Fourier) coefficients w.r.t. the basis $(\varphi_j)_{j \geqslant 1}$ decay sufficiently fast. More precisely, we assume that, for some parameters $s, R > 0$, the functions f and g belong to the set

$$\mathcal{H}_s(R) := \left\{ h \in \mathbb{L}^2([0,1]) : \sum_{j=1}^{+\infty} c_j(h)^2 j^{2s} \leqslant R^2 \right\} \, . \tag{21}$$

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 $\|f'\|_2 \leqslant R$.

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

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

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 (8)) in the truncated space. The next remark follows from direct calculations.

Remark 2. *Denote by $\mathbf{X}_d := \left(\langle \varphi_j, X \rangle\right)_{1 \leqslant j \leqslant d}$, $\boldsymbol{\theta}_d = (\theta_j)_{1 \leqslant j \leqslant d}$, and $\boldsymbol{\mu}_d = (\mu_j)_{1 \leqslant j \leqslant d}$ the versions of X, θ, and μ in the truncated space. Then,*

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

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 $\Phi_d^* = \mathbb{1}_{\{\eta_d \geqslant 1/2\}}$ 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.3. (The value of 4608 could most probably be improved.) We recall that $f_d = \Pi_d(f)$ and $g_d = \Pi_d(g)$.

8

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

$$\left(\Delta_d + 2\sqrt{\frac{d \log(n)}{n}} \right) \sqrt{\frac{d \log(n)}{n}} \leqslant \frac{\varepsilon}{48} \; . \tag{22}$$

Then, the classifiers $\widehat{\Phi}_d$ and $\Phi_d^\star$ defined by (16) and (18) satisfy

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

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 (21). 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 1. *There exist an absolute constant $c > 0$ and a constant $N_{s,R} \geqslant 86$ depending only on s and R such that the following holds true. For all $s, R > 0$ and all $n \geqslant N_{s,R}$, the classifier $\widehat{\Phi}_{d_n}$ defined by (16) with $d_n = \lfloor (R^2 n)^{\frac{1}{2s+1}} \rfloor$ satisfies*

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

The proof is postponed to Appendix A.4 and combines Lemmas 1 and 2 from the previous sections. A proof sketch is also provided below.

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 (Novelty of the bound)**.**

- *Taking $\Delta = 0$, we recover the worst-case bound of [7, Corollary 4.4] (where $u = 1/s$) up to logarithmic factors. As shown by Theorem 2 below, this slow rate is unimprovable for a small distance $\|f - g\|$.*

- *However, we obtain a significantly faster rate when $\|f - g\|$ is bounded from below. This improved rate is a consequence of the margin behavior (see, e.g., [2, 14]), but not of the choice of d_n that is oblivious to $\|f - g\|$.*

- *Continuous transitions from slow rates to faster rates were already derived in the past. For instance, for any supervised classification problem where the margin $|2\eta(X) - 1| \geqslant h$ is almost surely bounded from below, [29, 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, [30, 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.*

9

We finally note 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., [26]) after adapting it to the classification setting.

Sketch of the proof. We outline the main ingredients. For all $d \in \mathbb{N}^*$ and $0 < \varepsilon_1, \varepsilon_2 \leqslant 1/8$ satisfying the assumptions of Lemmas 2 and 1,

$$
\begin{aligned}
\mathcal{R}_{f,g}(\widehat{\Phi}_d) - \mathcal{R}_{f,g}(\Phi^\star) &= \mathcal{R}_{f,g}(\widehat{\Phi}_d) - \mathcal{R}_{f,g}(\Phi_d^\star) + \mathcal{R}_{f,g}(\Phi_d^\star) - \mathcal{R}_{f,g}(\Phi^\star) \\
&\lesssim \varepsilon_1 \left(1 \wedge \frac{\varepsilon_1}{\|f_d - g_d\|}\right) + \exp\left(-\frac{n\varepsilon_1^2}{4608\, d \log n}\right) + \frac{1}{n} + \varepsilon_2^2 + \varepsilon_2 \left(1 \wedge \frac{\varepsilon_2}{\|f - g\|}\right) \\
&\lesssim \min\left\{\varepsilon_1 + \varepsilon_2, \frac{\varepsilon_1^2 + \varepsilon_2^2}{\|f - g\|}\right\} + \exp\left(-\frac{n\varepsilon_1^2}{4608\, d \log n}\right) + \frac{1}{n} + \varepsilon_2^2 \\
&\lesssim \min\left\{\sqrt{\frac{d}{n}}\log(n) + \varepsilon_2, \frac{d\log^2(n)/n + \varepsilon_2^2}{\|f - g\|}\right\} + \frac{1}{n} + \varepsilon_2^2
\end{aligned}
$$

for d large enough, and for the choice of $\varepsilon_1 \approx \sqrt{d/n}\log(n)$. Following the comments at the end of Section 3.2, we also choose $\varepsilon_2 \approx Rd^{-s}$ (up to logarithmic factors). We obtain, omitting constant but also logarithmic factors:

$$
\mathcal{R}_{f,g}(\widehat{\Phi}_d) - \mathcal{R}_{f,g}(\Phi^\star) \lesssim \min\left\{\sqrt{\frac{d}{n}} + d^{-s}, \frac{d/n + d^{-2s}}{\|f - g\|}\right\} \lesssim \min\left\{n^{-\frac{s}{2s+1}}, \frac{n^{-\frac{2s}{2s+1}}}{\|f - g\|}\right\},
$$

where the last inequality is obtained with the optimal value of $d \approx n^{\frac{1}{2s+1}}$. A detailed and more formal proof (with, e.g., a rigorous treatment of the approximation $\|f_d - g_d\| \approx \|f - g\|$) can be found in Appendix A.4. $\qquad\square$

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 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 $\mathbf{X}_d \in \mathbb{R}^d$. 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 1 is minimax optimal up to logarithmic factors. The proof is postponed to Appendix B.

Theorem 2. *Consider the statistical model* (1) *and the set* $\mathcal{H}_s(R)$ *defined in* (21), *where* $s, R > 0$ *and where* $(\varphi_j)_{j \geqslant 1}$ *is any Hilbert basis of* $\mathbb{L}^2([0,1])$. *Then, every classifier* $\widehat{\Phi}$ *satisfies, for any number* $n \geqslant \max\{R^{1/s}, (32\log(2)+2)^{2s+1}/(3R^2/4)\}$ *of i.i.d. observations* $(X_i, Y_i)_{1 \leqslant i \leqslant n}$ *from* (1) *and all* $\Delta \in (0, R/2]$,

$$
\sup_{\substack{f,g \in \mathcal{H}_s(R) \\ \|f-g\| \geqslant \Delta}} \left\{\mathcal{R}_{f,g}(\widehat{\Phi}) - \inf_{\Phi} \mathcal{R}_{f,g}(\Phi)\right\} \geqslant
\begin{cases}
ce^{-2R^{2/(2s+1)}} R^{1/(2s+1)}\, n^{-s/(2s+1)} & \text{if } \Delta < R^{1/(2s+1)}\, n^{-s/(2s+1)} \\[2ex]
\dfrac{ce^{-2\Delta^2}}{\Delta}\, R^{2/(2s+1)}\, n^{-2s/(2s+1)} & \text{if } \Delta \geqslant R^{1/(2s+1)}\, n^{-s/(2s+1)}
\end{cases}
$$

for some absolute constant $c > 0$.

We note two minor differences between the upper and lower bounds: Theorem 1 involves extra logarithmic factors, while Theorem 2 involves an extra term of $e^{-2\Delta^2}$. Fortunately both terms have a minor influence (note that $e^{-2\Delta^2} \geqslant e^{-8R^2}$ 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 1 and 2 together imply that, for $n \geqslant N_{s,R}$ 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 $\mathcal{S} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, a number of neighbors k, a norm $\|.\|$ and a new incoming observation, the kNN classifier is defined as

$$\Phi_{n,k}(X) = \mathbb{1}_{\{\frac{1}{k}\sum_{j=1}^{k} Y_{(j)}(X) > 1/2\}}, \tag{23}$$

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

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

We refer the reader, e.g., to [16], [13] or [3] for more details.

We are interested below in the performances of the classifier $\Phi_{n,k}$ in this functionnal setting. For this purpose, we will use the recent contribution of [9] 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 α We shall consider first a finite d-dimensional case for our Gaussian translation model. In that case, Remark 2 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. $\{Y = 1\}$), 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:

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

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

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

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 [9], the rate of convergence of the kNN depends on a smoothness parameter α involved in the next inequality:

$$\forall x \in \mathbb{R}^d \qquad |\eta(B(x,r)) - \eta(x)| \leqslant L\mu(B(x,r))^\beta \tag{24}$$

where $\eta(B(x,r))$ refers to the mean value of η on $B(x,r)$ w.r.t. the distribution of the design X given by $\mu = \frac{1}{2}\gamma(.) + \frac{1}{2}\gamma(. - m)$. 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 Appendix C.1.

---

[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 6 (Appendix 6).

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

$$\forall r \leqslant \frac{1}{R} \qquad |\eta(B(x,r)) - \eta(x)| \leqslant L_R \mu(B(x,r))^{2/d}.$$

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

$$\forall r \leqslant 1 \quad \forall x \in B(0, R/2) \qquad |\eta(B(x,r)) - \eta(x)| \leqslant 60\pi ed R^2 e^{R^2/d} \mu(B(x,r))^{2/d},$$

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

$$L_R = 60\pi ed R^2 e^{R^2/d} \tag{25}$$

According to inequality (24) 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 [9] 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., [13] 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 [9] that, when the marginal law of X is compactly supported, the optimal calibration k_n^{opt} is:

$$\frac{1}{\sqrt{k_n^{opt}(d)}} = c \left(\frac{k_n^{opt}(d)}{n} \right)^{\beta_d} \quad \Leftrightarrow \quad k_n^{opt}(d) \sim n^{\frac{4}{4+d}} \tag{26}$$

where c refers to any non negative constant and $\beta_d = 2/d$ refers to the smoothness parameter of the model involved in Inequality (24). On the other hand, when $\beta_d = 1/d$, it is shown in [14] that (almost) optimal rates of convergence can be obtained in the non-compact case, choosing for instance

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

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 $d \in \mathbb{N}$, a constant C_1 exists such that*

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

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

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

The proof of this result is given in Appendix C.2.

Remark 4. *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 (26) 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):*

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

  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 $\mathcal{S}$. 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 $k_n^{opt}(\widehat{d})$ chosen as in Equation (26). It is important to note that the sample splitting strategy produces a choice $\widehat{d}$ independent on the samples in $\mathcal{S}_2$.

Theorem 3 below shows that any sample splitting strategy, every choice of $\widehat{d}$ will lead to bad performances of classification on model (1). The proof is postponed to Appendix C.3.

Theorem 3. *In the functional Gaussian translation model, any kNN classifier $\Phi_{k_n^{opt},n,\widehat{d}}$ satisfies*

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

The main conclusion of this section and of Theorem 3 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 choice of $k_n^{opt}(\widehat{d})$. Although this choice appears to be classic regarding the existing literature, 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. However, it should be kept in mind that the misclassification of the classifier $\widehat{\Phi}_{d_n}$ proposed in Equation (16) attains a polynomial rate of convergence, which is an important encouragement for its use instead of the use of a nonparametric classifier associated with a threshold rule.

A   Proof of the upper bounds

The goal of this section is to prove the polynomial upper bound of Theorem 1 together with the intermediate results of Proposition 1 and Lemma 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) are well separated.

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

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

Proof of Proposition 1. We use the Girsanov Equations (4) and (5) that define the likelihood ratio q_f and q_g. We therefore deduce that

$$
\begin{aligned}
\mathbb{P}_X &\left(\left| \eta(X) - \frac{1}{2} \right| \leqslant \varepsilon \right) \\
&= \mathbb{P}_X \left(\frac{|q_f(X) - q_g(X)|}{2(q_f(X) + q_g(X))} \leqslant \varepsilon \right) \\
&= \mathbb{P}_X \left(\left\{ \frac{|q_f(X) - q_g(X)|}{2(q_f(X) + q_g(X))} \leqslant \varepsilon \right\} \cap \{q_f(X) \leqslant q_g(X)\} \right) + \mathbb{P}_X \left(\left\{ \frac{|q_f(X) - q_g(X)|}{2(q_f(X) + q_g(X))} \leqslant \varepsilon \right\} \cap \{q_f(X) > q_g(X)\} \right) \\
&\leqslant \mathbb{P}_X \left(\left\{ \frac{|q_f(X) - q_g(X)|}{4q_g(X)} \leqslant \varepsilon \right\} \cap \{q_f(X) \leqslant q_g(X)\} \right) + \mathbb{P}_X \left(\left\{ \frac{|q_f(X) - q_g(X)|}{4q_f(X)} \leqslant \varepsilon \right\} \cap \{q_f(X) > q_g(X)\} \right) \\
&\leqslant \mathbb{P}_X \left(\left| \frac{q_f(X)}{q_g(X)} - 1 \right| \leqslant 4\varepsilon \right) + \mathbb{P}_X \left(\left| \frac{q_g(X)}{q_f(X)} - 1 \right| \leqslant 4\varepsilon \right) .
\end{aligned}
\tag{27}
$$

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

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

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

$$
\begin{aligned}
\mathbb{P}_X &\left(\left| \frac{q_f(X)}{q_g(X)} - 1 \right| \leqslant 4\varepsilon \right) \\
&= \mathbb{P} \left(\left| e^{\int_0^1 (f-g)(s)f(s)ds + \int_0^1 (f-g)(s)dW_s - \frac{1}{2}[\|f\|^2 - \|g\|^2]} - 1 \right| \leqslant 4\varepsilon \right) \mathbb{P}(Y = 1) \\
&\qquad + \mathbb{P} \left(\left| e^{\int_0^1 (f-g)(s)g(s)ds + \int_0^1 (f-g)(s)dW_s - \frac{1}{2}[\|f\|^2 - \|g\|^2]} - 1 \right| \leqslant 4\varepsilon \right) \mathbb{P}(Y = 0) \\
&= \frac{1}{2} \mathbb{P} \left(\left| e^{\frac{1}{2}\|f-g\|^2 + \int_0^1 (f-g)(s)dW_s} - 1 \right| \leqslant 4\varepsilon \right) + \frac{1}{2} \mathbb{P} \left(\left| e^{-\frac{1}{2}\|f-g\|^2 + \int_0^1 (f-g)(s)dW_s} - 1 \right| \leqslant 4\varepsilon \right) \\
&= \frac{1}{2} \mathbb{P} \left(\left| e^{\frac{1}{2}\Delta^2 + \Delta\xi} - 1 \right| \leqslant 4\varepsilon \right) + \frac{1}{2} \mathbb{P} \left(\left| e^{-\frac{1}{2}\Delta^2 + \Delta\xi} - 1 \right| \leqslant 4\varepsilon \right) ,
\end{aligned}
\tag{28}
$$

