Supplement to "Optimal functional supervised classification with separation condition"

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

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

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

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

1. Proof of the minimax lower bound (Theorem 4.1 of [4])

This section contains the proof of our minimax lower bound (Theorem 4.1 of [4]). 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 1.1 below. We will use several key technical ingredients gathered in Section 1.2.

1.1. Proof of Theorem 4.1 of [4]

Our lower bound strategy, in particular the way we reduce the classification problem to an estimation problem, is inspired from [11]. In the finite-dimensional setting, another type of reduction was carried out by [9] and [1].

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

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

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

Second case: $\Delta \geqslant R^{1/(2s+1)}\, n^{-s/(2s+1)}$. 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 $\Theta \subseteq \big\{\theta \in \mathbb{R}^d : \theta_1 = \Delta \text{ and } \sum_{j=2}^{d} \theta_j^2 j^{2s} \leqslant R^2 - \Delta^2\big\}$ 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 of [4].

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 (1.1) in [4], 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 (3.12) of [4]), 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,

$$\begin{aligned}\sup_{\substack{f,g\in\mathcal{H}_s(R)\\ \|f-g\|\geqslant\Delta}}\Big\{\mathcal{R}_{f,g}(\widehat{\Phi})-\inf_{\Phi}\mathcal{R}_{f,g}(\Phi)\Big\}&\geqslant\sup_{\theta\in\Theta}\Big\{\mathcal{R}_{f_\theta,0}(\widehat{\Phi})-\inf_{\Phi}\mathcal{R}_{f_\theta,0}(\Phi)\Big\}\\&=\sup_{\theta\in\Theta}\mathbb{E}_\theta\Big[\big|2\eta_\theta(X)-1\big|\mathbb{1}_{\widehat{\Phi}(X)\neq\Phi_\theta(X)}\Big]\,,\end{aligned}\tag{1.1}$$

where $\eta_\theta(x)=\mathbb{P}_\theta(Y=1|X=x)$ denotes the regression function corresponding to the statistical model (1.1) in [4] 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

$$\begin{aligned}\mathbb{E}_\theta\Big[\big|2\eta_\theta(X)-1\big|\mathbb{1}_{\widehat{\Phi}(X)\neq\Phi_\theta(X)}\Big]&\geqslant\delta\,\mathbb{P}_\theta\Big(\big\{|2\eta_\theta(X)-1|\geqslant\delta\big\}\cap\big\{\widehat{\Phi}(X)\neq\Phi_\theta(X)\big\}\Big)\\&\geqslant\delta\,\Big(\mathbb{P}_\theta\big(\widehat{\Phi}(X)\neq\Phi_\theta(X)\big)-\mathbb{P}_\theta\big(|2\eta_\theta(X)-1|<\delta\big)\Big)\\&\geqslant\delta\,\left(\mathbb{P}_\theta\big(\widehat{\Phi}(X)\neq\Phi_\theta(X)\big)-\frac{5\delta}{\Delta}\right),\end{aligned}\tag{1.2}$$

where the last inequality follows from Proposition 1 of [4]. 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 (1.1) of [4] 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 $\theta\in\Theta$,

$$\begin{aligned}\mathbb{P}_\theta\Big(\widehat{\Phi}(X)\neq\Phi_\theta(X)\Big)&=\mathbb{E}_\theta\Big\{\mathbb{P}_\theta\Big(\widehat{\Phi}(X)\neq\Phi_\theta(X)\,\Big|\,(X_i,Y_i)_{1\leqslant i\leqslant n}\Big)\Big\}\\&=\mathbb{E}_\theta\left\{\int\mathbb{1}_{\widehat{\Phi}(x)\neq\Phi_\theta(x)}\frac{\mathrm{d}Q_0(x)+\mathrm{d}Q_\theta(x)}{2}\right\}\\&\geqslant\frac{1}{2}\,\mathbb{E}_\theta\Big[\big\|\widehat{\Phi}-\Phi_\theta\big\|_{L^1(Q_0)}\Big]\,,\end{aligned}\tag{1.3}$$

