Sankhyā A : The Indian Journal of Statistics
https://doi.org/10.1007/s13171-025-00386-w

Isotropic Kernel Machine

Yann Guermeur
Campus Scientifique, Vandoeuvre-lès-Nancy Cedex, France
Nicolas Wicker
University of Lille, Villeneuve d'Ascq, France

Abstract

A new kernel machine for multi-class pattern recognition is introduced: the isotropic kernel machine. It is designed to make use of the isotropy of the class conditional densities in the feature space. We provide theoretical guarantees on its generalization error. This error is then assessed empirically, in the framework of a comparative study.

AMS (2000) subject classification. Primary 62H30; Secondary 68Q32.
Keywords and phrases. Margin multi-category classifiers, kernel machines, isotropy.

1 Introduction

The support vector machine (SVM), introduced by Cortes and Vapnik (1995), is the first and main kernel machine (Schölkopf and Smola 2002) for pattern classification. Over the last two decades, a great many multi-class extensions (M-SVMs) have been introduced (see Guermeur, 2012; Dogan et al., 2016 for a survey). They all share one basic feature: their decision boundaries are linear in the reproducing kernel Hilbert space (RKHS) (Berlinet and Thomas-Agnan 2004) of the kernel. The underlying idea is appealing: get the best of two worlds by combining the theoretical guarantees attached to linear classifiers with the gain of capacity induced by the kernelization. However, it is not without drawbacks, and recent studies have highlighted the fact that neural networks (Anthony and Bartlett 1999) could outperform kernel machines, for instance on classification tasks involving data with a low-dimensional representation. The phenomenon is especially noticeable when the descriptions are realizations of nearly isotropic random vectors (see for instance Ghorbani et al., 2020). Fortunately, it remains relevant to consider more complex decision boundaries in the RKHS. The rationale for this statement is the efficiency of famous quadratic classifiers,

Published online: 06 May 2025

such as (Fisher) quadratic discriminant analysis (QDA) (Hastie et al. 2008), and their kernelized extensions (Wang et al. 2008).

This article introduces a new multi-class kernel machine which aims at exploiting the isotropy of the class conditional distributions in the feature space. The isotropic kernel machine (IKM) can be seen as a kernelized extension of a linear classifier: the nearest centroid classifier (NCC). Like the quadratic classifiers, its decision boundaries in the feature space are nonlinear functions. Unlike them, its learning problem does not involve a parametric model of the data. We provide theoretical guarantees on its generalization error. This error is then assessed empirically, in the framework of a comparative study.

The paper is organized as follows. Section 2 is devoted to the definition and characterization of the new machine. The guarantees on its risk are provided in Section 3. Section 4 is devoted to the comparative study. At last, we draw conclusions and outline our ongoing research in Section 5. To make reading easier, all technical lemmas and proofs have been gathered in appendix.

2 The Machine

The new classifier is devised in the following theoretical framework.

2.1 Theoretical Framework We consider C-category pattern classification problems in the most general framework for empirical inference: *agnostic learning* (Kearns et al. 1994). Let $\mathcal{X}$ denote the description space and $\mathcal{Y}$ the set of categories. Since no specific hypotheses are made regarding the structure of $\mathcal{Y}$, this set is identified with the set of indices of the categories, i.e., the set of the integers ranging from 1 to C, hereafter denoted by $[\![1;C]\!]$. We denote by $(\mathcal{X},\mathcal{A}_{\mathcal{X}})$ and $(\mathcal{Y},\mathcal{A}_{\mathcal{Y}})$ the basic measurable spaces and by $\mathcal{A}_{\mathcal{X}}\otimes\mathcal{A}_{\mathcal{Y}}$ the tensor-product sigma-algebra on the Cartesian product $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$. P is the unknown probability measure on the measurable space $(\mathcal{X}\times\mathcal{Y},\mathcal{A}_{\mathcal{X}}\otimes\mathcal{A}_{\mathcal{Y}})$. Let $Z=(X,Y)$ be a random pair distributed according to P. The only access to P is via the m-sample $\mathbf{Z}_m=(Z_i)_{1\leqslant i\leqslant m}=((X_i,Y_i))_{1\leqslant i\leqslant m}$ made up of independent copies of Z (in short $\mathbf{Z}_m\sim P^m$).

A classifier is characterized by a triplet made up of a function class, a decision rule and an inductive principle. We now introduce the new kernel machine through the specification of the corresponding triplet.

2.2 Function Class and Decision Boundaries The architecture of the new machine is devised to capture the isotropy of the class conditional densities so as to fit the data more accurately than the kernelized linear classifiers,

while involving fewer parameters than the kernelized nonlinear classifiers. In the sequel, κ is a real-valued positive type function/kernel (Berlinet and Thomas-Agnan 2004) on $\mathcal{X}^2$ and $\left(\mathbf{H}_\kappa, \langle\cdot,\cdot\rangle_{\mathbf{H}_\kappa}\right)$ is its RKHS. Note that κ needs not be isotropic (Genton 2001).

Definition 1 (Function classes $\mathcal{H}_p$ and $\mathcal{H}$) Let κ be a kernel and $p \in [1,2]$. The function class $\mathcal{H}_p$ is the class of all real-valued functions h on $\mathcal{X}$ of the form

$$\forall x \in \mathcal{X},\ \ h\left(x\right) = R - a\left\|O - \kappa_x\right\|_{\mathbf{H}_\kappa}^p,$$

where $O \in \mathbf{H}_\kappa$, $R \in \mathbb{R}_+$ and $a \in \mathbb{R}_+^*$. Then, the function class at the basis of a C-category IKM is the class $\mathcal{H} = \bigcup_{p\in[1,2]} \mathcal{H}_p^C$.