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 $\ln(1 + 4\varepsilon) \leqslant 4\varepsilon$ and $\ln(1 - 4\varepsilon) \geqslant -8\varepsilon$ when $\varepsilon \leqslant 1/8$, the above probability can be upper bounded as

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

where the last inequality follows from $\mathbb{P}(a \leqslant \xi \leqslant b) \leqslant (b - a)/\sqrt{2\pi}$ and $12/\sqrt{2\pi} \leqslant 5$. Inverting the roles of f and g, we get by symmetry of the problem that the second term of (27) is also upper bounded by $5\varepsilon/\Delta$. This concludes the proof. $\qquad\square$

Remark 5. *Following the same proof strategy, it is easy to check that the same result hold in the truncated space, i.e., replacing η with η_d and Δ with $\Delta_d := \|\Pi_d(f - g)\|$. Namely, for all $d \in \mathbb{N}^*$ and all $0 < \varepsilon \leqslant 1/8$,*

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

In particular, Equation (28) 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.2 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

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

Proof of Lemma 1. We start with the well-known formula on the excess risk of any classifier (see, e.g., [16]):

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

Then, following a classical control of the excess risk (see, e.g., [14]),

$$
\begin{aligned}
\mathcal{R}_{f,g}(\Phi_d^\star) - \mathcal{R}_{f,g}(\Phi^\star) &= \mathbb{E}\left[|2\eta(X) - 1| \mathbb{1}_{\left\{\Phi_d^\star(X) \neq \Phi^\star(X)\right\}}\left[\mathbb{1}_{\{|\eta(X)-1/2|\leqslant\varepsilon\}} + \mathbb{1}_{\{|\eta(X)-1/2|>\varepsilon\}}\right]\right] \\
&\leqslant \underbrace{2\varepsilon\mathbb{P}\left(|\eta(X) - 1/2| \leqslant \varepsilon\right)}_{:=T_1} + \underbrace{\mathbb{P}\left(\{\Phi_d^\star(X) \neq \Phi^\star(X)\} \cap \{|\eta(X) - 1/2| > \varepsilon\}\right)}_{:=T_2} \quad (29)
\end{aligned}
$$

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 $0 < \varepsilon \leqslant 1/8$), we have

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

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

$$T_2 := \mathbb{P}\left(\{\Phi_d^\star(X) \neq \Phi^\star(X)\} \cap \{|\eta(X) - 1/2| > \varepsilon\}\right) \leqslant \mathbb{P}\left(|\eta_d(X) - \eta(X)| > \varepsilon\right).$$

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

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

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). To alleviate the notation, we set $f_d := \Pi_d(f)$ and $g_d := \Pi_d(g)$. Recall from (4) 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

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

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 vor any random variables Z_1, Z_2, Z_3. We get:

$T_{2,1}$

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

15

Taking the logarithm, we can see that:

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

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

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

$$(30)$$

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

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

where $\xi \sim \mathcal{N}\big(0, \|f_d - f\|^2\big) = \mathcal{N}(0, \delta_d^2)$. But, by a classical (sub)Gaussian tail bound stated, e.g., in [6, p.22], we get

$$S_1 \leqslant \exp\left(-\frac{(\varepsilon/8)^2}{2\delta_d^2}\right) = \exp\left(-\frac{\varepsilon^2}{128\delta_d^2}\right). \tag{31}$$

Combining the last inequality with the assumption

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

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

The second term S_2 introduced in (30) can be dealt similarly, except that we can no longer neglect the positive term $\big(\|f\|^2 - \|f_d\|^2\big)/2 = \delta_d^2/2$: considering again $\xi \sim \mathcal{N}(0, \delta_d^2)$, we have

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

where the last inequality follows from the same Gaussian concentration argument as in (31), and where the inequality before last is because $\delta_d^2/2 \leqslant \varepsilon/8$. Indeed, by the assumptions of Lemma 1,

$$\varepsilon \geqslant \sqrt{512\ln(8^2)}\max\{\delta_d, \widetilde{\delta}_d\} \geqslant 24\max\{\delta_d^2, \widetilde{\delta}_d^2\}, \tag{32}$$

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

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

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

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

16

where the second inequality follows from $\widetilde{\delta}_d \delta_d \leqslant \big(\widetilde{\delta}_d + \delta_d^2\big)/2$, where the third inequality is because $\max\{\delta_d^2, \widetilde{\delta}_d^2\} \leqslant \varepsilon/24$ (by (32)), and where the last inequality follows from the same Gaussian tail bound as the one used in (31) and from the assumption $\widetilde{\delta}_d^2 \leqslant \varepsilon^2 / \big(512 \ln(1/\varepsilon^2)\big)$.

A similar analysis shows that the last term S_4 introduced in (30) also satisfies $S_4 \leqslant \varepsilon^2$. Putting everything together, we finally get

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

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

A.3 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 2 (Section 3.3) that η_d and $\Phi_d^* = \mathbb{1}_{\{\eta_d \geqslant 1/2\}}$ correspond to the regression function and the Bayes classifier of the classification problem when the learner has only access to the projected input $\mathbf{X}_d := (\langle \varphi_j, X \rangle)_{1 \leqslant j \leqslant d}$. Since $\widehat{\Phi}_d(X)$ only depends on X through $\mathbf{X}_d$, its excess risk can be rewritten as

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

where the expectation is with respect to both the sample $(X_i, Y_i)_{1 \leqslant i \leqslant n}$ and the new input X. Now, for all $\varepsilon > 0$,

$$
\begin{aligned}
\mathcal{R}_{f,g}(\widehat{\Phi}_d) - \mathcal{R}_{f,g}(\Phi_d^\star) &= \mathbb{E}\Big[|2\eta_d(X) - 1|\mathbb{1}_{\{\widehat{\Phi}_d(X) \neq \Phi_d^\star(X)\}}\mathbb{1}_{\{|\eta_d(X) - 1/2| \leqslant \varepsilon\}}\Big] \\
&\quad + \mathbb{E}\Big[|2\eta_d(X) - 1|\mathbb{1}_{\{\widehat{\Phi}_d(X) \neq \Phi_d^\star(X)\}}\mathbb{1}_{\{|\eta_d(X) - 1/2| > \varepsilon\}}\Big] \\
&\leqslant 2\varepsilon\mathbb{P}_X\left(|\eta_d(X) - 1/2| \leqslant \varepsilon\right) + \mathbb{P}(|\widehat{\eta}_d(X) - \eta_d(X)| > \varepsilon) \,,
\end{aligned}
$$

where the last inequality follows from the inclusion $\big\{\widehat{\Phi}_d(X) \neq \Phi_d^\star(X)\big\} \cap \big\{|\eta_d(X) - 1/2| > \varepsilon\big\} \subseteq \big\{|\widehat{\eta}_d(X) - \eta_d(X)| > \varepsilon\big\}$ (because $\widehat{\Phi}_d(X) = \mathbb{1}_{\widehat{\eta}_d(X) \geqslant 1/2}$ and $\Phi_d^*(X) = \mathbb{1}_{\eta_d(X) \geqslant 1/2}$). We can now apply the adaptation of Proposition 1 to the truncated space (see Remark 5) to get

$$\mathcal{R}_{f,g}(\widehat{\Phi}_d) - \mathcal{R}_{f,g}(\Phi_d^\star) \leqslant 2\varepsilon\left(1 \wedge \frac{10\varepsilon}{\Delta_d}\right) + \mathbb{P}(|\widehat{\eta}_d(X) - \eta_d(X)| > \varepsilon) \,. \tag{33}$$

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

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

where, with a slight abuse of notation, the first probability $\mathbb{P}_f(\cdot)$ is with respect to both the sample $(X_i, Y_i)_{1 \leqslant i \leqslant n}$ drawn i.i.d. from (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 2 (Section 3.3), we have, setting $\boldsymbol{\theta}_d := (\theta_1, \ldots, \theta_d)$ and $\boldsymbol{\mu}_d := (\mu_1, \ldots, \mu_d)$,

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

Similarly, by (19), the estimated regression function $\widehat{\eta}_d$ can be rewritten as

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

Using simple algebra, we get

$$
\begin{aligned}
T_1 &:= \mathbb{P}_f(|\widehat{\eta}_d(X) - \eta_d(X)| > \varepsilon) \\
&= \mathbb{P}_f\left(\left|\frac{\widehat{F}_d(X)}{\widehat{F}_d(X) + \widehat{G}_d(X)} - \frac{F_d(X)}{F_d(X) + G_d(X)}\right| > \varepsilon\right) \\
&\leqslant \mathbb{P}_f\left(\left|\frac{\widehat{F}_d(X) - F_d(X)}{F_d(X) + G_d(X)} + \widehat{F}_d(X)\left(\frac{1}{\widehat{F}_d(X) + \widehat{G}_d(X)} - \frac{1}{F_d(X) + G_d(X)}\right)\right| > \varepsilon\right) \\
&\leqslant \mathbb{P}_f\left(\left|\frac{\widehat{F}_d(X) - F_d(X)}{F_d(X) + G_d(X)}\right| > \frac{\varepsilon}{3}\right) + \mathbb{P}_f\left(\left|\widehat{F}_d(X)\left(\frac{1}{\widehat{F}_d(X) + \widehat{G}_d(X)} - \frac{1}{F_d(X) + G_d(X)}\right)\right| > \frac{2\varepsilon}{3}\right) \\
&=: \mathbb{P}(A_1) + \mathbb{P}(A_2).
\end{aligned}
\tag{34}
$$

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

$$
\begin{aligned}
\mathbb{P}(A_1) &= \mathbb{P}_f\left(|\widehat{F}_d(X) - F_d(X)| > \frac{\varepsilon}{3}(F_d(X) + G_d(X))\right) \\
&\leqslant \mathbb{P}_f\left(|\widehat{F}_d(X) - F_d(X)| > \frac{\varepsilon}{3}F_d(X)\right) \\
&= \mathbb{P}_f\left(|\widehat{F}_d(X)/F_d(X) - 1| > \frac{\varepsilon}{3}\right) \\
&= \mathbb{P}_f\left(\left|e^{\langle \mathbf{X}_d - (\widehat{\theta} + \boldsymbol{\theta}_d)/2, \widehat{\theta} - \boldsymbol{\theta}_d\rangle} - 1\right| > \frac{\varepsilon}{3}\right).
\end{aligned}
\tag{35}
$$

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

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

Now, note from (12), (14)–(15), and Remark 1 that, under $\mathbb{P}_{\otimes^n} \otimes \mathbb{P}_f$ and on the event $\{N_1 > 0\}$, the random variables $\xi_j := X_{d,j} - \theta_{d,j} = \langle \varphi_j, X\rangle - \theta_j$, $1 \leqslant j \leqslant d$, and

$$
\zeta_j := \sqrt{N_1}\left(\widehat{\theta}_j - \theta_{d,j}\right) = \frac{1}{\sqrt{N_1}}\sum_{i=1}^{N_1} \varepsilon_{i,j}^1, \qquad 1 \leqslant j \leqslant d,
$$

are i.i.d. $\mathcal{N}(0,1)$ conditionally on $Y_{1:n} := (Y_1, \ldots, Y_n)$. (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 $\xi_1, \ldots, \xi_d, \zeta_1, \ldots, \zeta_d$ are i.i.d. $\mathcal{N}(0,1)$ (unconditionally).

Note also from Hoeffding's lemma (see, e.g., [6]) and $n/2 - \sqrt{n\log(n)/2} \geqslant n/4$ (because $n \geqslant 27$) that

$$
\mathbb{P}\left(N_1 < \frac{n}{4}\right) \leqslant \mathbb{P}\left(N_1 < \frac{n}{2} - \sqrt{\frac{n\log n}{2}}\right) \leqslant \frac{1}{n}.
\tag{36}
$$

Therefore, we deduce that

$$
\begin{aligned}
\mathbb{P}(A_1) \quad &\leqslant \quad \mathbb{P}_f\left(\left|\left\langle \mathbf{X}_d - \boldsymbol{\theta}_d, \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_d\right\rangle\right| > \frac{\varepsilon}{6} - \frac{\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_d\|^2}{2}, N_1 \geqslant \frac{n}{4}\right) + \mathbb{P}\left(N_1 < \frac{n}{4}\right) \\
&\leqslant \quad \mathbb{P}\left(\left|\sum_{j=1}^{d} \xi_j \zeta_j\right| \geqslant \frac{\sqrt{N_1}\varepsilon}{6} - \frac{\|\zeta\|^2}{2\sqrt{N_1}}, N_1 \geqslant \frac{n}{4}\right) + \frac{1}{n} \\
&\leqslant \quad \mathbb{P}\left(\left|\sum_{j=1}^{d} \xi_j \zeta_j\right| \geqslant \frac{\sqrt{n}\varepsilon}{12} - \frac{\|\zeta\|^2}{\sqrt{n}}\right) + \frac{1}{n} \\
&\leqslant \quad \mathbb{P}\left(\left|\sum_{j=1}^{d} \xi_j \zeta_j\right| \geqslant \frac{\sqrt{n}\varepsilon}{24}\right) + \mathbb{P}\left(\|\zeta\|^2 > \frac{n\varepsilon}{24}\right) + \frac{1}{n}. \quad (37)
\end{aligned}
$$

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 $(\xi_1, \ldots, \xi_d)$, we get

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

where the last inequality is because, conditionally on $(\xi_1, \ldots, \xi_d)$, 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 $\mathbb{P}(|Z| > z \,|\, \xi_1, \ldots, \xi_d) \leqslant 2 e^{-z^2/(2V)}$ for all $z > 0$. But, distinguishing whether $\sum_{j=1}^{d} \xi_j^2$ is below or above $4d \log n$, we obtain

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

where we used the concentration inequality for the χ^2 statistics of [25, Lemma 1]

$$
\forall x > 0, \qquad \mathbb{P}\left(\sum_{j=1}^{d} \xi_j^2 > d + 2\sqrt{dx} + 2x\right) \leqslant e^{-x} \quad (38)
$$

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

$$
d + 2\sqrt{d \log n} + 2\log n \leqslant 2d + 3\log n \leqslant 4d \log n. \quad (39)
$$

Plugging the above inequalities into (37), we finally obtain

$$
\mathbb{P}(A_1) \leqslant 2\exp\left(-\frac{n\varepsilon^2}{4608 d \log n}\right) + \mathbb{P}\left(\|\zeta\|^2 > \frac{n\varepsilon}{24}\right) + \frac{3}{n}. \quad (40)
$$