where the last inequality follows from the fact that $\mathbb{1}_{\widehat{\Phi}(x)\neq\Phi_\theta(x)}=\big|\widehat{\Phi}(x)-\Phi_\theta(x)\big|$ for all continuous functions $x:[0,1]\to\mathbb{R}$. Putting (1.1), (1.2), and (1.3) together, we finally get

$$\sup_{\substack{f,g\in\mathcal{H}_s(R)\\ \|f-g\|\geqslant\Delta}}\Big\{\mathcal{R}_{f,g}(\widehat{\Phi})-\inf_{\Phi}\mathcal{R}_{f,g}(\Phi)\Big\}\geqslant\frac{\delta}{2}\,\left(\sup_{\theta\in\Theta}\mathbb{E}_\theta\Big[\big\|\widehat{\Phi}-\Phi_\theta\big\|_{L^1(Q_0)}\Big]-\frac{10\delta}{\Delta}\right).\tag{1.4}$$

Step 2: a key combinatorial and geometrical argument In order to further bound (1.4) from below, we now specialize Θ to the set given by Lemma 1 in Appendix 1.2, whose proof combines Varshamov-Gilbert's lemma with simple but key geometrical arguments in dimension two. More precisely, we use Lemma 1 in Appendix 1.2 with $\varepsilon = c/\sqrt{n}$ and $d = \big\lfloor \big((R^2 - \Delta^2)\, n\big)^{1/(2s+1)} \big\rfloor$, for some absolute constant $c \in (0, 1]$ to be determined later. Two remarks are in order:

- We have $d \geqslant \big((R^2 - \Delta^2)n\big)^{1/(2s+1)} - 1 \geqslant 32\log(2) + 1$ 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 1 holds true.
- The condition $\Delta \geqslant \sqrt{d}\,\varepsilon$ of Lemma 1 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 1 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 $\theta \neq \theta' \in \Theta$,

$$\big\|\Phi_\theta - \Phi_{\theta'}\big\|_{L^1(Q_0)} \geqslant \frac{\sqrt{d-1}\,\varepsilon}{4\pi\Delta}\, e^{-\Delta^2}\,. \tag{1.5}$$

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 $d \leqslant \big((R^2 - \Delta^2)\, n\big)^{1/(2s+1)}$ and $\varepsilon \leqslant 1/\sqrt{n}$. Therefore, $\Theta \subseteq \big\{\theta \in \mathbb{R}^d : \theta_1 = \Delta \text{ and } \sum_{j=2}^d \theta_j^2 j^{2s} \leqslant R^2 - \Delta^2\big\}$ 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 (1.4) and setting

$$\widehat{\theta} \in \operatorname*{arg\,min}_{\theta\in\Theta} \big\|\widehat{\Phi} - \Phi_\theta\big\|_{L^1(Q_0)}\,,$$

we can see that

$$\begin{aligned}\sup_{\substack{f,g\in\mathcal{H}_s(R)\\ \|f-g\|\geqslant\Delta}} \Big\{\mathcal{R}_{f,g}(\widehat{\Phi}) - \inf_{\Phi} \mathcal{R}_{f,g}(\Phi)\Big\} &\geqslant \frac{\delta}{2}\left(\sup_{\theta\in\Theta} \mathbb{E}_\theta\Big[\mathbb{1}_{\{\widehat{\theta}\neq\theta\}} \big\|\widehat{\Phi} - \Phi_\theta\big\|_{L^1(Q_\mu)}\Big] - \frac{10\delta}{\Delta}\right) \\ &\geqslant \frac{\delta}{2}\left(\frac{\sqrt{d-1}\,\varepsilon}{8\pi\Delta}\, e^{-\Delta^2} \sup_{\theta\in\Theta} \mathbb{P}_\theta\Big(\widehat{\theta} \neq \theta\Big) - \frac{10\delta}{\Delta}\right)\,, \end{aligned} \tag{1.6}$$

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

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

by a combination of Inequality (1.5), the definition of $\widehat{\theta}$, and the triangle inequality.