The dedication to the exploitation of isotropy rests on the fact that for every function in $\mathcal{H}$, the level surfaces of the component functions (associated with the different categories) are hyperspheres of the RKHS. All these component functions are associated with the same value of p. The reason for this simplification is two-fold. On the one hand, we could not identify a (real-world) problem calling for a more general choice. On the other hand, the restriction is highly beneficial to the capacity control.

Definition 2 (Decision rule) For every function $h = (h_k)_{1\leqslant k\leqslant C} \in \mathcal{H}$, a *decision rule* dr_h is specified in the following way:

$$\forall x \in \mathcal{X},\ \ \begin{cases} \left|\underset{1\leqslant k\leqslant C}{\operatorname{argmax}}\, h_k\left(x\right)\right| = 1 \Longrightarrow \mathrm{dr}_h\left(x\right) = \underset{1\leqslant k\leqslant C}{\operatorname{argmax}}\, h_k\left(x\right) \\ \left|\underset{1\leqslant k\leqslant C}{\operatorname{argmax}}\, h_k\left(x\right)\right| > 1 \Longrightarrow \mathrm{dr}_h\left(x\right) = * \end{cases}$$

where $|\cdot|$ returns the cardinality of its argument and $*$ stands for a dummy category.

Definitions 1 and 2 make it clear that the IKM is a (kernelized) extension of the NCC. Indeed, let the function h of $\mathcal{H}$ be characterized by $p \in [1,2]$, the vectors $\mathbf{O}_C = (O_k)_{1\leqslant k\leqslant C} \in \left(\mathbf{H}_\kappa\right)^C$, $\mathbf{R}_C = (R_k)_{1\leqslant k\leqslant C} \in \left(\mathbb{R}_+\right)^C$ and $\mathbf{a}_C = (a_k)_{1\leqslant k\leqslant C} \in \left(\mathbb{R}_+^*\right)^C$. Then if $\mathcal{X}$ is included in a Euclidean space, a function h implementing the NCC is given by $p \in [1,2]$,

$$\mathbf{O}_C = \left(\frac{1}{|\mathcal{Y}_k|}\sum_{\{i:\ Y_i=k\}} X_i\right)_{1\leqslant k\leqslant C}$$

where $\mathcal{Y}_k = \{i : Y_i = k\}$, $\mathbf{R}_C = 0_C$ and $\mathbf{a}_C = \mathbb{1}_C$. When $\mathcal{X} \subset \mathbb{R}^d$, an example of isotropic class conditional distributions for which the classifier can produce the Bayes decision boundaries is the following one:

$$\forall k \in [\![1; C]\!], \quad f_k(x) = \frac{\Gamma\left(\frac{d}{2}\right)}{\Gamma(d)} \frac{1}{2\left(\sqrt{\pi} b_k\right)^d} \exp\left(-\frac{\|x - \mu_k\|_2}{b_k}\right), \tag{1}$$

where Γ stands for Euler's Gamma function. Then, denoting by $\{\pi_k : 1 \leqslant k \leqslant C\}$ the set of prior probabilities, the boundary between categories k and l computed by the Bayes classifier is given by:

$$\frac{\|x - \mu_k\|_2}{b_k} - \frac{\|x - \mu_l\|_2}{b_l} = \ln\left(\frac{\pi_k}{\pi_l}\left(\frac{b_l}{b_k}\right)^d\right).$$

Consequently, setting $K = \min_{1 \leqslant l \leqslant C} \ln\left(\frac{\pi_l}{b_l^d}\right)$, the IKM will be optimal for a linear kernel (Euclidean dot product), with $p^* = 1$, $\mathbf{O}_C^* = (\mu_k)_{1 \leqslant k \leqslant C}$, $\mathbf{R}_C^* = \left(\ln\left(\frac{\pi_k}{b_k^d}\right) - K\right)_{1 \leqslant k \leqslant C}$ and $\mathbf{a}_C^* = \left(b_k^{-1}\right)_{1 \leqslant k \leqslant C}$.

Let us now consider the case when the class conditional densities are normal, and given by:

$$\forall k \in [\![1; C]\!], \quad f_k(x) = \frac{1}{\left(\sqrt{2\pi}\sigma_k\right)^d} \exp\left(-\frac{\|x - \mu_k\|_2^2}{2\sigma_k^2}\right). \tag{2}$$

This time, the analytical expression of the optimal decision boundary between categories k and l is:

$$\frac{\|x - \mu_k\|_2^2}{2\sigma_k^2} - \frac{\|x - \mu_l\|_2^2}{2\sigma_l^2} = \ln\left(\frac{\pi_k}{\pi_l}\left(\frac{\sigma_l}{\sigma_k}\right)^d\right).$$

Consequently, setting $K = \min_{1 \leqslant l \leqslant C} \ln\left(\frac{\pi_l}{\sigma_l^d}\right)$, the IKM will be optimal for a linear kernel, with $p^* = 2$, $\mathbf{O}_C^* = (\mu_k)_{1 \leqslant k \leqslant C}$, $\mathbf{R}_C^* = \left(\ln\left(\frac{\pi_k}{\sigma_k^d}\right) - K\right)_{1 \leqslant k \leqslant C}$ and $\mathbf{a}_C^* = \left(\left(2\sigma_k^2\right)^{-1}\right)_{1 \leqslant k \leqslant C}$.