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

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

The first term has already been studied above (see (35) and the following inequalities) and thus satisfies the same upper bound as $\mathbb{P}(A_1)$ in (40). As for the second term, following the same lines as those leading to (37), we can see that

$$
\mathbb{P}_f\left(\left|G_d(X) - \widehat{G}_d(X)\right| > \frac{\varepsilon}{3}G_d(X)\right)
$$

$$
\leqslant \mathbb{P}_f\left(\left|\left\langle \mathbf{X}_d - \frac{\widehat{\mu} + \boldsymbol{\mu}_d}{2}, \widehat{\mu} - \boldsymbol{\mu}_d\right\rangle\right| > \frac{\varepsilon}{6}\right)
$$

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

$$
\leqslant \mathbb{P}_f\left(\left|\left\langle \mathbf{X}_d - \boldsymbol{\theta}_d, \widehat{\mu} - \boldsymbol{\mu}_d\right\rangle\right| > \frac{\varepsilon}{6} - \underbrace{\left(\|\boldsymbol{\theta}_d - \boldsymbol{\mu}_d\| + \frac{\|\widehat{\mu} - \boldsymbol{\mu}_d\|}{2}\right)\|\widehat{\mu} - \boldsymbol{\mu}_d\|}_{\leqslant 4\left(\Delta_d + 2\sqrt{\frac{d\log(n)}{n}}\right)\sqrt{\frac{d\log(n)}{n}}\text{ w.p. } \geqslant 1 - 2/n}\right)
$$

$$
\leqslant \mathbb{P}_f\left(\left|\left\langle \mathbf{X}_d - \boldsymbol{\theta}_d, \widehat{\mu} - \boldsymbol{\mu}_d\right\rangle\right| > \frac{\varepsilon}{12}\right) + \frac{2}{n},
$$

where we used (36) and (38)–(39) again, and where the last inequality holds true whenever

$$
\left(\Delta_d + 2\sqrt{\frac{d\log(n)}{n}}\right)\sqrt{\frac{d\log(n)}{n}} \leqslant \frac{\varepsilon}{48}. \tag{41}
$$

Mimicking what we did to derive (37), we then get

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

Putting everything together, we can see that, provided (41) holds,

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

Conclusion Combining all results above, we get, under condition (41),

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

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

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

To conclude the proof, we note that, if (41) holds true, then $n\varepsilon/24 \geqslant 4d\log n \geqslant d + 2\sqrt{d\log n} + 2\log n$ (by (39)), so that $\mathbb{P}\left(\|\zeta\|^2 > n\varepsilon/24\right) \leqslant 1/n$ by (38). $\qquad\square$

A.4 Proof of Theorem 1 (excess risk of $\widehat{\Phi}_{d_n}$)

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

$$
\mathcal{R}_{f,g}(\widehat{\Phi}_{d_n}) - \inf_{\Phi}\mathcal{R}_{f,g}(\Phi) \leqslant \begin{cases} c\, R^{\frac{1}{2s+1}} n^{-\frac{s}{2s+1}}\log(n) & \text{if } \Delta < R^{\frac{1}{2s+1}} n^{-\frac{s}{2s+1}}\log(n) \\ \dfrac{c}{\Delta} R^{\frac{2}{2s+1}} n^{-\frac{2s}{2s+1}}\log^2(n) & \text{if } \Delta \geqslant R^{\frac{1}{2s+1}} n^{-\frac{s}{2s+1}}\log(n) \end{cases} \tag{42}
$$

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 (42) is non-increasing in Δ.

Recall that $\Phi = \mathbb{1}_{\eta \geqslant 1/2}$ is the Bayes (optimal) classifier and that $\Phi^\star_{d_n}$ is the Bayes classifier in the d_n-dimensional truncated space (see Remark 2 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,

$$
\begin{aligned}
\mathcal{R}_{f,g}&(\widehat{\Phi}_{d_n}) - \inf_\Phi \mathcal{R}_{f,g}(\Phi) \\
&= \mathcal{R}_{f,g}(\widehat{\Phi}_{d_n}) - \mathcal{R}_{f,g}(\Phi^\star_{d_n}) + \mathcal{R}_{f,g}(\Phi^\star_{d_n}) - \mathcal{R}_{f,g}(\Phi^\star) \\
&\leqslant 2\varepsilon_1\left(1 \wedge \frac{10\varepsilon_1}{\Delta_{d_n}}\right) + 6\exp\left(-\frac{n\varepsilon_1^2}{4608\, d_n \log n}\right) + \frac{13}{n} + 12\varepsilon_2^2 + 2\varepsilon_2\left(1 \wedge \frac{10\varepsilon_2}{\Delta}\right) \\
&\leqslant 2\varepsilon_1\left(1 \wedge \frac{10\varepsilon_1}{\Delta_{d_n}}\right) + \frac{19}{n} + 12\varepsilon_2^2 + 2\varepsilon_2\left(1 \wedge \frac{10\varepsilon_2}{\Delta}\right),
\end{aligned}
\tag{43}
$$

where $\Delta_{d_n} := \|f_{d_n} - g_{d_n}\|$, and where we assumed that $\varepsilon_1^2 \geqslant 4608\, d_n \log^2(n)/n$ (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 $N_{s,R} \geqslant 1/R^2$, so that $d_n := \lfloor (R^2 n)^{\frac{1}{2s+1}} \rfloor \geqslant 1$ for all $n \geqslant N_{s,R}$. 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 $n \geqslant N_{s,R}$ that

$$
\sqrt{\frac{d_n}{n}} \leqslant R^{\frac{1}{2s+1}} n^{-\frac{s}{2s+1}} \leqslant Rd_n^{-s} \leqslant 2R^{\frac{1}{2s+1}} n^{-\frac{s}{2s+1}}.
\tag{44}
$$

We now choose ε_1 and ε_2 so as to minimize (43), while meeting the assumptions of Lemmas 2 and 1.

- We choose
$$
\varepsilon_1 := 48\left(\Delta_{d_n} + 2\sqrt{\frac{d_n \log(n)}{n}} + \sqrt{2\log(n)}\right)\sqrt{\frac{d_n \log(n)}{n}}.
$$

  This entails that $0 < \varepsilon_1 \leqslant 1/8$ for all $n \geqslant N_{s,R}$ (provided $N_{s,R}$ is chosen large enough), that Assumption (22) of Lemma 2 holds true, and that the requirement $\varepsilon_1^2 \geqslant 4608\, d_n \log^2(n)/n$ above is met.

- We choose
$$
\varepsilon_2 := 32Rd_n^{-s}\sqrt{\log\frac{1}{32Rd_n^{-s}}}.
$$

  Choosing $N_{s,R}$ large enough, we can guarantee for all $n \geqslant N_{s,R}$ that $0 < \varepsilon_2 \leqslant 1/8$, as well as $\log\!\big[1/(32Rd_n^{-s})\big] \geqslant 1$ so that $\varepsilon_2 \geqslant 32Rd_n^{-s}$ and therefore $\varepsilon_2 \geqslant 32Rd_n^{-s}\sqrt{\log(1/\varepsilon_2)}$, i.e.,

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

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

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

  Combining the above inequalities implies that Assumption (20) of Lemma 1 is met.

Before plugging the values of ε_1 and ε_2 into (43), we compare Δ_{d_n} with Δ:

$$
\Delta_{d_n} := \|f_{d_n} - g_{d_n}\| \geqslant \|f - g\| - \|f - f_{d_n}\| - \|g - g_{d_n}\| \geqslant \Delta - 2Rd_n^{-s} \geqslant \frac{\Delta}{10}
\tag{45}
$$

whenever $\Delta \geqslant (20/9)Rd_n^{-s}$. By (44) a sufficient condition is that $\Delta \geqslant (40/9)R^{\frac{1}{2s+1}} n^{-\frac{s}{2s+1}}$ or even that $\Delta \geqslant R^{\frac{1}{2s+1}} n^{-\frac{s}{2s+1}} \log(n)$ (provided $N_{s,R} \geqslant e^{40/9} \approx 85.2$). This is the threshold value we use below, since it makes the righ-hand side of (42) 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 (43) and discard the (relatively large) terms $10\,\varepsilon_1/\Delta_{d_n}$ and $10\,\varepsilon_2/\Delta$. We obtain, noting that $12\varepsilon_2^2 \leqslant 12\varepsilon_2/8 \leqslant 2\varepsilon_2$:

$$
\begin{aligned}
\mathcal{R}_{f,g}(\widehat{\Phi}_{d_n}) - \inf_{\Phi} \mathcal{R}_{f,g}(\Phi) &\leqslant 2\varepsilon_1 + \frac{19}{n} + 12\varepsilon_2^2 + 2\varepsilon_2 \leqslant 2\varepsilon_1 + 4\varepsilon_2 + \frac{19}{n} \\
&\leqslant 96\left(\Delta_{d_n} + 2\sqrt{\frac{d_n \log(n)}{n}} + \sqrt{2\log(n)}\right)\sqrt{\frac{d_n \log(n)}{n}} + 128 R d_n^{-s}\sqrt{\log\frac{1}{32 R d_n^{-s}}} + \frac{19}{n} \\
&\leqslant 96\left(2R + 2R^{\frac{1}{2s+1}} n^{-\frac{s}{2s+1}}\sqrt{\log(n)} + \sqrt{2\log(n)}\right) R^{\frac{1}{2s+1}} n^{-\frac{s}{2s+1}}\sqrt{\log(n)} \\
&\quad + 256 R^{\frac{1}{2s+1}} n^{-\frac{s}{2s+1}}\sqrt{\log\frac{n^{\frac{s}{2s+1}}}{32 R^{\frac{1}{2s+1}}}} + \frac{19}{n} \\
&\leqslant c_1 R^{\frac{1}{2s+1}} n^{-\frac{s}{2s+1}} \log(n)\,,
\end{aligned}
\tag{46}
$$

where the inequality before last follows from (44) and from $\Delta_{d_n} \leqslant \Delta \leqslant \|f\| + \|g\| \leqslant 2R$ (since $f, g \in \mathcal{H}_s(R)$), and where (46) holds for all $n \geqslant N_{s,R}$ provided the absolute constant $c_1 > 0$ and the constant $N_{s,R}$ are chosen large enough.

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

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 (43) and then (45) that, for some absolute constants $c_2, c_3 > 0$,

$$
\begin{aligned}
\mathcal{R}_{f,g}(\widehat{\Phi}_{d_n}) - \inf_{\Phi} \mathcal{R}_{f,g}(\Phi) &\leqslant \frac{20\varepsilon_1^2}{\Delta_{d_n}} + \frac{19}{n} + 12\varepsilon_2^2 + \frac{20\varepsilon_2^2}{\Delta} \leqslant \frac{200\varepsilon_1^2}{\Delta} + \frac{20\varepsilon_2^2}{\Delta} + 12\varepsilon_2^2 + \frac{19}{n} \\
&\leqslant c_2 R^{\frac{2}{2s+1}} n^{-\frac{2s}{2s+1}}\left(\frac{\log^2(n)}{\Delta} + \frac{\log(n)}{\Delta} + \log(n)\right) + \frac{19}{n} \\
&\leqslant \frac{c_3 R^{\frac{2}{2s+1}} n^{-\frac{2s}{2s+1}} \log^2(n)}{\Delta}\,,
\end{aligned}
\tag{47}
$$

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

Conclusion: We derive (42) by combining (46) and (47) and by choosing $c := \max\{c_1, c_3\}$. This concludes the proof of Theorem 1.

B Proof of the minimax lower bound (Theorem 2)

This section contains the proof of our minimax lower bound (Theorem 2). We will pay a specific attention to the influence of the separation distance $\Delta = \|f - g\|$ on the misclassification rate. We directly start with the proof in Section B.1 below. We will use several key technical ingredients gathered in Section B.2.

B.1 Proof of Theorem 2

First case: $\Delta < R^{1/(2s+1)}\, n^{-s/(2s+1)}$. Note that

$$
\left\{(f,g) \in \mathcal{H}_s(R) \times \mathcal{H}_s(R) : \|f - g\| \geqslant \Delta\right\} \supseteq \left\{(f,g) \in \mathcal{H}_s(R) \times \mathcal{H}_s(R) : \|f - g\| \geqslant R^{1/(2s+1)}\, n^{-s/(2s+1)}\right\}.
$$

Therefore, taking the supremum over all such functions, we directly obtain a lower bound on the minimax excess risk by applying the lower bound $\left(ce^{-2\Delta^2}/\Delta\right) R^{2/(2s+1)} n^{-2s/(2s+1)}$ of the second case below with $\Delta = R^{1/(2s+1)}\, n^{-s/(2s+1)}$. This yields the desired lower bound of $ce^{-2R^{2/(2s+1)}} R^{1/(2s+1)} n^{-s/(2s+1)}$.

<u>Second case</u>: $\Delta \geqslant R^{1/(2s+1)} n^{-s/(2s+1)}$. We proceed in three main steps.

Step 1: reduction to a finite-dimensional $\mathbb{L}^1$-estimation problem, and some notation.
Finite-dimensional construction. Let $\widehat{\Phi}$ be any classifier built from the sample $(X_i, Y_i)_{1 \leqslant i \leqslant n}$. As is usual when deriving nonparametric lower bounds, we restrict the supremum over all $f, g \in \mathcal{H}_s(R)$ to a well-chosen finite-dimensional subset. More precisely, in what follows, we restrict our attention to functions $f : [0,1] \to \mathbb{R}$ and $g : [0,1] \to \mathbb{R}$ of the form:

$$\forall t \in \mathbb{R}, \qquad f(t) = f_\theta(t) := \sum_{j=1}^d \theta_j \varphi_j(t), \quad \theta \in \Theta, \qquad \text{and} \qquad g(t) = 0,$$

for some $d \in \mathbb{N}^*$ and some parameter set $\Theta \subseteq \big\{ \theta \in \mathbb{R}^d : \theta_1 = \Delta$ and $\sum_{j=2}^d \theta_j^2 j^{2s} \leqslant R^2 - \Delta^2 \big\}$ to be made more precise in Step 2 below. Note that $\langle f_\theta, \varphi_j \rangle = \theta_j$, so that the notation θ_j is consistent with that of Section 3.1.