We now lower bound the worst-case testing error $\sup_{\theta\in\Theta} \mathbb{P}_\theta\big(\widehat{\theta} \neq \theta\big)$. 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 $\mathbb{P}_\theta\big(\widehat{\theta} \neq \theta\big) = P_\theta\big(\widehat{\theta} \neq \theta\big)$. We can thus use Fano's inequality (cf. Lemma 6 in Appendix 1.2.3) with the events $A_\theta = \big\{\widehat{\theta} = \theta\big\}$, the 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:

$$\inf_{\theta\in\Theta} P_\theta\big(\widehat{\theta} = \theta\big) \leqslant \frac{1}{|\Theta|} \sum_{\theta\in\Theta} P_\theta\big(\widehat{\theta} = \theta\big) \leqslant \frac{\dfrac{1}{|\Theta|}\displaystyle\sum_{\theta\in\Theta} \mathrm{KL}\big(P_\theta, P_{\theta_0}\big) + \log 2}{\log|\Theta|}\,. \tag{1.7}$$

Using the chain rule for the Kullback-Leibler divergence, and following similar computations as in Section 2 of [4] (application of Girsanov's formula), we can see that, for all $\theta \in \Theta$,

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

where we used the fact that $\theta \in \Theta \subseteq \{\Delta\} \times \{-\varepsilon, \varepsilon\}^{d-1}$ and $\theta_0 := (\Delta, 0, \dots, 0)$. Combining (1.7) 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 (1.6), we finally get

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

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 4.1 of [4].

1.2. 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 4.1 of [4]. 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 1. *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 $\theta \neq \theta' \in \Theta$,*

$$\big\|\Phi_\theta - \Phi_{\theta'}\big\|_{L^1(Q_0)} \geqslant \frac{\sqrt{d-1}\,\varepsilon}{4\pi\Delta}\, e^{-\Delta^2}\,, \tag{1.8}$$

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 1.2.2 below. We first state three intermediary results.

1.2.1. Intermediary results

The following lemma shows that, for the d-dimensional construction of Section 1.1 (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 2. *Consider the statistical construction of Section 1.1 (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 $\widetilde{W} := \big(\widetilde{W}_j\big)_{1\leqslant j\leqslant d} \in \mathbb{R}^d$. Then, the Bayes classifier $\Phi_\theta = \mathbb{1}_{\{\eta_\theta \geqslant 1/2\}}$ satisfies*

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

Proof. The result follows directly from the calculations of Section 2.1 of [4] (application of Girsanov's formula). Indeed, using (2.3) of [4] and the fact that $g = 0$ and $\|f_\theta\| = \|\theta\|$, we obtain

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

which concludes the proof. □

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 1.

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 3. *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 5 in Appendix 1.2.3), and define*

$$\Theta := \big\{\Delta\big\} \times \big(\varepsilon\Gamma\big) = \Big\{(\Delta, \varepsilon u_1, \varepsilon u_2, \ldots, \varepsilon u_{d-1}) : (u_1, \ldots, u_{d-1}) \in \Gamma\Big\} \subset \mathbb{R}^d \,. \tag{1.9}$$

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

$$\cos\Big(\angle(\theta,\, \theta')\Big) = \frac{\langle \theta, \theta'\rangle}{\|\theta\|\,\|\theta'\|} = \frac{\Delta^2 + \varepsilon^2 \sum_{j=1}^m u_j u'_j}{\sqrt{\Delta^2 + m\varepsilon^2}\,\sqrt{\Delta^2 + m\varepsilon^2}} = \frac{\Delta^2 + \varepsilon^2 \sum_{j=1}^m u_j u'_j}{\Delta^2 + m\varepsilon^2} \,. \tag{1.10}$$

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, $\cos\big(\angle(\theta,\, \theta')\big) \geqslant 0$, 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 5 in Appendix 1.2.3), 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 (1.10) yields

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

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

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

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 $\angle(\theta,\,\theta') \geqslant \sin\big(\angle(\theta,\,\theta')\big) = \sqrt{\sin^2\big(\angle(\theta,\,\theta')\big)}$ since $\angle(\theta,\,\theta') \in [0,\pi]$:

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

□

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

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

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 $\angle(u,v) := \arccos\big(\langle u,v\rangle/(\|u\|\,\|v\|)\big) \in (0,\pi)$ 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 4. *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 $\mathcal{C} - z = \big\{x - z :\, x \in \mathcal{C}\big\}$ and using the inequality $\|z+u\|^2 \leqslant 2\|z\|^2 + 2\|u\|^2$, we get

$$\begin{aligned}
\gamma_2(\mathcal{C}) &= \frac{1}{2\pi}\int_{\mathcal{C}} e^{-\|x\|^2/2}\,\mathrm{d}x = \frac{1}{2\pi}\int_{\mathcal{C}-z} e^{-\|z+u\|^2/2}\,\mathrm{d}u \geqslant \frac{e^{-\|z\|^2}}{2\pi}\int_{\mathcal{C}-z} e^{-\|u\|^2}\mathrm{d}u \\
&= \frac{e^{-\|z\|^2}}{2\pi}\, 2\int_0^{\mathcal{A}} \left(\int_0^{+\infty} r e^{-r^2}\mathrm{d}r\right)\mathrm{d}\alpha = \frac{e^{-\|z\|^2}}{2\pi}\,\mathcal{A}\,,
\end{aligned}$$

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. □

1.2.2. Proof of Lemma 1

We now prove Lemma 1 using the intermediary results of the previous subsection. We use the same notation as in Section 1.1. Let $\Gamma \subseteq \{-1,1\}^{d-1}$ be a set provided by Varshamov-Gilbert's lemma in dimension $m = d-1$ (cf. Lemma 5 in Appendix 1.2.3). Next we show that the set

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

satisfies the statement of Lemma 1. 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 $\theta \neq \theta' \in \Theta$,

$$\big\|\Phi_\theta - \Phi_{\theta'}\big\|_{L^1(Q_0)} \geqslant \frac{\sqrt{d-1}\,\varepsilon}{4\pi\Delta}\, e^{-\Delta^2}\,, \tag{1.11}$$

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

Proof of (1.11)**.** 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 2, 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 3 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.

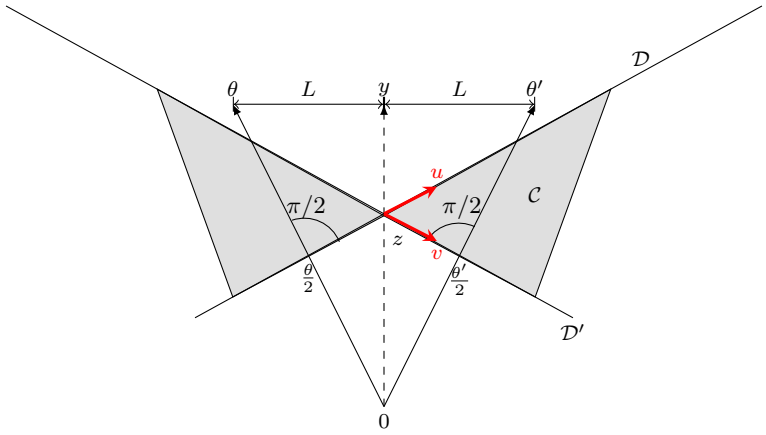

Figure 1. The main objects of interest on the plane $\mathcal{P}$.

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

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

Therefore,

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

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 4 and then Lemma 3,

$$\mathbb{P}(A) \;\geqslant\; \frac{\angle(\theta,\theta')}{2\pi}\, e^{-(z_1^2+z_2^2)} \geqslant \frac{e^{-(z_1^2+z_2^2)}\sqrt{d-1}\,\varepsilon}{4\pi\Delta}\,. \tag{1.12}$$

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 3) 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 (1.12) concludes the proof. □

1.2.3. 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 $\{-1,1\}^m$, exponentionally many of them are almost opposite from one another. A proof can be found, e.g., in [10, Lemma 4.7].