The meaning of the vector $\mathbf{R}_C$ can appear more clearly if the class conditional distributions have finite support. A good illustration is provided by the following elementary situation, where the data live in $\mathbb{R}^2$ and follow truncated multivariate normal distributions whose supports are balls

$\mathcal{B}_k$ centered on the corresponding means μ_k. To avoid degenerate situations, their radii r_k ars supposed to be large enough so that they intersect one another. To simplify further, the prior probabilities and the standard deviations σ_k are supposed to be all equal, to C^{-1} and $2^{-\frac{1}{2}}$ respectively. As a consequence, the basic density functions $\tilde{f}_k$ are given by:

$$\forall k \in [\![1; C]\!], \quad \tilde{f}_k(x) = \frac{1}{\pi} \exp\left(-\|x - \mu_k\|_2^2\right),$$

and denoting

$$\forall k \in [\![1; C]\!], \quad Z_k = \int_{\mathcal{B}_k} \tilde{f}_k(x)\, \mathrm{d}x,$$

their truncated variants f_k take the form

$$\forall k \in [\![1; C]\!], \quad \begin{cases} f_k(x) = \frac{1}{Z_k} \tilde{f}_k(x) & \text{if } x \in \mathcal{B}_k \\ f_k(x) = 0 & \text{otherwise} \end{cases}.$$

Since the integrals Z_k do not depend on the means μ_k, applying the transform to polar coordinates produces:

$$\begin{aligned} \forall k \in [\![1; C]\!], \quad Z_k &= \frac{1}{\pi} \int_0^{2\pi} \int_0^{r_k} e^{-r^2} r \mathrm{d}r \mathrm{d}\theta \\ &= \left[-e^{-r^2}\right]_0^{r_k} \\ &= 1 - e^{-r_k^2}. \end{aligned}$$

As a consequence, the analytical expression of the optimal decision boundary between categories k and l is:

$$\|x - \mu_k\|_2^2 - \|x - \mu_l\|_2^2 = \ln\left(\frac{1 - e^{-r_l^2}}{1 - e^{-r_k^2}}\right).$$

Once more, the Bayes classifier is a function in the class $\mathcal{H}$. It corresponds to a linear kernel, with $p^* = 2$, $\mathbf{O}_C^* = (\mu_k)_{1 \leqslant k \leqslant C}$, $\mathbf{R}_C^* = \left(\ln\left(\frac{1-e^{-r_{\max}^2}}{1-e^{-r_k^2}}\right)\right)_{1 \leqslant k \leqslant C}$ and $\mathbf{a}_C^* = \mathbb{1}_C$. Not only are the values of the parameters R_k^* different from the squares of the radii r_k, but they even vary in the opposite way. Generally speaking, there is no (direct) connections between the component functions of the classifier and the corresponding minimum enclosing balls.

2.3 Function Selection We have stated in introduction that the function selection does not rely on a parametric model of the distributions of the populations. The inferential principle implemented to derive the classifier from the data simply consists in minimizing a data-fit term based on the notion of (analytical) margin. This calls for the selection of a margin loss function. We use the parameterized truncated hinge loss, applied to the margin functions, a choice that bears the advantage to ensure Fisher consistency (see Section 3).

Definition 3 (Margin operatorρ) Let $\mathcal{G}$ be a class of functions from $\mathcal{X}$ into $\mathbb{R}^C$. Define ρ as an operator on $\mathcal{G}$ such that:

$$\begin{aligned} \rho : \mathcal{G} &\longrightarrow \rho_{\mathcal{G}} \\ g &\mapsto \quad \rho_g \end{aligned}$$

$$\forall\, (x,k) \in \mathcal{Z},\ \ \rho_g\,(x,k) = \frac{1}{2}\left(g_k\,(x) - \max_{l \neq k} g_l\,(x)\right).$$

The function ρ_g is the *margin function* associated with g.

Definition 4 (Parameterized truncated hinge loss $\phi_{2,\gamma}$) For $\gamma \in \mathbb{R}_+^*$, the *parameterized truncated hinge loss* $\phi_{2,\gamma}$ is defined by:

$$\forall t \in \mathbb{R},\ \ \phi_{2,\gamma}\,(t) = \mathbb{1}_{\{t \leqslant 0\}} + \left(1 - \frac{t}{\gamma}\right)\mathbb{1}_{\{t \in (0,\gamma]\}}.$$

With these definitions at hand, the learning problem can be defined as follows.

Definition 5 (Learning problem of the IKM) Let κ be a kernel and $\mathcal{H}$ the function class associated with κ according to Definition 1. For $\mathbf{z}_m = (z_i)_{1 \leqslant i \leqslant m} \in \mathcal{Z}^m$ $\gamma \in \mathbb{R}_+^*$ and $\lambda \in \mathbb{R}_+^*$, the C-category IKM associated with κ, $\mathbf{z}_m$, γ and λ is obtained by solving the following optimization problem:

Problem 1

$$\min_{h \in \mathcal{H}} \left\{ \sum_{i=1}^{m} \phi_{2,\gamma} \circ \rho_h\,(z_i) + \lambda \left\| \mathbf{R}_C \right\|_1 \right\}$$

$$\text{s.t. } \forall k \in [\![1; C]\!],\ O_k \in \operatorname{conv}\left(\{\kappa_{x_i} : y_i = k\}\right),$$

where the function *conv* returns the convex hull of its argument.