Some notation. The notation we choose for this proof differs slightly from that of the rest of the paper. We write $\mathbb{P}_\theta$ for the joint distribution of the training and test samples $\big((X_i, Y_i)_{1 \leqslant i \leqslant n}, (X, Y) \big)$ when the true parameter is θ, and denote by $\mathbb{E}_\theta$ the corresponding expectation. We also denote by Q_θ the distribution of the process $(Z_t)_{0 \leqslant t \leqslant 1}$ defined by $dZ_t = f_\theta(t)dt + dW(t)$. We define the $\mathbb{L}^1$-norm of h by

$$\|h\|_{L^1(Q_0)} := \int |h(x)| \mathrm{d}Q_0(x) = \mathbb{E}\big[|h(W)|\big].$$

Finally, for $X = (X(t))_{0 \leqslant t \leqslant 1}$ solution of (1), we set

$$\widetilde{X}_j := \langle \varphi_j, X \rangle = \int_0^1 \varphi_j(t) dX(t).$$

Note that when X is a standard Brownian motion on $[0,1]$, then $(\widetilde{X}_j)_{j \geqslant 1}$, are independent standard Gaussian random variables (since $(\varphi_j)_{j \geqslant 1}$ is an orthonormal basis).

Reduction to an $\mathbb{L}^1$-estimation problem. Note that $g = 0 \in \mathcal{H}_s(R)$ and $\{f_\theta : \theta \in \Theta\} \subseteq \mathcal{H}_s(R)$ (see the definition in (21)), and that $\|f_\theta - 0\| = \|\theta\| \geqslant \Delta$ for all $\theta \in \Theta$ (we use the notation $\|.\|$ both in $\mathbb{L}^2([0,1])$ and in $\mathbb{R}^d$). Therefore,

$$\sup_{\substack{f,g \in \mathcal{H}_s(R) \\ \|f-g\| \geqslant \Delta}} \Big\{ \mathcal{R}_{f,g}(\widehat{\Phi}) - \inf_\Phi \mathcal{R}_{f,g}(\Phi) \Big\} \geqslant \sup_{\theta \in \Theta} \Big\{ \mathcal{R}_{f_\theta, 0}(\widehat{\Phi}) - \inf_\Phi \mathcal{R}_{f_\theta, 0}(\Phi) \Big\}$$

$$= \sup_{\theta \in \Theta} \mathbb{E}_\theta \Big[\big|2\eta_\theta(X) - 1\big| \mathbb{1}_{\widehat{\Phi}(X) \neq \Phi_\theta(X)} \Big], \tag{48}$$

where $\eta_\theta(x) = \mathbb{P}_\theta(Y = 1|X = x)$ denotes the regression function corresponding to the statistical model (1) with $f = f_\theta$ and $g = 0$, and where $\Phi_\theta(x) = \mathbb{1}_{\eta_\theta(x) \geqslant 1/2}$ is the associated Bayes classifier.

But, for all $\theta \in \Theta$ and any $\delta \in (0, 1/4)$ (to be chosen later), we have

$$\mathbb{E}_\theta \Big[\big|2\eta_\theta(X) - 1\big| \mathbb{1}_{\widehat{\Phi}(X) \neq \Phi_\theta(X)} \Big] \geqslant \delta \, \mathbb{P}_\theta \Big(\big\{ |2\eta_\theta(X) - 1| \geqslant \delta \big\} \cap \big\{ \widehat{\Phi}(X) \neq \Phi_\theta(X) \big\} \Big)$$

$$\geqslant \delta \, \Big(\mathbb{P}_\theta \big(\widehat{\Phi}(X) \neq \Phi_\theta(X) \big) - \mathbb{P}_\theta \big(|2\eta_\theta(X) - 1| < \delta \big) \Big)$$

$$\geqslant \delta \, \Big(\mathbb{P}_\theta \big(\widehat{\Phi}(X) \neq \Phi_\theta(X) \big) - \frac{5\delta}{\Delta} \Big), \tag{49}$$

where the last inequality follows from Proposition 1. Next, we use a conditional argument to handle the probability above given the training sample $(X_i, Y_i)_{1 \leqslant i \leqslant n}$: the process $X = (X(t))_{0 \leqslant t \leqslant 1}$ defined in (1) is independent from the training sample and has distribution $(Q_0 + Q_\theta)/2$ under $\mathbb{P}_\theta$ (recall that Q_θ denotes the distribution of the process $(Z_t)_{0 \leqslant t \leqslant 1}$ defined by $dZ_t = f_\theta(t)dt + dW(t)$). Therefore, for all $\theta \in \Theta$,

$$\mathbb{P}_\theta \Big(\widehat{\Phi}(X) \neq \Phi_\theta(X) \Big) = \mathbb{E}_\theta \Big\{ \mathbb{P}_\theta \Big(\widehat{\Phi}(X) \neq \Phi_\theta(X) \Big| (X_i, Y_i)_{1 \leqslant i \leqslant n} \Big) \Big\}$$

$$= \mathbb{E}_\theta \Big\{ \int \mathbb{1}_{\widehat{\Phi}(x) \neq \Phi_\theta(x)} \frac{\mathrm{d}Q_0(x) + \mathrm{d}Q_\theta(x)}{2} \Big\}$$

$$\geqslant \frac{1}{2} \, \mathbb{E}_\theta \Big[\big\| \widehat{\Phi} - \Phi_\theta \big\|_{L^1(Q_0)} \Big], \tag{50}$$

where the last inequality follows from the fact that $\mathbb{1}_{\widehat{\Phi}(x) \neq \Phi_\theta(x)} = \left|\widehat{\Phi}(x) - \Phi_\theta(x)\right|$ for all continuous functions $x : [0,1] \to \mathbb{R}$. Putting (48), (49), and (50) together, we finally get

$$\sup_{\substack{f,g \in \mathcal{H}_s(R) \\ \|f-g\| \geqslant \Delta}} \left\{ \mathcal{R}_{f,g}(\widehat{\Phi}) - \inf_\Phi \mathcal{R}_{f,g}(\Phi) \right\} \geqslant \frac{\delta}{2} \left(\sup_{\theta \in \Theta} \mathbb{E}_\theta \left[\left\| \widehat{\Phi} - \Phi_\theta \right\|_{L^1(Q_0)} \right] - \frac{10\delta}{\Delta} \right). \tag{51}$$

Step 2: a key combinatorial and geometrical argument In order to further bound (51) from below, we now specialize Θ to the set given by Lemma 3 in Appendix B.2, whose proof combines Varshamov-Gilbert's lemma with simple but key geometrical arguments in dimension two. More precisely, we use Lemma 3 in Appendix B.2 with $\varepsilon = c/\sqrt{n}$ and $d = \lfloor ((R^2 - \Delta^2)n)^{1/(2s+1)} \rfloor$, for some absolute constant $c \in (0,1]$ to be determined later. Two remarks are in order:

- We have $d \geqslant ((R^2 - \Delta^2)n)^{1/(2s+1)} - 1 \geqslant 32\log(2) + 1$ by the assumption $n \geqslant (32\log(2) + 2)^{2s+1}/(3R^2/4) \geqslant (32\log(2) + 2)^{2s+1}/(R^2 - \Delta^2)$ since $\Delta \leqslant R/2$. In particular the condition $d \geqslant 7$ in Lemma 3 holds true.

- The condition $\Delta \geqslant \sqrt{d}\,\varepsilon$ of Lemma 3 holds since by assumption on Δ, we have

$$\Delta \geqslant R^{1/(2s+1)}\, n^{-s/(2s+1)} = \sqrt{(R^2\,n)^{1/(2s+1)}/n} \geqslant \sqrt{d/n} \geqslant \sqrt{d}\,\varepsilon\,,$$

  by definition of d and ε.

We can thus apply Lemma 3 and find a subset $\Theta \subseteq \{\Delta\} \times \{-\varepsilon,\,\varepsilon\}^{d-1} \subseteq \mathbb{R}^d$ of cardinality $|\Theta| \geqslant e^{(d-1)/8} \geqslant 2$ such that, for all $\theta \neq \theta' \in \Theta$,

$$\left\| \Phi_\theta - \Phi_{\theta'} \right\|_{L^1(Q_0)} \geqslant \frac{\sqrt{d-1}\,\varepsilon}{4\pi\Delta}\, e^{-\Delta^2}. \tag{52}$$

Note that our construction of Θ meets our earlier requirement: for all $\theta \in \Theta$, we have $\sum_{j=2}^d \theta_j^2 j^{2s} \leqslant (d-1)\varepsilon^2 d^{2s} \leqslant d^{2s+1}\varepsilon^2 \leqslant R^2 - \Delta^2$ by definition of $d \leqslant ((R^2 - \Delta^2)n)^{1/(2s+1)}$ and $\varepsilon \leqslant 1/\sqrt{n}$. Therefore, $\Theta \subseteq \{ \theta \in \mathbb{R}^d : \theta_1 = \Delta \text{ and } \sum_{j=2}^d \theta_j^2 j^{2s} \leqslant R^2 - \Delta^2 \}$ as assumed at the beginning of this proof.

Step 3: Reduction to a testing problem with finitely-many hypotheses We now use a classical tool in nonparametric statistics since we reduce the problem to a multiple-hypotheses testing problem. More precisely, using (51) and setting

$$\widehat{\theta} \in \arg\min_{\theta \in \Theta} \left\| \widehat{\Phi} - \Phi_\theta \right\|_{L^1(Q_0)},$$

we can see that

$$\sup_{\substack{f,g \in \mathcal{H}_s(R) \\ \|f-g\| \geqslant \Delta}} \left\{ \mathcal{R}_{f,g}(\widehat{\Phi}) - \inf_\Phi \mathcal{R}_{f,g}(\Phi) \right\} \geqslant \frac{\delta}{2} \left(\sup_{\theta \in \Theta} \mathbb{E}_\theta \left[\mathbb{1}_{\{\widehat{\theta} \neq \theta\}} \left\| \widehat{\Phi} - \Phi_\theta \right\|_{L^1(Q_\mu)} \right] - \frac{10\delta}{\Delta} \right)$$

$$\geqslant \frac{\delta}{2} \left(\frac{\sqrt{d-1}\,\varepsilon}{8\pi\Delta}\, e^{-\Delta^2} \sup_{\theta \in \Theta} \mathbb{P}_\theta \left(\widehat{\theta} \neq \theta \right) - \frac{10\delta}{\Delta} \right), \tag{53}$$

where in the last inequality we used the fact that, on the event $\{\widehat{\theta} \neq \theta\}$, we necessarily have

$$\left\| \widehat{\Phi} - \Phi_\theta \right\|_{L^1(Q_0)} \geqslant \frac{\sqrt{d-1}\,\varepsilon}{8\pi\Delta}\, e^{-\Delta^2}$$

by a combination of Inequality (52), the definition of $\widehat{\theta}$, and the triangle inequality.

We now lower bound the worst-case testing error $\sup_{\theta \in \Theta} \mathbb{P}_\theta(\widehat{\theta} \neq \theta)$. Since $\widehat{\theta}$ only depends on the training sample $(X_i, Y_i)_{1 \leqslant i \leqslant n}$, whose distribution we denote by P_θ, we can write $\mathbb{P}_\theta(\widehat{\theta} \neq \theta) = P_\theta(\widehat{\theta} \neq \theta)$. We can thus use Fano's inequality (cf. Lemma 8 in Appendix B.2.3) with the events $A_\theta = \{\widehat{\theta} = \theta\}$, the

24

distributions P_θ, $\theta \in \Theta$, and the reference distribution $\mathbb{Q} = P_{\theta_0}$, where $\theta_0 := (\Delta, 0, \ldots, 0) \in \mathbb{R}^d$. We obtain:

$$\inf_{\theta \in \Theta} P_\theta\big(\widehat{\theta} = \theta\big) \leqslant \frac{1}{|\Theta|} \sum_{\theta \in \Theta} P_\theta\big(\widehat{\theta} = \theta\big) \leqslant \frac{\dfrac{1}{|\Theta|} \sum_{\theta \in \Theta} \mathrm{KL}\big(P_\theta, P_{\theta_0}\big) + \log 2}{\log |\Theta|} \ . \tag{54}$$

Using the chain rule for the Kullback-Leibler divergence, and following similar computations as in Section 2 (application of Girsanov's formula), we can see that, for all $\theta \in \Theta$,

$$\mathrm{KL}\big(P_\theta, P_{\theta_0}\big) = n\left(\mathrm{KL}\big(\mathrm{Ber}(1/2), \mathrm{Ber}(1/2)\big) + \frac{\mathrm{KL}(Q_\theta, Q_{\theta_0}) + \mathrm{KL}(Q_0, Q_0)}{2}\right) = \frac{n\|\theta - \theta_0\|^2}{4} = \frac{n(d-1)\varepsilon^2}{4} \, ,$$

where we used the fact that $\theta \in \Theta \subseteq \{\Delta\} \times \{-\varepsilon, \varepsilon\}^{d-1}$ and $\theta_0 := (\Delta, 0, \ldots, 0)$. Combining (54) with the Kullback-Leibler upper bound above, and recalling that $|\Theta| \geqslant e^{(d-1)/8}$, we get

$$\inf_{\theta \in \Theta} P_\theta\big(\widehat{\theta} = \theta\big) \leqslant \frac{n(d-1)\varepsilon^2/4 + \log 2}{(d-1)/8} \leqslant 2c^2 + \frac{1}{4} \, ,$$

where the last inequality follows from $\varepsilon = c/\sqrt{n}$ and $d \geqslant 32\log(2) + 1$. As a consequence, choosing $c := 1/(2\sqrt{2})$,

$$\sup_{\theta \in \Theta} P_\theta\big(\widehat{\theta} \neq \theta\big) \geqslant 1 - 2c^2 - \frac{1}{4} = \frac{1}{2} \ .$$

Plugging the last lower bound into (53), we finally get

$$\sup_{\substack{f, g \in \mathcal{H}_s(R) \\ \|f-g\| \geqslant \Delta}} \left\{ \mathcal{R}_{f,g}(\widehat{\Phi}) - \inf_\Phi \mathcal{R}_{f,g}(\Phi) \right\} \geqslant \frac{5\delta}{\Delta}\left(\frac{\sqrt{d-1}\,\varepsilon}{160\pi}\,e^{-\Delta^2} - \delta\right) = \frac{(d-1)\,\varepsilon^2}{20480\pi^2\Delta}\,e^{-2\Delta^2}$$

with the particular choice of $\delta = \sqrt{d-1}\,\varepsilon\,e^{-\Delta^2}/(320\pi)$. We conclude the proof by substituting the values of $\varepsilon = c/\sqrt{n}$ and $d - 1 = \big\lfloor \big((R^2 - \Delta^2)\,n\big)^{1/(2s+1)} \big\rfloor - 1 \geqslant (6/8)\big((R^2 - \Delta^2)\,n\big)^{1/(2s+1)}$ (since $\lfloor x \rfloor - 1 \geqslant 6x/8$ for all $x \geqslant 7$) and by using the fact that $R^2 - \Delta^2 \geqslant 3R^2/4$ (since $\Delta \leqslant R/2$). Note also that, by the assumption $n \geqslant R^{1/s}$, we have $\delta < 1/4$ as required in the analysis. This concludes the proof of Theorem 2.