Lemma 5 (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 [7, Chapter VII, Lemma 1.1] or [3, Theorem 2.11.1] (see also Proposition 1 in the recent survey [6]).

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

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

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

2. Truncated nearest neighbor strategy (Theorem 4.2 of [4])

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.

2.1. 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 (4.2) of [4]). 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 of [4]. According to the definition of the smoothness parameter given in Equation (4.2) of [4], we compute the average value of η on a ball $B(x,r)$ and compare it to $\eta(x)$:

$$\begin{aligned}
&\eta(B(x,r)) - \eta(x) \\
&= \frac{1}{\mu(B(x,r)}\int_{B(x,r)} \eta(s)d\mu(s) - \frac{\gamma(x)}{\gamma(x)+\gamma(x-m)}, \\
&= \frac{2}{\int_{B(x,r)}\gamma(s)+\gamma(s-m)ds}\int_{B(x,r)}\frac{\gamma(s)}{\gamma(s)+\gamma(s-m)}\frac{1}{2}[\gamma(s)+\gamma(s-m)]ds - \frac{\gamma(x)}{\gamma(x)+\gamma(x-m)}, \\
&= \frac{\gamma(B(x,r))}{\gamma(B(x,r))+\gamma(B(x-m,r))} - \frac{\gamma(x)}{\gamma(x)+\gamma(x-m)}, \\
&= \frac{[\gamma(x)+\gamma(x-m)]\gamma(B(x,r)) - \gamma(x)[\gamma(B(x,r))+\gamma(B(x-m,r))]}{[\gamma(x)+\gamma(x-m)][\gamma(B(x,r))+\gamma(B(x-m,r))]}, \\
&= \frac{\gamma(x-m)\gamma(B(x,r)) - \gamma(x)\gamma(B(x-m,r))}{[\gamma(x)+\gamma(x-m)][\gamma(B(x,r))+\gamma(B(x-m,r))]}. \qquad (2.1)
\end{aligned}$$

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:

$$\exp(-\|x\|^2/2)\gamma(B(0,r)) \leqslant \gamma(B(x,r)) \leqslant \gamma(B(0,r)). \qquad (2.2)$$

In particular, we have (see [8]) when $r \longrightarrow 0$ that

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

but the r.h.s. of (2.2) is tight only for x close to 0. Expanding the denominator of (2.1), we obtain that

$$\begin{aligned}
&|\eta(B(x,r)) - \eta(x)| \\
&= \frac{|\gamma(x-m)\gamma(B(x,r)) - \gamma(x)\gamma(B(x-m,r))|}{\gamma(x)\gamma(B(x,r)) + \gamma(x)\gamma(B(x-m,r)) + \gamma(x-m)\gamma(B(x,r)) + \gamma(x-m)\gamma(B(x-m,r))} \\
&\leqslant \frac{|\gamma(x-m)\gamma(B(x,r)) - \gamma(x)\gamma(B(x-m,r))|}{\gamma(x)\gamma(B(x-m,r)) + \gamma(x-m)\gamma(B(x,r))}. \qquad (2.3)
\end{aligned}$$

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

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

where the last line comes from the symmetry of the Gaussian distribution. Using this last inequality in Inequality (2.3) yields:

$$|\eta(B(x,r)) - \eta(x)| \leqslant r^2\left[\|x-m\|^2 + \|x\|^2\right]. \tag{2.4}$$

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 (2.2) on the volume of shifted balls entails

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

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:

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

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:

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

□

2.2. 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 of [4]. 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 of [4] gives $\beta_d = 2/d$ in our situation. From Proposition 2 of [4], the value of L_R given in (4.3) of [4], 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 [2]: for any (β_d, L_R)-smooth distribution (see the dependency on β_d in Equation (4.2) of [4]), 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 $\delta > 0$:

$$\begin{aligned}
\mathbb{E}\mathcal{R}(\Phi_{k_n,n,d}) - \mathcal{R}(\Phi_d^\star) &\geqslant \kappa\delta\varepsilon\mathbb{E}_X\left[\mathbb{1}_{\{\delta\varepsilon<|\eta(X)-1/2|<\varepsilon\}}\mathbb{1}_{\{|\eta(X)-1/2|<\frac{1}{\sqrt{k_n}}-L_R\left(\frac{k_n+\sqrt{k_n}+1}{n}\right)^\beta\}}\mathbb{1}_{\{\|X\|\leqslant R_d\}}\right] \\
&\geqslant \kappa\delta\varepsilon\mathbb{E}_X\left[\mathbb{1}_{\{\delta\varepsilon<|\eta(X)-1/2|<\varepsilon\}}\mathbb{1}_{\{|\eta(X)-1/2|<\left(\frac{k_n}{n}\right)^{2/d}\left[2d-d\left(1+k_n^{-1/2}+k_n^{-1}\right)^{2/d}\right]\}}\mathbb{1}_{\{\|X\|\leqslant R_d\}}\right] \\
&\geqslant \kappa\delta\varepsilon\mathbb{E}_X\left[\mathbb{1}_{\{\delta\varepsilon<|\eta(X)-1/2|<\varepsilon\}}\mathbb{1}_{\{|\eta(X)-1/2|<\frac{d}{2}\left(\frac{k_n}{n}\right)^{2/d}\}}\mathbb{1}_{\{\|X\|\leqslant R_d\}}\right], \qquad (2.5)
\end{aligned}$$

where we used that $k \leqslant K_n$. To obtain the best achievable lower bound in (2.5), 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 2 shows that a δ exists (independent on n) such that

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

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. □

Finally, we emphasize that we can easily derive an upper bound associated with the statement of Proposition 3 of [4]. A straightforward application of Theorem 4.3 of [5] in our setting yields a $\log(n)^{-2s}$ upper bound for the rate of convergence of the misclassification of the kNN.

2.3. Proof of Theorem 4.2 of [4]

2.3.1. 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 1. *Let X distributed according to the model (1.1) of [4] 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:

- 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. □

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 2. *Let X given by (1.1) of [4] 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 1. Now, we use Proposition 1 in [4] 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. □

***Remark* 2.1.** Proposition 2 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 $\{\eta < 1/2\}$.

2.3.2. Logarithmic rate of Nearest Neighbor rule

This last paragraph is devoted to the proof of Theorem 4.2 in [4], 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 4.2 in [4]. 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:

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

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

$$\mathcal{R}_{f,g}(\Phi_{k,n,d}) - \mathcal{R}_{f,g}(\Phi^\star) = \mathcal{R}_{f,g}(\Phi_{k,n,d}) - \mathcal{R}_{f,g}(\Phi^\star_d) + \mathcal{R}_{f,g}(\Phi^\star_d) - \mathcal{R}_{f,g}(\Phi^\star), \tag{2.6}$$

where $\Phi^\star_d$ is the Bayes classification rule with the Gaussian d-dimensional model that involves the first d frequencies. Proposition 3 of [4] shows that if $\Delta^2 = \|f - g\|_2^2$, then a constant $c_{\Delta,1}$ exists such that:

$$\mathcal{R}_{f,g}(\Phi_{k,n}) - \mathcal{R}_{f,g}(\Phi^\star_d) \geqslant c_{\Delta,1} n^{-\frac{4}{d+4}}. \tag{2.7}$$

We now focus on the second term of (2.6). Since Y is distributed according to a Bernoulli distribution $\mathcal{B}(1/2)$, we have:

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

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:

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

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

$$\mathbb{P}_f[\Phi_d^\star = 1] - \mathbb{P}_f[\Phi^\star = 1] \geqslant c_{\Delta,2}\, d^{-2s}. \tag{2.8}$$

Gathering Equations (2.7) and (2.8), 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. □

References

[1] T. Cai and L. Zhang. High-dimensional linear discriminant analysis: optimality, adaptive algorithm and missing data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(4):675–705, 2019.

[2] 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.

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

[4] S. Gadat, S. Gerchinovitz, and C. Marteau. Optimal functional supervised classification with separation condition. *Submitted to Bernoulli*.

[5] 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.

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

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

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

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

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

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