Problem 1 is a nonconvex programming problem. To solve it, one can make use of the fact that the parameterized truncated hinge loss is a difference of convex functions (Le Thi and Pham 2005). Its originality rests in the constraints on the centers O_k, inspired by the idea to derive the classifier

from the NCC. They also opportunely provide a representer theorem (for the component functions). Let $\boldsymbol{\theta}_m = (\theta_i)_{1 \leqslant i \leqslant m} \in [0,1]^m$ be the vector of the convex combinations, so that:

$$\forall k \in [\![1; C]\!],\ O_k = \sum_{\{i: y_i = k\}} \theta_i \kappa_{x_i}.$$

Then by application of the reproducing property,

$$\begin{aligned} h_k(x) &= R_k - a_k \langle O_k - \kappa_x, O_k - \kappa_x \rangle_{\mathbf{H}_\kappa}^{\frac{p}{2}} \\ &= R_k - a_k \left(\sum_{\{(i,j): y_i = y_j = k\}} \theta_i \theta_j \kappa(x_i, x_j) - 2 \sum_{\{i: y_i = k\}} \theta_i \kappa(x_i, x) + \kappa(x, x) \right)^{\frac{p}{2}}. \end{aligned} \tag{3}$$

An equivalent but more tractable formulation of Problem 1 is obtained by introduction of slack variables. This formulation is Problem 2.

Problem 2

$$\min_{h \in \mathcal{H}, \boldsymbol{\xi}_m = (\xi_i)_{1 \leqslant i \leqslant m} \in \mathbb{R}_+^m} \{ \|\boldsymbol{\xi}_m\|_1 + \lambda \|\mathbf{R}_C\|_1 \}$$

$$\text{s.t.} \begin{cases} \forall i \in [\![1; m]\!],\ \max\left\{0, \frac{1}{\gamma} \rho_h(z_i)\right\} \geqslant 1 - \xi_i \\ \forall k \in [\![1; C]\!],\ O_k \in \operatorname{conv}(\{\kappa_{x_i} : y_i = k\}) \end{cases}.$$

3 Guarantees on the Generalization Error

In this section, we establish an asymptotic property of the generalization error, the Fisher consistency, and an upper bound on this quantity holding (with high probability) for a finite value of the sample size m, a guaranteed risk. Central in their formulations are the concepts of risk and margin risk, that we define now.

Definition 6 (Risk and margin risk) Let $\mathcal{G}$ be a class of functions from $\mathcal{X}$ into $\mathbb{R}^C$. The *expected risk* of any function $g \in \mathcal{G}$, $L(g)$, is given by:

$$L(g) = \mathbb{E}_{(X,Y) \sim P} \left[\mathbb{1}_{\{\rho_g(X,Y) \leqslant 0\}} \right] = P(\mathrm{dr}_g(X) \neq Y).$$

Its *empirical risk* measured on the m-sample $\mathbf{Z}_m$ is:

$$L_m(g) = \mathbb{E}_{Z' \sim P_m} \left[\mathbb{1}_{\{\rho_g(Z') \leqslant 0\}} \right] = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\{\rho_g(Z_i) \leqslant 0\}}$$

(where P_m is the empirical measure supported on $\mathbf{Z}_m$). Given a class of margin loss functions ϕ_γ parameterized by $\gamma \in (0,1]$, for every (ordered) pair $(g,\gamma) \in \mathcal{G} \times (0,1]$, the *risk with margin* γ of g, $L_\gamma(g)$, is defined as:

$$L_\gamma(g) = \mathbb{E}_{Z\sim P}\left[\phi_\gamma \circ \rho_g(Z)\right].$$

$L_{\gamma,m}(g)$ designates the corresponding empirical risk, measured on the m-sample $\mathbf{Z}_m$:

$$L_{\gamma,m}(g) = \mathbb{E}_{Z'\sim P_m}\left[\phi_\gamma \circ \rho_g\left(Z'\right)\right] = \frac{1}{m}\sum_{i=1}^{m} \phi_\gamma \circ \rho_g(Z_i).$$

The first property we establish is Fisher consistency (Liu 2007).

3.1 Fisher Consistency

Proposition 1 *Let $\mathcal{G}$ be the class of all the functions from $\mathcal{X}$ into $\mathbb{R}^C$. The minimizer g^* of $\mathbb{E}_{(X,Y)}\left[\phi_{2,\gamma} \circ \rho_g(X,Y)\right]$ over $\mathcal{G}$ satisfies the following:*

$$\forall x \in \mathcal{X},\ \exists k(x) \in \underset{1\leqslant k\leqslant C}{\operatorname{argmax}} P(Y=k \mid X=x):\ \rho_{g^*}(x,k(x)) \geqslant \gamma.$$

The learning problem of the IKM (Problem 1) is an implementation of the empirical risk minimization inductive principle, with a restriction corresponding to the constraints on the points O_k. Thus, if the minimizer h^* of the margin risk on $\mathcal{H}$ is such that for every k in $[\![1;C]\!]$, the point O_k^* belongs to the convex hull of the support of the distribution of category k (the constraints do not affect the asymptotic behaviour), then Proposition 1 implies that the estimation error of the machine goes to zero as m goes to infinity. Suppose further that the approximation error is null, which is the case, for instance, if the class densities are given by Eqs. 1 or 2. Then asymptotically, the function selection returns a function h^* whose decision rule is that of the Bayes classifier (its risk is minimal).

We now introduce the guaranteed risk.

3.2 Guaranteed Risk In order to derive the upper bound on the probability of error of the classifier, the following assumptions are made.