B.2 A key combinatorial and geometrical lemma

In this section, we provide a key combinatorial and geometrical lemma to derive the minimax lower bound of Theorem 2. Indeed, the next result guarantees the existence of a parameter set $\Theta \subset \mathbb{R}^d$ such that—when ε is chosen small enough—it is statistically hard to estimate the true value of the parameter $\theta \in \Theta$, while all Bayes classifiers Φ_θ and $\Phi_{\theta'}$, $\theta \neq \theta' \in \Theta$, are sufficiently far from one another, thus leading to a large classification excess risk.

Lemma 3. *Let $d \geqslant 7$, $\varepsilon > 0$, and $\Delta \geqslant \sqrt{d}\,\varepsilon$. There exists a subset $\Theta \subseteq \{\Delta\} \times \{-\varepsilon, \varepsilon\}^{d-1} \subseteq \mathbb{R}^d$ of cardinality $|\Theta| \geqslant e^{(d-1)/8} \geqslant 2$ such that, for all $\theta \neq \theta' \in \Theta$,*

$$\big\|\Phi_\theta - \Phi_{\theta'}\big\|_{L^1(Q_0)} \geqslant \frac{\sqrt{d-1}\,\varepsilon}{4\pi\Delta}\,e^{-\Delta^2} \, , \tag{55}$$

where Q_0 denotes the distribution of a standard Brownian motion $W = (W(t))_{0 \leqslant t \leqslant 1}$ on $[0, 1]$, and where $\|h\|_{L^1(Q_0)} := \mathbb{E}\big[|h(W)|\big]$.

The proof is provided in Section B.2.2 below. We first state three intermediary results.

B.2.1 Intermediary results

The following lemma shows that, for the d-dimensional construction of Section B.1 (Step 1), the Bayes classifier Φ_θ only depends on the d random variables $\widetilde{X}_j := \int_0^1 \varphi_j(t)dX(t)$, $1 \leqslant j \leqslant d$, and takes the form of a simple linear classifier in $\mathbb{R}^d$. We recall that $(\varphi_j)_{j \geqslant 1}$ is any Hilbert basis of $\mathbb{L}^2([0,1])$ and that $f_\theta = \sum_{j=1}^d \theta_j \varphi_j$.

Lemma 4. *Consider the statistical construction of Section B.1 (Step 1). Let $W = (W(t))_{0 \leqslant t \leqslant 1}$ be a standard Brownian motion and define $\widetilde{W}_j := \int_0^1 \varphi_j(t) dW(t)$ as well as $\widetilde{W} := \left(\widetilde{W}_j\right)_{1 \leqslant j \leqslant d} \in \mathbb{R}^d$. Then, the Bayes classifier $\Phi_\theta = \mathbb{1}_{\eta_\theta \geqslant 1/2}$ satisfies*

$$
\Phi_\theta(W) = \begin{cases} 0 & \text{if} \quad \|\widetilde{W} - \theta\| > \|\widetilde{W}\| \\ 1 & \text{if} \quad \|\widetilde{W} - \theta\| \leqslant \|\widetilde{W}\| \end{cases} \qquad \text{almost surely.}
$$

Proof. The result follows directly from the calculations of Section 2.1 (application of Girsanov's formula). Indeed, using (6) and the fact that $g = 0$ and $\|f_\theta\| = \|\theta\|$, we obtain

$$
\eta_\theta(W) \geqslant 1/2 \iff \int_0^1 f_\theta(t) dW(t) \geqslant \frac{\|f_\theta\|^2}{2}
$$
$$
\iff \widetilde{\theta} \cdot \widetilde{W} \geqslant \frac{\|\theta\|^2}{2}
$$
$$
\iff \|\widetilde{W} - \theta\|^2 \leqslant \|\widetilde{W}\|^2 \,,
$$

which concludes the proof. $\qquad\square$

The above lemma shows that the Bayes classifier Φ_θ corresponds to a linear classifier in $\mathbb{R}^d$ (after projecting onto $(\varphi_j)_{1 \leqslant j \leqslant d}$). The next lemma provides a lower bound on the angle between the hyperplanes associated with two linear classifiers Φ_θ and $\Phi_{\theta'}$, for $\theta \neq \theta' \in \Theta$. This result will be crucial in our proof of the lower bound of Lemma 3.

We recall that the (undirected) internal angle between two non-zero vectors $\theta, \theta' \in \mathbb{R}^d$ is given by

$$
\angle(\theta, \theta') := \arccos\left(\frac{\langle \theta, \theta' \rangle}{\|\theta\| \, \|\theta'\|}\right) \in [0, \pi] \, ;
$$

this angle is in particular well defined for all $\theta, \theta' \in \Theta$ (since $0 \notin \Theta$ by construction).

Lemma 5. *Let $d \geqslant 7$, $\varepsilon > 0$, and $\Delta \geqslant \sqrt{d}\varepsilon$. Let $\Gamma \subseteq \{-1, 1\}^{d-1}$ be a set provided by Varshamov-Gilbert's lemma in dimension $m = d - 1$ (see, e.g., Lemma 7 in Appendix B.2.3), and define*

$$
\Theta := \{\Delta\} \times (\varepsilon\Gamma) = \left\{ (\Delta, \varepsilon u_1, \varepsilon u_2, \dots, \varepsilon u_{d-1}) : (u_1, \dots, u_{d-1}) \in \Gamma \right\} \subset \mathbb{R}^d \,. \tag{56}
$$

Then, for all $\theta \neq \theta' \in \Theta$, the internal angle $\angle(\theta, \theta')$ between the vectors θ and θ' is bounded by

$$
\frac{\sqrt{d-1}\,\varepsilon}{2\Delta} \leqslant \angle(\theta, \theta') \leqslant \frac{\pi}{2} \,.
$$

Proof. Let $\theta \neq \theta' \in \Theta$. By (56) we can write $\theta = (\Delta, \varepsilon u_1, \dots, \varepsilon u_{d-1})$ and $\theta' = (\Delta, \varepsilon u'_1, \dots, \varepsilon u'_{d-1})$ with $u \neq u' \in \Gamma$. We also set $m = d - 1$. We have

$$
\cos\left(\angle(\theta, \theta')\right) = \frac{\langle \theta, \theta' \rangle}{\|\theta\| \, \|\theta'\|} = \frac{\Delta^2 + \varepsilon^2 \sum_{j=1}^m u_j u'_j}{\sqrt{\Delta^2 + m\varepsilon^2}\sqrt{\Delta^2 + m\varepsilon^2}} = \frac{\Delta^2 + \varepsilon^2 \sum_{j=1}^m u_j u'_j}{\Delta^2 + m\varepsilon^2} \,. \tag{57}
$$

Note that $u_j u'_j \in \{-1, 1\}$ so that $\Delta^2 + \varepsilon^2 \sum_{j=1}^m u_j u'_j \geqslant \Delta^2 - m\varepsilon^2 \geqslant 0$ because we assumed that $\Delta \geqslant \sqrt{d}\varepsilon$. Therefore, $\cos\left(\angle(\theta, \theta')\right) \geqslant 0$, which in turn entails that $\angle(\theta, \theta') \leqslant \pi/2$ since $\angle(\theta, \theta') \in [0, \pi]$ by definition.

We now prove the lower bound on $\angle(\theta, \theta')$. By construction of Γ (Lemma 7 in Appendix B.2.3), we have $u_j u'_j \in \{-1, 1\}$ and $\sum_{j=1}^m \mathbb{1}_{\{u_j \neq u'_j\}} \geqslant m/4$, so that $\sum_{j=1}^m u_j u'_j \leqslant -m/4 + 3m/4 = m/2$. Substituting this upper bound in (57) yields

$$
\cos\left(\angle(\theta, \theta')\right) \leqslant \frac{\Delta^2 + m\varepsilon^2/2}{\Delta^2 + m\varepsilon^2} = 1 - \frac{m\varepsilon^2/2}{\Delta^2 + m\varepsilon^2} \,.
$$

Using the former result $\cos\left(\angle(\theta, \theta')\right) \geqslant 0$ and the last inequality above, we obtain

$$
\sin^2\left(\angle(\theta, \theta')\right) = 1 - \cos^2\left(\angle(\theta, \theta')\right) \geqslant 1 - \cos\left(\angle(\theta, \theta')\right) \geqslant \frac{m\varepsilon^2/2}{\Delta^2 + m\varepsilon^2} \geqslant \frac{m\varepsilon^2}{4\Delta^2} \,,
$$

where we again used $m = d - 1 \leqslant d$ and our assumption on Δ: $\sqrt{m}\,\varepsilon \leqslant \sqrt{d}\,\varepsilon \leqslant \Delta$. We conclude the proof by noting that $\angle(\theta, \theta') \geqslant \sin(\angle(\theta, \theta')) = \sqrt{\sin^2(\angle(\theta, \theta'))}$ since $\angle(\theta, \theta') \in [0, \pi]$:

$$\angle(\theta, \theta') \geqslant \frac{\sqrt{m}\varepsilon}{2\Delta} = \frac{\sqrt{d-1}\varepsilon}{2\Delta}\,.$$

$\square$

Our third and last lemma in this subsection provides a lower bound on the Gaussian measure of a double cone in dimension 2. We say that $\mathcal{C} \subset \mathbb{R}^2$ is an *open double cone with apex* $z \in \mathbb{R}^2$ if it is of the form

$$\mathcal{C} = \left\{ z + au + bv : (a, b) \in \mathbb{R}_+^{\star 2} \cup \mathbb{R}_-^{\star 2} \right\}$$

for some linearly independent vectors $u, v \in \mathbb{R}^2$. It is clear that there is not a one-to-one correspondence between (u, v) and $\mathcal{C}$ (several pairs (u, v) correspond to the same $\mathcal{C}$). However, the value of the internal angle $\angle(u, v) := \arccos(\langle u, v \rangle / (\|u\| \, \|v\|)) \in (0, \pi)$ between u and v is the same for all pairs (u, v) that correspond to $\mathcal{C}$. We thus call $\angle(u, v)$ the *angle of the open double cone* $\mathcal{C}$.

Lemma 6. *Let $\mathcal{C} \subset \mathbb{R}^2$ be an open double cone with apex $z \in \mathbb{R}^2$ and angle $\mathcal{A} \in (0, \pi)$. Then, the measure of $\mathcal{C}$ with respect to the standard Gaussian distribution $\gamma_2 = \mathcal{N}(0, \mathbb{I}_{2 \times 2})$ on $\mathbb{R}^2$ is lower bounded by*

$$\gamma_2(\mathcal{C}) \geqslant \frac{\mathcal{A}}{2\pi}\, e^{-\|z\|^2}\,.$$

We emphasize that rather intuitively, the above lower bound is proportional to the angle $\mathcal{A}$ and decreases exponentially fast with $\|z\|^2$. (The constant of 1 appearing in the exponential could certainly be optimized, but this one is sufficient for our purposes.)

Proof. We carry out a change of variables by a translation around z: writing $\mathcal{C} - z = \{x - z : x \in \mathcal{C}\}$ and using the inequality $\|z + u\|^2 \leqslant 2\|z\|^2 + 2\|u\|^2$, we get

$$\gamma_2(\mathcal{C}) = \frac{1}{2\pi} \int_{\mathcal{C}} e^{-\|x\|^2/2}\,\mathrm{d}x = \frac{1}{2\pi} \int_{\mathcal{C}-z} e^{-\|z+u\|^2/2}\,\mathrm{d}u \geqslant \frac{e^{-\|z\|^2}}{2\pi} \int_{\mathcal{C}-z} e^{-\|u\|^2}\mathrm{d}u$$

$$= \frac{e^{-\|z\|^2}}{2\pi}\, 2 \int_0^{\mathcal{A}} \left(\int_0^{+\infty} r e^{-r^2}\mathrm{d}r \right) \mathrm{d}\alpha = \frac{e^{-\|z\|^2}}{2\pi}\, \mathcal{A}\,,$$

where the second line is obtained by parameterizing $\mathcal{C} - z$ with polar coordinates and by noting that $\mathcal{C} - z$ is an open double cone of angle $\mathcal{A}$ pointed at the origin. This concludes the proof. $\square$

B.2.2   Proof of Lemma 3

We now prove Lemma 3 using the intermediary results of the previous subsection. We use the same notation as in Section B.1. Let $\Gamma \subseteq \{-1, 1\}^{d-1}$ be a set provided by Varshamov-Gilbert's lemma in dimension $m = d - 1$ (cf. Lemma 7 in Appendix B.2.3). Next we show that the set

$$\Theta := \{\Delta\} \times (\varepsilon\Gamma) = \left\{ (\Delta, \varepsilon u_1, \varepsilon u_2, \ldots, \varepsilon u_{d-1}) : (u_1, \ldots, u_{d-1}) \in \Gamma \right\} \subset \mathbb{R}^d$$

satisfies the statement of Lemma 3. We can already see that its cardinality is $|\Theta| = |\Gamma| \geqslant e^{m/8} \geqslant e^{(d-1)/8}$. It remains to prove that, for all $\theta \neq \theta' \in \Theta$,

$$\left\| \Phi_\theta - \Phi_{\theta'} \right\|_{L^1(Q_0)} \geqslant \frac{\sqrt{d-1}\,\varepsilon}{4\pi\Delta}\, e^{-\Delta^2}\,, \tag{58}$$

where Q_0 denotes the distribution of a standard Brownian motion $W = (W(t))_{0 \leqslant t \leqslant 1}$ on $[0, 1]$, and where $\|h\|_{L^1(Q_0)} := \mathbb{E}[|h(W)|]$.