Hypothesis 1 *The kernel κ is supposed to be such that $\sup_{x\in\mathcal{X}} \|\kappa_x\|_{\mathbf{H}_\kappa} \leqslant \frac{1}{2}$. The norms of the centers O are bounded from above by the same value. At last, the parameters R are bounded from above by 1.*

The assumption on the norms of the functions κ_x is not restrictive since it is always possible to standardize the kernel. Once this first assumption made, then the second one is a consequence of the constraints of Problem 1.

Combining them by means of Minkowski inequality implies that the term $\|O - \kappa_x\|_{\mathbf{H}_\kappa}^p$ is always (for every value of p) upper bounded by 1.

The capacity measure involved in our guaranteed risk is a Rademacher complexity (Bartlett and Mendelson 2002; Koltchinskii and Panchenko 2002).

Definition 7 (Rademacher complexity) Let $(\mathcal{T}, \mathcal{A}_\mathcal{T}, P_\mathcal{T})$ be a probability space and let T be a random variable distributed according to $P_\mathcal{T}$. For $n \in \mathbb{N}^*$, let $\mathbf{T}_n = (T_i)_{1 \leqslant i \leqslant n}$ be an n-sample made up of independent copies of T and let $\boldsymbol{\sigma}_n = (\sigma_i)_{1 \leqslant i \leqslant n}$ be a Rademacher sequence. Let $\mathcal{F}$ be a class of real-valued functions with domain $\mathcal{T}$. The *empirical Rademacher complexity* of $\mathcal{F}$ given $\mathbf{T}_n$ is

$$\hat{R}_n(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(T_i) \,\middle|\, \mathbf{T}_n\right].$$

The *Rademacher complexity* of $\mathcal{F}$ is

$$R_n(\mathcal{F}) = \mathbb{E}_{\mathbf{T}_n \sim P_\mathcal{T}^n}\left[\hat{R}_n(\mathcal{F})\right].$$

The basic supremum inequality involving a Rademacher complexity can be seen as an application of Theorem 9.2 in Mohri et al. (2018).

Theorem 1 *Let $\mathcal{H}$ be the class of functions given by Definition 1. For a fixed $\gamma \in (0, 1]$ and a fixed $\delta \in (0, 1)$, with P^m-probability at least $1 - \delta$,*

$$\sup_{h \in \mathcal{H}} \left(L(h) - L_{\gamma,m}(h)\right) \leqslant \frac{2C}{\gamma} R_m\left(\bigcup_{p \in [1,2]} \mathcal{H}_p\right) + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2m}}, \tag{4}$$

where the margin loss function defining the empirical margin risk is the parameterized truncated hinge loss (Definition 4).

With Theorem 1 at hand, deriving a guaranted risk for the IKM boils down to bounding from above $R_n\left(\bigcup_{p \in [1,2]} \mathcal{H}_p\right)$. Handling the union over all the possible values for p raises a difficulty, all the more since among these values, only one can be treated with results from the literature: the value 2. This problem is taken care of by two new lemmas that should be of interest in their own right: a structural result on covering numbers, Lemma 4, and an extension of Talagrand's contraction lemma (see for instance Lemma 5.7 in Mohri et al., 2018), Lemma 7. This yields to the following bound.

Lemma 1 *Let the function classes $\mathcal{H}_p$ be those of Definition 1, under Hypothesis 1. Suppose that there exists $\Lambda > 0$ such that $\sup_{h \in \mathcal{H}_p} a \leqslant \Lambda$. Then for $n \geqslant 2$,*

$$R_n\left(\bigcup_{p\in[1,2]} \mathcal{H}_p\right) \leqslant \frac{1}{2\sqrt{n}} + 2\Lambda\left(\frac{\log_2(n)}{n}\right)^{\frac{1}{4}}\left[1 + \frac{K}{\sqrt{\log_2(n)}}\sqrt{\ln\left(10\left(\frac{n}{\log_2(n)}\right)^{\frac{1}{4}}\right)}\right], \tag{5}$$

where $K = 138240$.

The combination of Theorem 1 and Lemma 1 produces a guaranteed risk whose convergence rate is a $O\left(\left(\frac{\ln(m)}{m}\right)^{\frac{1}{4}}\right)$. The value of the exponent is a function of the value of the lower endpoint of the interval in which p takes its value. Precisely, should this endpoint be set to $p_0 \in (1, 2]$, then the exponent would (improve to) become $\frac{p_0}{4}$. This behaviour can be exploited to perform capacity control, for instance through the choice of the initial feasible solution for Problem 1. Indeed, it suggests to use the (kernelized) NCC with a specific value of p, the largest possible one: 2. Then, training should be monitored so as to avoid considering lower values unless the prediction accuracy requires it, in application of some kind of structural risk minimization.

4 Comparative Study

In this study, the IKM is compared with four popular classifiers whose decision boundaries in the feature space are well characterized: the M-SVM of Weston and Watkins (1998), hereafter referred to as the WW-M-SVM, the kernel linear discriminant analysis (KFDA), the QDA and a multilayer perceptron (MLP). The WW-M-SVM is used with a Gaussian kernel. Our implementation of this machine can be found at the following address: https://members.loria.fr/YGuermeur, with our implementation of the MLP. The KFDA is the one of the package in R (Yang et al. 2004) and the QDA is the one of scikit-learn.