Proof of (58). Let $\theta \neq \theta' \in \Theta$. Let $W = (W(t))_{0 \leqslant t \leqslant 1}$ be a standard Brownian motion on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Noting that $\big|\Phi_\theta(W) - \Phi_{\theta'}(W)\big| = \mathbb{1}_{\Phi_\theta(W) \neq \Phi_{\theta'}(W)}$ a.s., we have

$$\begin{aligned}
\big\|\Phi_\theta - \Phi_{\theta'}\big\|_{L^1(Q_0)} &= \mathbb{P}\big(\Phi_\theta(W) \neq \Phi_{\theta'}(W)\big) \\
&= \mathbb{P}\Big(\big\{\|\widetilde{W} - \theta\| \leqslant \|\widetilde{W}\| < \|\widetilde{W} - \theta'\|\big\} \cup \big\{\|\widetilde{W} - \theta'\| \leqslant \|\widetilde{W}\| < \|\widetilde{W} - \theta\|\big\}\Big) \\
&\geqslant \mathbb{P}\Big(\underbrace{\big\{\|\widetilde{W} - \theta\| < \|\widetilde{W}\| < \|\widetilde{W} - \theta'\|\big\} \cup \big\{\|\widetilde{W} - \theta'\| < \|\widetilde{W}\| < \|\widetilde{W} - \theta\|\big\}}_{=:A} \Big),
\end{aligned}$$

where the line before last follows from Lemma 4, and where we recall that $\widetilde{W} := \big(\widetilde{W}_j\big)_{1 \leqslant j \leqslant d} \in \mathbb{R}^d$ with $\widetilde{W}_j := \int_0^1 \varphi_j(t)dW(t)$. In order to bound $\mathbb{P}(A)$ from below, we project (orthogonally) all points in $\mathbb{R}^d$ onto the unique plane $\mathcal{P}$ that contains 0 and the non-colinear vectors θ and θ' (note from Lemma 5 that $0 < \angle(\theta, \theta') \leqslant \pi/2 < \pi$). As shown in Figure 1, we define $z \in \mathcal{P}$ as the intersection between the perpendicular bisectors $\mathcal{D}$ and $\mathcal{D}'$ of the segments $[0, \theta]$ and $[0, \theta']$ on the plane $\mathcal{P}$. Writing $r_{-\pi/2}$ for the rotation of angle $-\pi/2$ on the plane $\mathcal{P}$, we also consider the unit vectors $u = r_{-\pi/2}\big(\theta/\|\theta\|\big)$ and $v = r_{-\pi/2}\big(\theta'/\|\theta'\|\big)$ that support the lines $\mathcal{D}$ and $\mathcal{D}'$ respectively.
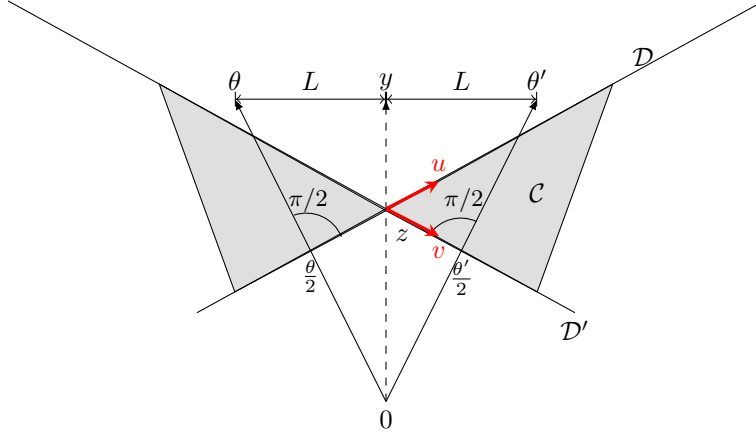

Figure 1: The main objects of interest on the plane $\mathcal{P}$.

Writing $\widetilde{W}_\mathcal{P}$ for the orthogonal projection of $\widetilde{W} \in \mathbb{R}^d$ onto $\mathcal{P}$, we can see that

$$\mathbb{P}(A) = \mathbb{P}\Big(\widetilde{W}_\mathcal{P} \in \mathcal{C}\Big) \quad \text{with} \quad \mathcal{C} := \Big\{z + au + bv : (a, b) \in \mathbb{R}_+^{*2} \cup \mathbb{R}_-^{*2}\Big\}.$$

Let (e_1, e_2) be any orthonormal basis of $\mathcal{P}$. Decomposing any $w \in \mathcal{P}$ as $w = w^1 e_1 + w^2 e_2$ (and similarly for u and v), we can see that

$$w \in \mathcal{C} \iff (w^1, w^2) \in \underbrace{\Big\{(z^1, z^2) + a(u^1, u^2) + b(v^1, v^2) : (a, b) \in \mathbb{R}_+^{*2} \cup \mathbb{R}_-^{*2}\Big\}}_{=:\widetilde{\mathcal{C}}}.$$

Therefore,

$$\mathbb{P}(A) = \mathbb{P}\Big(\big(\widetilde{W}_\mathcal{P}^1, \widetilde{W}_\mathcal{P}^2\big) \in \widetilde{\mathcal{C}}\Big) = \gamma_2\big(\widetilde{\mathcal{C}}\big),$$

where $\gamma_2 = \mathcal{N}(0, \mathbb{I}_{2 \times 2})$ denotes the standard Gaussian distribution on $\mathbb{R}^2$. The last equality holds true because $W = (W(t))_{0 \leqslant t \leqslant 1}$ is a standard Brownian motion so that the $\widetilde{W}_j = \int_0^1 \varphi_j(t)dW(t)$, $1 \leqslant j \leqslant d$, are independent $\mathcal{N}(0, 1)$ random variables (because the φ_j are orthonormal), so that $\big(\widetilde{W}^1, \widetilde{W}^2\big)$ is a standard two-dimensional Gaussian vector (because e_1 and e_2 are orthonormal).

Now, we note that the subset $\widetilde{\mathcal{C}} \subset \mathbb{R}^2$ is an open double cone with apex (z_1, z_2). Since (e_1, e_2) is an orthonormal basis of $\mathcal{P}$, the angle of $\widetilde{\mathcal{C}}$ is equal to $\angle(u, v) = \angle\big(r_{-\pi/2}\big(\theta/\|\theta\|\big), r_{-\pi/2}\big(\theta'/\|\theta'\|\big)\big) = \angle(\theta, \theta')$. Therefore, applying Lemma 6 and then Lemma 5,

$$\mathbb{P}(A) \geqslant \frac{\angle(\theta, \theta')}{2\pi} e^{-(z_1^2 + z_2^2)} \geqslant \frac{e^{-(z_1^2 + z_2^2)}\sqrt{d-1}\,\varepsilon}{4\pi\Delta} \, . \tag{59}$$

We conclude the proof by upper bounding $z_1^2 + z_2^2 = \|z\|^2$ as follows. First note from Figure 1 that

$$\cos\left(\frac{\angle(\theta,\,\theta')}{2}\right) = \frac{\|\theta\|/2}{\|z\|} \qquad \text{so that} \qquad \|z\| = \frac{\|\theta\|}{2\cos\left(\frac{\angle(\theta,\,\theta')}{2}\right)} \, .$$

But, from the inequality $0 \leqslant \angle(\theta, \theta')/2 \leqslant \pi/4$ (see Lemma 5) we get that $\cos\big(\angle(\theta, \theta')/2\big) \geqslant 1/\sqrt{2}$, so that $\|z\| \leqslant \|\theta\|/\sqrt{2}$, i.e.,

$$z_1^2 + z_2^2 \leqslant \frac{\|\theta\|^2}{2} = \frac{\Delta^2 + (d-1)\varepsilon^2}{2} \leqslant \Delta^2$$

by the assumption $\Delta \geqslant \sqrt{d}\varepsilon$. Combining $\|z\|^2 \leqslant \Delta^2$ with Equation (59) concludes the proof. $\qquad\square$

B.2.3   Two well-known lemmas

The next combinatorial result is known as Varshamov-Gilbert's lemma. It provides a lower bound on the packing entropy of the m-dimensional hypercube $\{-1, 1\}^m$ endowed with the Hamming metric, at scale $m/4$. This result indicates that among the 2^m corners of $\{-1, 1\}^m$, exponentially many of them are almost opposite from one another. A proof can be found, e.g., in [28, Lemma 4.7].

Lemma 7 (Varshamov-Gilbert's lemma). *Let $m \geqslant 1$. There exists a subset $\Gamma \subseteq \{-1, 1\}^m$ of cardinality $|\Gamma| \geqslant e^{m/8}$ such that*

$$\forall x \neq y \in \Gamma, \quad \sum_{j=1}^{m} \mathbb{1}_{\{x_j \neq y_j\}} > \frac{m}{4} \, .$$

The next lemma is a well-known version of Fano's inequality that follows, e.g., from [18, Chapter VII, Lemma 1.1] or [12, Theorem 2.11.1] (see also Proposition 1 in the recent survey [15]).

We recall that the Kullback-Leibler divergence $\mathrm{KL}(\mathbb{P}, \mathbb{Q})$ between two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ on the same measurable space $(E, \mathcal{B})$ is defined by

$$\mathrm{KL}(\mathbb{P}, \mathbb{Q}) := \begin{cases} \displaystyle\int_E \ln\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right) \mathrm{d}\mathbb{P} & \text{if } \mathbb{P} \text{ is absolutely continuous with respect to } \mathbb{Q}; \\ +\infty & \text{otherwise.} \end{cases}$$

Lemma 8 (Fano's inequality). *Let $(E, \mathcal{B})$ be any measurable space and $N \geqslant 2$. Let $(A_1, \ldots, A_N)$ be a measurable partition of $(E, \mathcal{B})$ and $(\mathbb{P}_1, \ldots, \mathbb{P}_N)$ a family of probability distributions on $(E, \mathcal{B})$. Then,*

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{P}_i(A_i) \leqslant \frac{\displaystyle\inf_{\mathbb{Q}} \frac{1}{N}\sum_{i=1}^{N} \mathrm{KL}(\mathbb{P}_i, \mathbb{Q}) + \log 2}{\log N} \, ,$$

where the infimum is over all probability distributions $\mathbb{Q}$ on $(E, \mathcal{B})$.

C   Truncated nearest neighbor strategy (Theorem 3)

This appendix section gathers the proof of the lower bound of the nearest neighbor method used with a sample-splitting thresholding strategy, i.e., half of the learning sample is used to choose a thresholding dimension $\widehat{d}_n$ and then the nearest neighbor classifier is computed on the remaining part of the samples. Therefore, $\widehat{d}_n$ is choosen independently from the second part of the samples.

C.1 Smoothness of thee Gaussian translation model

This paragraph is devoted to the computation of the smoothness index β_d involved in the Gaussian translation model in dimension $d \in \mathbb{N}^\star$ (see, e.g., Equation 24). Below, γ will refer to the density of the d-dimensional standard Gaussian random variable and we omit the dependency in d to alleviate the notations.

Proof of Proposition 2. According to the definition of the smoothness parameter given in Equation 24, we compute the average value of η on a ball $B(x, r)$ and compare it to $\eta(x)$:

$$
\begin{aligned}
&\eta(B(x,r)) - \eta(x) \\
&= \frac{1}{\mu(B(x,r))} \int_{B(x,r)} \eta(s) d\mu(s) - \frac{\gamma(x)}{\gamma(x) + \gamma(x-m)}, \\
&= \frac{2}{\int_{B(x,r)} \gamma(s) + \gamma(s-m) ds} \int_{B(x,r)} \frac{\gamma(s)}{\gamma(s) + \gamma(s-m)} \frac{1}{2}[\gamma(s) + \gamma(s-m)] ds - \frac{\gamma(x)}{\gamma(x) + \gamma(x-m)}, \\
&= \frac{\gamma(B(x,r))}{\gamma(B(x,r)) + \gamma(B(x-m,r))} - \frac{\gamma(x)}{\gamma(x) + \gamma(x-m)}, \\
&= \frac{[\gamma(x) + \gamma(x-m)]\gamma(B(x,r)) - \gamma(x)[\gamma(B(x,r)) + \gamma(B(x-m,r))]}{[\gamma(x) + \gamma(x-m)][\gamma(B(x,r)) + \gamma(B(x-m,r))]}, \\
&= \frac{\gamma(x-m)\gamma(B(x,r)) - \gamma(x)\gamma(B(x-m,r))}{[\gamma(x) + \gamma(x-m)][\gamma(B(x,r)) + \gamma(B(x-m,r))]}.
\end{aligned}
\tag{60}
$$

It is then necessary to compare $\gamma(B(x,r))$ with $\gamma(x)\lambda(B_r)$ where $\lambda(B_r)$ is the Lebesgue measure of the centered ball of radius r in $\mathbb{R}^d$. For this purpose, we can use the well known convexity inequality on Gaussian measures of shifted balls:

$$
\exp(-\|x\|^2/2)\gamma(B(0,r)) \leqslant \gamma(B(x,r)) \leqslant \gamma(B(0,r)).
\tag{61}
$$

In particular, we have (see [21]) when $r \longrightarrow 0$ that

$$
\gamma(B(x,r)) \sim \exp(-\|x\|^2/2)\gamma(B(0,r)),
$$

but the r.h.s. of (61) is tight only for x close to 0. Expanding the denominator of (60), we obtain that

$$
\begin{aligned}
&|\eta(B(x,r)) - \eta(x)| \\
&= \frac{|\gamma(x-m)\gamma(B(x,r)) - \gamma(x)\gamma(B(x-m,r))|}{\gamma(x)\gamma(B(x,r)) + \gamma(x)\gamma(B(x-m,r)) + \gamma(x-m)\gamma(B(x,r)) + \gamma(x-m)\gamma(B(x-m,r))} \\
&\leqslant \frac{|\gamma(x-m)\gamma(B(x,r)) - \gamma(x)\gamma(B(x-m,r))|}{\gamma(x)\gamma(B(x-m,r)) + \gamma(x-m)\gamma(B(x,r))}.
\end{aligned}
\tag{62}
$$

Concerning the numerator, a simple change of variable leads to

$$
\begin{aligned}
&\gamma(x-m)\gamma(B(x,r)) - \gamma(x)\gamma(B(x-m,r)) \\
&= (2\pi)^{-d} \int_{B(0,r)} \left\{ e^{-\|x-m\|^2/2} e^{-\|x-s\|^2/2} - e^{-\|x\|^2/2} e^{-\|x-m-s\|^2/2} \right\} ds.
\end{aligned}
$$

For all $x \in \mathbb{R}^d$ and $s \in B(0,r)$, the term inside the integral above may be written as

$$
e^{-\|x-m\|^2/2} e^{-\|x-s\|^2/2} - e^{-\|x\|^2/2} e^{-\|x-m-s\|^2/2} = e^{-\|x-m\|^2/2 - \|x\|^2/2} e^{-\|s\|^2/2} \left[e^{\langle x,s \rangle} - e^{\langle x-m,s \rangle} \right].
$$

We can use the following upper bound for any real value a:

$$
|e^a - 1 - a| \leqslant \frac{a^2 e^{|a|}}{2},
$$

with $a = \langle x, s \rangle$ and $a = \langle x-m, s \rangle$ and deduce that

$$
|e^{\langle x,s \rangle} - e^{\langle x-m,s \rangle} - \langle m,s \rangle| \leqslant \frac{s^2}{2} \left(\|x-m\|^2 e^{|\langle x-m,s \rangle|} + \|x\|^2 e^{|\langle x,s \rangle|} \right).
$$