As for IKM, the algorithm used to solve Problem 2 rests on a heuristic. For values of p and $\mathbf{a}_C$ taken on a grid it alternately optimizes vector $\mathbf{R}_C$ while keeping vector $\mathbf{O}_C$, i.e., vector $\boldsymbol{\theta}_m$, fixed, and vice versa. The optimization of $\mathbf{R}_C$ relies on a linear problem and to optimize vector $\boldsymbol{\theta}_m$, the principle of the sequential minimal optimization (Platt 1999) is applied. An index k of category and two distinct indices i and j in $[\![1; m]\!]$ satisfying

$y_i = y_j = k$ are chosen randomly. Then, the values of the components θ_i and θ_j are optimized while the other ones remain unchanged.

A first group of experiments assesses the behaviour of the IKM when the kernel is the Euclidean dot product (linear kernel). This kernel is then replaced with a Gaussian kernel.

4.1 IKM with a Linear Kernel This study involves two artificial problems (for which the Bayes classifier is available). In both cases, all the prior probabilities of the categories are equal (i.e., equal to $1/C$).

The first problem is a two-dimensional binary classification problem: the *bullseye*, or more precisely the half bullseye (so that the task could be handled efficiently by all five classifiers considered). The class conditional densities are provided by the following formula:

$$\forall k \in \{1,2\}\,,\ \forall x \in \mathbb{R}_+ \times \mathbb{R},\quad f_k(x) = \int_{-\pi/2}^{\pi/2} \int_0^{+\infty} \frac{\exp\left(-\frac{(x_1 - r\cos\theta)^2}{2\sigma_k^2} - \frac{(x_2 - r\sin\theta)^2}{2\sigma_k^2}\right)}{2\pi^2\sigma_k^2} \frac{r}{\lambda_k} \exp\left(-\frac{r^2}{2\lambda_k}\right) \mathrm{d}r\mathrm{d}\theta.$$

We chose $\lambda_1 = 1$, $\lambda_2 = 3$ and $\sigma_1 = \sigma_2 = 0.2$. The decision boundary produced by the IKM, a section of a branch of an hyperbola-like curve since the value obtained for the exponent p is 1, is depicted in Fig. 1.

The second problem involves categories whose distributions are given by Eq. 1 (generalized Laplace distributions - gLaplace). Two configurations are considered. The first one corresponds to $C = 2$, $d = 2$, $\mu_1^T = (0,0)$, $\mu_2^T = (5,5)$, and $\mathbf{b}_2^T = (1,10)$. The second one is given by $C = 3$, $d = 2$, $\mu_1^T = (0,0)$, $\mu_2^T = (5,5)$, $\mu_3^T = (5,-5)$, and $\mathbf{b}_3^T = (1,10,3)$.

The recognition rates are reported in Table 1.

The performances of the IKM and the WW-M-SVM are similar, and globally superior to those of the three other classifiers. According to the two-sample proportion test, their superiority over the MLP on the third data set is statistically significant with confidence exceeding 0.95. This is noteworthy since the generalized Laplace distributions are isotropic.

4.2 IKM with a Gaussian Kernel The IKM with a Gaussian kernel is assessed on nine data sets from the literature. Among those data sets, six are from the UCI repository website [1]: the "SPECT Heart Data Set" (SPECT), the "Glass Identification Data Set" (glass), the "Car Evaluation Data Set" (car), the "Arcene Data Set", the "MicroMass Data Set" and the "LSVT Voice Rehabilitation Data Set" (LSVT). The "Hipparcos-1 dataset"

[1] https://archive.ics.uci.edu/ml/index.php

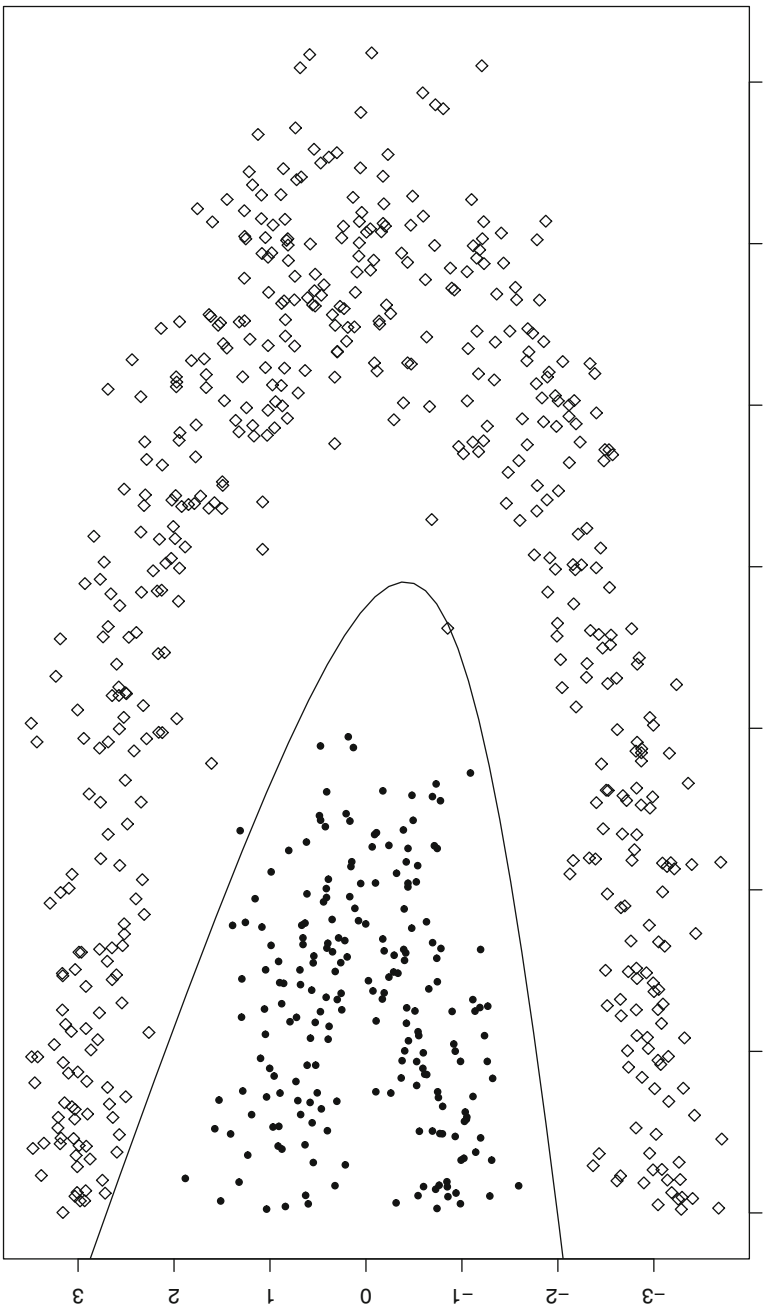

Figure 1: Decision boundary of the IKM with a linear kernel trained to solve the half bullseye problem

Table 1: Evaluation of the IKM with a linear kernel

data set	C	d	nb. of training/test data	IKM	WW-M-SVM	KFDA	QDA	MLP
bullseye	2	2	700/700	99.7	100.0	99.9	97.3	99.9
gLaplace	2	2	2000/2000	96.6	96.5	94.0	96.1	96.5
gLaplace	3	2	2400/2400	82.9	83.1	81.5	82.7	80.8

Table 2: Evaluation of the IKM with a Gaussian kernel

data set	C	d	IKM	WW-M-SVM	KFDA	QDA	MLP
SPECT	2	22	84.68	84.32	82.01	78.60	84.27
glass	6	9	68.72	68.40	62.77	26.23	55.14
hipparcos	14	22	82.20	82.32	78.83	70.96	80.55
car	4	6	98.47	98.84	83.63	3.75	91.49
Arcene	2	10000	77.00	74.00	56.00	51.00	63.00
MicroMass	20	1300	80.92	65.67	10.50	4.00	57.09
clean	2	168	97.26	96.84	56.52	61.99	74.16
USPS	10	256	90.80	89.10	87.00	29.1	89.70
LSVT	2	309	85.76	82.56	-	67.35	66.67

(hipparcos) is borrowed from the hipparcos-1 catalogue (ESA 1997) while the "Clean data set" (clean) was introduced in Vanschoren et al. (2014). The USPS data set is a subset of size 1000 of the set introduced by Schölkopf et al. (1997).

The recognition rates obtained by means of a five-fold cross-validation are reported in Table 2.

The IKM yields similar results to those of the WW-M-SVM. Its performance is all the better as the dimension of the data is larger. It is especially by far the best classifier for the two data sets living in the highest dimensional spaces: Arcene and MicroMass. This is all the more remarkable as little effort has been spent so far on the optimization process.

5 Conclusions and Ongoing Research

A new kernel machine for pattern classification has been introduced: the isotropic kernel machine. It aims at exploiting isotropy through a nonparametric approach derived from the nearest centroid classifier. Its decision boundaries are nonlinear in the reproducing kernel Hilbert space of the kernel. We give a bound on its learning risk using new tools that should prove interesting in their own right. The initial experimental results are promising.

Our ongoing research follows three directions. The first one is the derivation of sharper multi-class risk bounds. The second one is the design of automatic methods for the choice of the hyperparameters: the kernel κ, the margin parameter γ and the regularization coefficient λ. At last, the third one is the optimization of the training algorithm.

Funding Information None to be declared.

Declarations

Conflicts of Interest None to be declared.

References

Anthony, M. and Bartlett, P. (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge.

Bartlett, P. and Mendelson, S. (2002). Rademacher and Gaussian complexities: risk bounds and structural results. *J. Mach. Learn. Res.* **3**, 463–482.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, Boston.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* **20**, 273–297.

Dogan, U., Glasmachers T. and Igel C. (2016). A unified view on multi-class support vector classification. *J. Mach. Learn. Res.* **17**, 1–32.

Dudley, R. (1967). The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *J. Funct. Anal.* **1**, 290–330.

Dudley, R. (1984). A course on empirical processes. In *Ecole d'Eté de Probabilités de Saint-Flour XII - 1982*, (P. Hennequin, ed.). Lecture Notes in Mathematics, Springer-Verlag, vol. 1097, pp. 1–142.

ESA (1997). ESA SP-1200: The HIPPARCOS and TYCHO catalogues. Astrometric and photometric star catalogues derived from the ESA HIPPARCOS space astrometry mission ESA, Noordwijk.

Genton, M. (2001). Classes of kernels for machine learning: a statistics perspective. *J. Mach. Learn. Res.* **2**, 299–312.

Ghorbani, B., Mei, S., Misiakiewicz T. and Montanari A. (2020). When do neural networks outperform kernel methods? In *NeurIPS* 34.

Guermeur, Y. (2012). A generic model of multi-class support vector machine. *Int. J. Intell. Inf. Database Syst.* **6**, 555–577.

Guermeur, Y. (2017). L_p-norm Sauer-Shelah lemma for margin multi-category classifiers. *J. Comput. Syst. Sci.* **89**, 450–473.