Therefore, we obtain

$$
\begin{aligned}
&|\gamma(x-m)\gamma(B(x,r)) - \gamma(x)\gamma(B(x-m,r))| \\
&\leqslant \quad \gamma(x)\gamma(x-m)\int_{B(0,r)} e^{-\|s\|^2/2}\langle m,s\rangle ds \\
&\quad + \frac{r^2}{2}\gamma(x)\gamma(x-m)\left[\|x-m\|^2\int_{B(0,r)}e^{-\frac{\|s\|^2}{2}}e^{|\langle x-m,s\rangle|}ds + \|x\|^2\int_{B(0,r)}e^{-\frac{\|s\|^2}{2}}e^{|\langle x,s\rangle|}ds\right] \\
&= \quad \frac{r^2}{2}\gamma(x)\gamma(x-m)\left[\|x-m\|^2\int_{B(0,r)}e^{-\frac{\|s\|^2}{2}}e^{|\langle x-m,s\rangle|}ds + \|x\|^2\int_{B(0,r)}e^{-\frac{\|s\|^2}{2}}e^{|\langle x,s\rangle|}ds\right] \\
&\leqslant \quad \frac{r^2}{2}\gamma(x)\gamma(x-m)\|x-m\|^2\left(\int_{B(0,r)}e^{-\frac{\|s\|^2}{2}}e^{\langle x-m,s\rangle}ds + \int_{B(0,r)}e^{-\frac{\|s\|^2}{2}}e^{-\langle x-m,s\rangle}ds\right) \\
&\quad + \frac{r^2}{2}\gamma(x)\gamma(x-m)\|x\|^2\left(\int_{B(0,r)}e^{-\frac{\|s\|^2}{2}}e^{\langle x,s\rangle}ds + \int_{B(0,r)}e^{-\frac{\|s\|^2}{2}}e^{-\langle x,s\rangle}ds\right) \\
&= \quad \frac{r^2}{2}\left[\|x-m\|^2\gamma(x)[\gamma(B(x-m,r)) + \gamma(B(m-x,r))] + \|x\|^2\gamma(x-m)[\gamma(B(x,r)) + \gamma(B(-x,r))]\right] \\
&= \quad r^2\left[\|x-m\|^2\gamma(x-m)\gamma(B(x,r)) + \|x\|^2\gamma(x)\gamma(B(x-m,r))\right],
\end{aligned}
$$

where the last line comes from the symmetry of the Gaussian distribution. Using this last inequality in Inequality (62) yields:

$$|\eta(B(x,r)) - \eta(x)| \leqslant r^2\left[\|x-m\|^2 + \|x\|^2\right]. \tag{63}$$

Now, we should remark that

$$\gamma(B(0,r)) = \int_{B(0,r)}\frac{e^{-|u|^2/2}}{\sqrt{2\pi}^d}du \geqslant e^{-r^2/2}(2\pi)^{-d/2}\lambda(B(0,r)) \geqslant e^{-r^2/2}(2\pi)^{-d/2}r^d\frac{\pi^{d/2}}{\Gamma(d/2+1)},$$

where we used the direct computation of the Lebesgue volume of the unit ball in $\mathbb{R}^d$

$$\lambda(B(0,1)) = \frac{\pi^{d/2}}{\Gamma(d/2+1)}.$$

Therefore, we obtain that

$$r^2 \leqslant \left(\frac{\gamma(B(0,r))e^{r^2/2}(2\pi)^{d/2}\Gamma(d/2+1)}{\pi^{d/2}}\right)^{2/d} = 2e^{r^2/d}\Gamma(d/2+1)^{2/d}\gamma(B(0,r))^{2/d}.$$

Then, Equation (61) on the volume of shifted balls entails

$$
\begin{aligned}
\forall x \in \mathbb{R}^d \quad \forall r > 0 \qquad r^2 &\leqslant \quad 2e^{r^2/d}\Gamma(d/2+1)^{2/d}\left(\frac{\gamma(B(x,r))e^{\|x\|^2/2} + \gamma(B(x-m,r))e^{\|x-m\|^2/2}}{2}\right)^{2/d} \\
&\leqslant \quad 2e^{r^2/d}\Gamma(d/2+1)^{2/d}\left[\gamma(x)^{-1} + \gamma(x-m)^{-1}\right]^{2/d}\mu(B(x,r))^{2/d}.
\end{aligned}
$$

Using the Stirling formula, we have

$$\Gamma(d/2+1) \leqslant 2\sqrt{2\pi}(d/2+1)^{d/2+1/2}e^{-d/2-1}.$$

We then plug-in this upper bound in the previous inequality and we deduce that:

$$
\begin{aligned}
r^2 &\leqslant \quad 2e^{r^2/d}\frac{d}{2}\left(2\sqrt{2\pi}(1+2/d)^{d/2+1/2}e^{-d/2-1}\right)^{2/d}\left[\gamma(x)^{-1} + \gamma(x-m)^{-1}\right]^{2/d}\mu(B(x,r))^{2/d} \\
&\leqslant \quad de^{r^2/d}\left[\gamma(x)^{-1} + \gamma(x-m)^{-1}\right]^{2/d}\mu(B(x,r))^{2/d}\sup_{d'\geqslant 1}\left\{\left(2\sqrt{2\pi}(1+2/d')^{d'/2+1/2}e^{-d'/2-1}\right)^{2/d'}\right\}.
\end{aligned}
$$

Some straightforward algebra yields:

$$\sup_{d'\geqslant 1}\left\{\left(2\sqrt{2\pi}(1+2/d')^{d'/2+1/2}e^{-d'/2-1}\right)^{2/d'}\right\} \leqslant 72\pi e^{-3} \leqslant 12,$$

which entails that:

$$|\eta(B(x,r)) - \eta(x)| \leqslant 12 d e^{r^2/d} \left[\|x - m\|^2 + \|x\|^2 \right] \left[\gamma(x)^{-1} + \gamma(x-m)^{-1} \right]^{2/d} \mu(B(x,r))^{2/d}.$$

$\square$

C.2 Analysis of the Nearest Neighbor classifier in finite dimension

Below, $\Phi_{k,n}$ refers to the k nearest neighbor classifier given a n sample $\mathcal{D}_n := (X_1, Y_1), \ldots, (X_n, Y_n)$ in $\mathbb{R}^d$ with a Gaussian translation model.

Proof of Proposition 3. We begin with a classical decomposition of the excess risk, we have:

$$\mathcal{R}_{f,g}(\Phi_{k,n,d}) - \mathcal{R}_{f,g}(\Phi_d^\star) = \mathbb{E}\left[|2\eta_d(X) - 1| \mathbb{1}_{\Phi_{k,n,d}(X) \neq \Phi_d^\star(X)} \right].$$

Consider a small ε, whose value will be fixed later on. For any $\delta > 0$, we use the simple lower bound

$$
\begin{aligned}
\mathcal{R}_{f,g}(\Phi_{k,n,d}) - \mathcal{R}_{f,g}(\Phi_d^\star) &\geqslant \mathbb{E}\left[|2\eta_d(X) - 1| \mathbb{1}_{\delta\varepsilon < |\eta(X) - 1/2| < \varepsilon} \mathbb{1}_{\Phi_{k,n,d}(X) \neq \Phi_d^\star(X)} \right], \\
&\geqslant \delta\varepsilon \mathbb{E}\left[\mathbb{1}_{\delta\varepsilon < |\eta(X) - 1/2| < \varepsilon} \mathbb{1}_{\Phi_{k,n,d}(X) \neq \Phi_d^\star(X)} \right], \\
&\geqslant \delta\varepsilon \mathbb{E}_X\left[\mathbb{1}_{\delta\varepsilon < |\eta(X) - 1/2| < \varepsilon} \mathbb{E}_{\otimes n}\left[\mathbb{1}_{\Phi_{k,n}(X) \neq \Phi_d^\star(X)} \right] \right], \\
&\geqslant \delta\varepsilon \mathbb{E}_X\left[\mathbb{1}_{\delta\varepsilon < |\eta(X) - 1/2| < \varepsilon} \mathbb{E}_{\otimes n}\left[\mathbb{1}_{\Phi_{k,n}(X) \neq \Phi_d^\star(X)} \right] \mathbb{1}_{\{\|X\| \leqslant R_d\}} \right],
\end{aligned}
$$

where $R_d := \tau\sqrt{d}$ for some $\tau > 0$. Proposition 2 gives $\beta_d = 2/d$ in our situation. From Proposition 2, the value of L_R given in (25), and the choice of $R = R_d$, we know that a $\tau > 0$ exists such that $L_{R_d} = d$. It is important to notice that R is independent of n.

We now use Lemma 5, Lemma 17 and Lemma 18 of [9]: for any (β_d, L_R)-smooth distribution (see the dependency on β_d in Equation 24), then a constant $\kappa > 0$ exists such that for any k and n:

$$\mathbb{P}_{\otimes n}\left[\Phi_{k,n}(X) \neq \Phi_d^\star(X) \,\middle|\, |\eta(X) - 1/2| \leqslant \frac{1}{\sqrt{k}} - L_{R_d}\left(\frac{k + \sqrt{k} + 1}{n} \right)^{\beta_d} \right] \geqslant \kappa.$$

According to our choice of k_n and R_d, we then have for any $\delta > 0$:

$$
\begin{aligned}
\mathbb{E}\mathcal{R}(\Phi_{k_n,n,d}) - \mathcal{R}(\Phi_d^\star) &\geqslant \kappa\delta\varepsilon \mathbb{E}_X\left[\mathbb{1}_{\delta\varepsilon < |\eta(X) - 1/2| < \varepsilon} \mathbb{1}_{|\eta(X) - 1/2| < \frac{1}{\sqrt{k_n}} - L_R\left(\frac{k_n + \sqrt{k_n} + 1}{n} \right)^\beta} \mathbb{1}_{\{\|X\| \leqslant R_d\}} \right] \\
&\geqslant \kappa\delta\varepsilon \mathbb{E}_X\left[\mathbb{1}_{\delta\varepsilon < |\eta(X) - 1/2| < \varepsilon} \mathbb{1}_{|\eta(X) - 1/2| < \left(\frac{k_n}{n} \right)^{2/d}\left[2d - d\left(1 + k_n^{-1/2} + k_n^{-1}\right)^{2/d} \right]} \mathbb{1}_{\{\|X\| \leqslant R_d\}} \right] \\
&\geqslant \kappa\delta\varepsilon \mathbb{E}_X\left[\mathbb{1}_{\delta\varepsilon < |\eta(X) - 1/2| < \varepsilon} \mathbb{1}_{|\eta(X) - 1/2| < \frac{d}{2}\left(\frac{k_n}{n} \right)^{2/d}} \mathbb{1}_{\{\|X\| \leqslant R_d\}} \right], \quad (64)
\end{aligned}
$$

where we used that $k \leqslant K_n$. To obtain the best achievable lower bound in (64), ε has to be chosen as large as possible. We are driven to the choice (ε depends on n and d):

$$\varepsilon_n = \frac{1}{2} d \left(\frac{k_n}{n} \right)^{2/d}.$$

Then one has for any value of δ smaller than 1:

$$
\begin{aligned}
\mathcal{R}_{f,g}(\Phi_{k,n}) - \mathcal{R}(\Phi_d^\star) &\geqslant c_\delta \varepsilon_n \mathbb{E}_X\left[\mathbb{1}_{\delta\varepsilon_n < |\eta(X) - 1/2| < \varepsilon_n} \mathbb{1}_{\{\|X\| \leqslant R_d\}} \right], \\
&\geqslant c_\delta \varepsilon_n \mathbb{P}_X\left(\{\delta\varepsilon_n < |\eta(X) - 1/2| < \varepsilon_n\} \cap \{\|X\| \leqslant R_d\} \right)
\end{aligned}
$$

Again, we shall use the margin property of the Gaussian translation model: Theorem 5 in Appendix A shows that a δ exists (independent on n) such that

$$\mu\left(\delta t \leqslant \left| \eta(X) - \frac{1}{2} \right| \leqslant t \right) \geqslant \check{c}_\delta \, t,$$

32

where $\check{c}$ is a small enough positive constant. In the same time, there exists a constant C_τ such that

$$\mathbb{P}(\|X\| \leqslant \tau\sqrt{d}) \geqslant C_\tau.$$

The last bound of the excess risk above together with the previous inequality lead to a lower bound of the order ε_n^2: a constant C_1 independent on n and d exists such that

$$\mathbb{E}\mathcal{R}(\Phi_{k,n,d}) - \mathcal{R}(\Phi_d^\star) \geqslant C_1 d^2 \left(\frac{k}{n}\right)^{4/d} \geqslant \frac{C_1}{k}$$

We stress that this lower bound is uniform for any $k \leqslant K_n$ which leads to the desired result. The upper bound involved in the statement of Proposition 3 is a simple consequence of Theorem 4.3 of [14]. □

C.3 Proof of Theorem 3

C.3.1 Technical result

Below, we establish a complementary result with a lower bound on the probability involved in the margin condition. This will make it possible to derive a lower bound of the nearest neighbour classifier.

Proposition 4. *Let X distributed according to the model (1) and for any fixed $\Delta = \|f - g\|_2$, then:*

$$\forall \varepsilon < 1/4 \qquad \mathbb{P}\left(\left|\eta(X) - \frac{1}{2}\right| \leqslant \varepsilon\right) \geqslant (2\pi)^{-1/2}\left[\frac{\varepsilon}{\Delta}e^{-(1+\Delta/2)^2/2} \wedge \frac{e^{-1/2}}{2}\right].$$

Proof. To alleviate the notations, we skip the dependency on X and write $\eta - 1/2 = \frac{q_f - q_g}{2(q_f + q_g)}$. We then repeat the arguments used above:

$$
\begin{aligned}
\mathbb{P}\left(\left|\eta - \frac{1}{2}\right| \leqslant \varepsilon\right) &= \mathbb{P}\left(\frac{|q_f - q_g|}{2(q_f + q_g)} \leqslant \varepsilon\right) \\
&= \mathbb{P}\left(\frac{q_f - q_g}{2(q_f + q_g)} \leqslant \varepsilon, q_f > q_g\right) + \mathbb{P}\left(\frac{q_g - q_f}{2(q_f + q_g)} \leqslant \varepsilon, q_f < q_g\right) \\
&\geqslant \mathbb{P}\left(\frac{q_f - q_g}{2q_f} \leqslant \varepsilon, q_f > q_g\right) + \mathbb{P}\left(\frac{q_g - q_f}{2q_g} \leqslant \varepsilon, q_f < q_g\right) \\
&= \mathbb{P}\left(0 \leqslant 1 - \frac{q_g}{q_f} \leqslant 2\varepsilon\right) + \mathbb{P}\left(0 \leqslant 1 - \frac{q_f}{q_g} \leqslant \varepsilon\right) \\
&= \mathbb{P}\left(\log(1 - 2\varepsilon) \leqslant \log\left(\frac{q_g}{q_f}\right) \leqslant 0\right) + \mathbb{P}\left(\log(1 - 2\varepsilon) \leqslant \log\left(\frac{q_f}{q_g}\right) \leqslant 0\right)
\end{aligned}
$$

We compute a lower bound of the first bound (the second term being handled similarly. For $\varepsilon < 1/4$, it can be checked that $\log(1 - 2\varepsilon) < -\varepsilon$. Therefore, we have

$$\mathbb{P}\left(\log(1 - 2\varepsilon) \leqslant \log\left(\frac{q_g}{q_f}\right) \leqslant 0\right) \geqslant \mathbb{P}\left(-\varepsilon \leqslant \log\left(\frac{q_g}{q_f}\right) \leqslant 0\right)$$

Using again the conditional distribution of $X|Y$ and that Y is distributed according to a Bernoulli distribution $\mathcal{B}(1/2)$, we have

$$\mathbb{P}\left(-\varepsilon \leqslant \log\left(\frac{q_g}{q_f}\right) \leqslant 0\right) = \frac{1}{2}\mathbb{P}\left(-\varepsilon \leqslant \frac{\Delta^2}{2} + \Delta\xi \leqslant 0\right) + \frac{1}{2}\mathbb{P}\left(-\varepsilon \leqslant -\frac{\Delta^2}{2} + \Delta\xi \leqslant 0\right),$$

where $\Delta = \|f - g\|_2$ and ξ is distributed according to $\mathcal{N}(0, 1)$. We can conclude that

$$\mathbb{P}\left(\left|\eta - \frac{1}{2}\right| \leqslant \varepsilon\right) \geqslant \frac{1}{2}\int_{-\frac{\varepsilon}{\Delta} - \frac{\Delta}{2}}^{-\Delta/2} \frac{e^{-t^2/2}}{\sqrt{2\pi}}dt + \frac{1}{2}\int_{-\frac{\varepsilon}{\Delta} + \frac{\Delta}{2}}^{\Delta/2} \frac{e^{-t^2/2}}{\sqrt{2\pi}}dt.$$

Then, we split our study into two cases:

33

- If $\varepsilon \leqslant \Delta$, then $\forall t \in [-\frac{\varepsilon}{\Delta} - \frac{\Delta}{2}, \frac{\Delta}{2}]$ and $\frac{e^{-t^2/2}}{\sqrt{2\pi}} \geqslant \frac{e^{-(1+\Delta/2)^2/2}}{\sqrt{2\pi}}$ and in this case:

$$\mathbb{P}\left(\left|\eta - \frac{1}{2}\right| \leqslant \varepsilon\right) \geqslant \frac{e^{-(1+\Delta/2)^2/2}}{\sqrt{2\pi}}\frac{\varepsilon}{\Delta}$$

- If $\varepsilon > \Delta$,

$$
\begin{aligned}
\mathbb{P}\left(\left|\eta - \frac{1}{2}\right| \leqslant \varepsilon\right) &\geqslant \frac{1}{2}\int_{-\frac{\varepsilon}{\Delta}}^{-\Delta/2} \frac{e^{-t^2/2}}{\sqrt{2\pi}}dt + \frac{1}{2}\int_{-\frac{\varepsilon}{\Delta}}^{0} \frac{e^{-t^2/2}}{\sqrt{2\pi}}dt \\
&\geqslant \int_{-\frac{\varepsilon}{\Delta}}^{-\Delta/2} \frac{e^{-t^2/2}}{\sqrt{2\pi}}dt \\
&\geqslant (2\pi)^{-1/2}\left[\int_{-1}^{0} e^{-t^2/2}dt - \frac{\Delta}{2}\right] \\
&\geqslant \frac{e^{-1/2}}{2\sqrt{2\pi}},
\end{aligned}
$$

where the last bound comes from the fact that $\int_{-1}^0 e^{-t^2/2}dt \geqslant e^{-1/2}$ while $\Delta < \varepsilon < 1/4 < e^{-1/2}$. This ends the proof of the Proposition. $\qquad\square$

A key consequence is the lower bound of the area of the crown $\delta\varepsilon \leqslant |\eta - 1/2| \leqslant \varepsilon$ for δ small enough.

Proposition 5. *Let X given by (1) and for any fixed $\Delta = \|f - g\|_2$, if we set $\delta = \frac{e^{-(1+\Delta/2)^2/2}}{2\sqrt{2\pi}}$, then:*

$$\forall \varepsilon \leqslant \frac{1}{4} \wedge \Delta \qquad \mathbb{P}\left(\delta\varepsilon \leqslant \left|\eta(X) - \frac{1}{2}\right| \leqslant \varepsilon\right) \geqslant \delta\frac{\varepsilon}{\Delta}.$$

Proof. For a given $c > 0$, we introduce $\delta = \frac{e^{-(1+\Delta/2)^2/2}}{c\sqrt{2\pi}}$ and use the decomposition

$$
\begin{aligned}
\mathbb{P}\left(\delta\varepsilon \leqslant \left|\eta(X) - \frac{1}{2}\right| \leqslant \varepsilon\right) &= \mathbb{P}\left(\left|\eta(X) - \frac{1}{2}\right| \leqslant \varepsilon\right) - \mathbb{P}\left(\left|\eta(X) - \frac{1}{2}\right| \leqslant \delta\varepsilon\right) \\
&\geqslant c\delta\frac{\varepsilon}{\Delta} - \mathbb{P}\left(\left|\eta(X) - \frac{1}{2}\right| \leqslant \delta\varepsilon\right),
\end{aligned}
$$

where the last line comes from Proposition 4. Now, we use Proposition 1 to conclude that

$$\mathbb{P}\left(\delta\varepsilon \leqslant \left|\eta(X) - \frac{1}{2}\right| \leqslant \varepsilon\right) \geqslant (c-1)\delta\frac{\varepsilon}{\Delta}.$$

We now choose $c = 2$ and obtain the desired result. $\qquad\square$

Remark 6. *Proposition 5 states that when Δ is small, the measure of the uncertainty area for the classification ($\eta \simeq 1/2$) has an important mass although this measure decreases linearly with the inverse of Δ. This result is intuitive and translates the fact that for large values of Δ, the classification problem is easy (the two classes are well separated) and there is a steep transition from $\{\eta > 1/2\}$ to $\{\eta < 1/2\}$.*

C.3.2 Logarithmic rate of Nearest Neighbor rule

This last paragraph is devoted to the proof of Theorem 3, which shows that a sample splitting strategy used with the NN rule is not efficient with a logarithmic decrease of the misclassification rate.

Proof of Theorem 3. Since the truncation is chosen once for all at the beginning of the classification process with a sample-splitting strategy, our elementary starting point is given by:

$$\mathcal{R}_{f,g}(\widehat{\Phi}_{kNN}^{\widehat{d}}) - \mathcal{R}_{f,g}(\Phi^\star) \geqslant \min_{d \in \mathbb{N}} \mathcal{R}_{f,g}(\widehat{\Phi}_{kNN}^{d}) - \mathcal{R}_{f,g}(\Phi^\star).$$

For any frequency threshold $d \in \mathbb{N}$, we decompose the excess risk as:

$$\mathcal{R}_{f,g}(\Phi_{k,n,d}) - \mathcal{R}_{f,g}(\Phi^\star) = \mathcal{R}_{f,g}(\Phi_{k,n,d}) - \mathcal{R}_{f,g}(\Phi_d^\star) + \mathcal{R}_{f,g}(\Phi_d^\star) - \mathcal{R}_{f,g}(\Phi^\star), \qquad (65)$$

where $\Phi_d^\star$ is the Bayes classification rule with the Gaussian d-dimensional model that involves the first d frequencies. Proposition 3 shows that if $\Delta^2 = \|f - g\|_2^2$, then a constant $c_{\Delta,1}$ exists such that:

$$\mathcal{R}_{f,g}(\Phi_{k,n}) - \mathcal{R}_{f,g}(\Phi_d^\star) \geqslant c_{\Delta,1} n^{-\frac{4}{d+4}}. \qquad (66)$$

We now focus on the second term of (65). Since Y is distributed according to a Bernoulli distribution $\mathcal{B}(1/2)$, we have:

$$\mathcal{R}_{f,g}(\Phi_d^\star) - \mathcal{R}_{f,g}(\Phi^\star) = \frac{1}{2}\left(\mathbb{P}_f[\Phi_d^\star = 1] - \mathbb{P}_f[\Phi^\star = 1]\right) + \frac{1}{2}\left(\mathbb{P}_g[\Phi_d^\star = 0] - \mathbb{P}_g[\Phi^\star = 0]\right).$$

We compute the first term (the second term is handled similarly). Let f, g be fixed function belonging to $\mathcal{H}_s(R)$ which will be made precise latter on. We define $\Delta_d^2 = \|g - f\|_{d,2}^2$ the L^2 norm of $g - f$ restricted to the first d coefficients. If ξ is a standard Gaussian random variable, we have:

$$\mathbb{P}_f[\Phi_d^\star(X) = 1] = \mathbb{P}_f\left[\langle X - f, g - f\rangle_d > \frac{\|g - f\|_{d,2}^2}{2}\right] = \mathbb{P}\left(\xi\Delta_d > \frac{\Delta_d^2}{2}\right)$$

In the meantime, the second probability can be computed as

$$\mathbb{P}_f[\Phi^\star(X) = 1] = \mathbb{P}_f\left[\langle X - f, g - f\rangle > \frac{\|g - f\|_2^2}{2}\right] = \mathbb{P}\left(\xi\Delta > \frac{\Delta^2}{2}\right).$$

Hence, we deduce that

$$\mathbb{P}_f[\Phi_d^\star(X) = 1] - \mathbb{P}_f[\Phi^\star(X) = 1] = \int_{\Delta_d/2}^{\Delta} \gamma(s)ds \geqslant \gamma(\Delta)\frac{\Delta - \Delta_d}{2} = \gamma(\Delta)\frac{\Delta^2 - \Delta_d^2}{2(\Delta + \Delta_d)} \geqslant \frac{\Delta^2 - \Delta_d^2}{4\Delta}\gamma(\Delta).$$

We can then find f and g such that $\Delta^2 < 1$ and $\Delta^2 - \Delta_d \sim d^{-2s}$ because f and g shall belong to the Sobolev space $\mathcal{H}_s(R)$. Hence, we deduce the following lower bound on the excess risk between the truncated Bayes rule and the non parametric Bayes rule: a constant $c_{\Delta,2}$ exists such that

$$\mathbb{P}_f[\Phi_d^\star = 1] - \mathbb{P}_f[\Phi^\star = 1] \geqslant c_{\Delta,2}\, d^{-2s}. \qquad (67)$$

Gathering Equations (66) and (67), we deduce that

$$\mathcal{R}_{f,g}(\widehat{\Phi}_{k,n,\widehat{d}}) - \mathcal{R}_{f,g}(\Phi^\star) \geqslant c_{\Delta,3} \min_{d \in \mathbb{N}^\star}\left[d^{-2s} + n^{-\frac{4}{4+d}}\right].$$

We then optimize our lower bound with respect to d and we obtain the conclusion of the proof. $\qquad \square$

References

[1] C. Abraham, G. Biau, and B. Cadre. On the kernel rule for function classification. *Ann. Inst. Statist. Math*, pages 619–633, 2003.

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

[3] G. Biau and L. Devroye. *Lectures on the nearest neighbor method.* Springer Series in the Data Sciences. Springer, Cham, 2015.

[4] G. Biau and E. Scornet. A random forest guided tour. *TEST*, 25(2):197–227, 2016.

[5] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat.*, 9:323–375, 2005.

[6] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities.* Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[7] B. Cadre. Supervised classification of diffusion paths. *Math. Methods Statist.*, 22(3):213–225, 2013.

[8] F. Cerou and A. Guyader. Nearest neighbor classification in infinite dimension. *ESAIM, P. & S.*, pages 340–355, 2006.

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

[10] T. Chonavel. *Statistical Signal Processing.* Springer-Verlag, New-York, 2002.

[11] T.M. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory*, 13(1):21–27, 1967.

[12] T.M. Cover and J.A. Thomas. *Elements of information theory.* John Wiley & Sons, second edition, 2006.

[13] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York).* Springer-Verlag, New York, 1996.

[14] S. Gadat, T. Klein, and C. Marteau. Classification in general finite dimensional spaces with the k-nearest neighbor rule. *Ann. Statist.*, 44(3):982–1009, 2016.

[15] S. Gerchinovitz, P. Ménard, and G. Stoltz. Fano's inequality for random variables. *arXiv:1702.05985*, 2017.

[16] L. Győrfi. On the rate of convergence of nearest neighbor rules. *IEEE Trans. Inform. Theory*, 24(4):509–512, 1978.

[17] I. Ibragimov and R. Khasminskii. *Statistical Estimation: Asymptotic Theory.* Springer-Verlag, New-York, 1981.

[18] I. A. Ibragimov and R. Z. Has'minskii. *Statistical Estimation: Asymptotic Theory*, volume 16. Springer-Verlag New York, 1981.

[19] N. Ikeda and S. Watanabe. *Stochastic differential equations and diffusion processes*, volume 24 of *North-Holland Mathematical Library.* North-Holland Publishing Co., Amsterdam; Kodansha, Ltd., Tokyo, second edition, 1989.

[20] Y. Ingster. Minimax nonparametric detection of signals in white gaussian noise. *Problems of Information Transmission*, 18:130–140, 1982.

[21] J. Kuelbs, W.V. Li, and W. Linde. The Gaussian measure of shifted balls. *Probab. Theory Related Fields*, 98(2):143–162, 1994.

[22] S.R. Kulkarni and S.E. Posner. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Trans. Inform. Theory*, 41(4):1028–1039, 1995.

[23] D. Lamberton and B. Lapeyre. *Introduction to Stochastic Calculus Applied to Finance.* Chapman and Hall, CRC Press, London, 1996.

[24] R. Lande, S. Engen, and Saether. *Stochastic Populations Dynamics in Ecology and Conservation.* Oxford University Press Inc., New-York, 2003.

[25] B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, 28(5):1302–1338, 2000.

[26] O. Lepskii. On a problem of adaptive estimation in gaussian white noise. *Theory of Probability and its Applications*, 35:454–466, 1990.

[27] E. Mammen and A.B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, 1999.

[28] P. Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007.

[29] P. Massart and E. Nédélec. Risk bounds for statistical learning. *Ann. Statist.*, 34(5):2326–2366, 2006.

[30] A. Rakhlin, K. Sridharan, and A.B. Tsybakov. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2):789–824, 2017.

[31] R. Samworth. Optimal weighted nearest neighbour classifiers. *Ann. Statist.*, 40:2733–2763, 2012.

[32] I. Steinwart and A. Christmann. *Support vector machines.* Information Science and Statistics. Springer, New York, 2008.