Hastie, T., Tibshirani R. and Friedman J. (2008). *The Elements of Statistical Learning*. Springer

Kearns, M. and Schapire, R. (1994). Efficient distribution-free learning of probabilistic concepts. *J. Comput. Syst. Sci.* **48**, 464–497.

Kearns, M., Schapire R. and Sellie L. (1994). Toward efficient agnostic learning. *Mach. Learn.* **17**, 115–141.

Kolmogorov, A. and Tihomirov, V. (1961). ϵ-entropy and ϵ-capacity of sets in functional spaces. *Amer. Math. Soc. Transl. Ser. 2* **17**, 277–364.

Koltchinskii, V. and Panchenko, D. (2002). Empirical margin distributions and bounding the generalization error of combined classifiers. *Ann. Stat.* **30**, 1–50.

Le Thi, H. and Pham, D. (2005). The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Ann. Oper. Res.* **133**, 23–46.

Liu, Y. (2007). Fisher consistency of multicategory support vector machines. In *Eleventh International Conference on Artificial Intelligence and Statistics*, pp. 289–296.

Mendelson, S. (2002). Rademacher averages and phase transitions in Glivenko-Cantelli classes. *IEEE Trans. Inf. Theory* **48**, 251–263.

Mendelson, S. (2003). A few notes on statistical learning theory. In *Advanced Lectures on Machine Learning*, (S. Mendelson and A. Smola, eds.). Springer-Verlag, Berlin, Heidelberg, New York, chap. 1, pp. 1–40.

Mendelson, S. and Vershynin, R. (2003). Entropy and the combinatorial dimension. *Invent. Math.* **152**, 37–55.

Mohri, M., Rostamizadeh A. and Talwalkar, A. (2018). *Foundations of Machine Learning*, 2nd edn. The MIT Press, Cambridge, MA.

Platt, J. (1999). Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods, Support Vector Learning*, (B. Schölkopf, C. Burges and A. Smola eds.). The MIT Press, Cambridge, MA, chap. 12, pp. 185–208.

Schölkopf, B. and Smola, A. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, Cambridge, MA.

Schölkopf, B., Sung, K.K., Burges, C., Girosi, F., Niyogi, P., Poggio T. and Vapnik V. (1997). Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Trans. Signal Process.* **45**, 2758–2765.

Vanschoren, J., van Rijn, J., Bischl B. and Torgo L. (2014). OpenML: networked science in machine learning. *ACM SIGKDD Explor. Newslett.* **15**, 49–60.

Wang, J., Plataniotis, K., Lu J. and Venetsanopoulos A. (2008). Kernel quadratic discriminant analysis for small sample size problem. *Pattern. Recog.* **41**, 1528–1538.

Weston, J. and Watkins, C. (1998). Multi-class support vector machines. Tech. Rep. CSD-TR-98-04, Royal Holloway, University of London, Department of Computer Science.

Williamson, R., Smola A. and Schölkopf B. (2001). Generalization performance of regularization networks and support vector machines *via* entropy numbers of compact operators. *IEEE Trans. Inf. Theory.* **47**, 2516–2532.

Yang, J., Jin, Z., Yang, J., Zhang D. and Frangi A. (2004). Essence of kernel fisher discriminant: KPCA plus LDA. *Pattern Recog.* **37**, 2097–2100.

Appendix A: Proof of the Fisher Consistency

Proof By disintegration (see Lemma 1.2.1 in Dudley, 1984), there exists a map $x \mapsto P(\cdot \mid x)$ from $\mathcal{X}$ into the set of all probability measures on $\mathcal{Y}$ such that P is the joint distribution of $(P(\cdot \mid x))_{x \in \mathcal{X}}$ and of the marginal

distribution $P_{\mathcal{X}}$ of P on $\mathcal{X}$. Consequently,

$$\begin{aligned}\mathbb{E}_{(X,Y)}\left[\phi_{2,\gamma}\circ\rho_g\left(X,Y\right)\right] &= \int_{\mathcal{X}\times\mathcal{Y}}\phi_{2,\gamma}\circ\rho_g\left(x,y\right)dP(x,y)\\ &= \int_{\mathcal{X}}\left\{\sum_{k=1}^{C}\phi_{2,\gamma}\circ\rho_g\left(x,k\right)P\left(Y=k\mid X=x\right)\right\}dP_{\mathcal{X}}(x)\\ &= \mathbb{E}_X\left[\sum_{k=1}^{C}\phi_{2,\gamma}\circ\rho_g\left(X,k\right)P\left(Y=k\mid X\right)\right],\end{aligned}$$

from which it springs that

$$\forall x\in\mathcal{X},\quad g^*\in\underset{g\in\mathcal{G}}{\operatorname{argmin}}\sum_{k=1}^{C}\phi_{2,\gamma}\circ\rho_g\left(x,k\right)P\left(Y=k\mid X=x\right).$$

Given $x\in\mathcal{X}$ and $g\in\mathcal{G}$, by definition of ρ_g, there is at most one value of k in $[\![1;C]\!]$ such that $\rho_g\left(x,k\right)>0$. Suppose that there is none. Then according to Definition 4,

$$\begin{aligned}\sum_{k=1}^{C}\phi_{2,\gamma}\circ\rho_g\left(x,k\right)P\left(Y=k\mid X=x\right) &= \sum_{k=1}^{C}P\left(Y=k\mid X=x\right)\\ &= 1.\end{aligned}\tag{6}$$

Suppose on the contrary that there exists $k^*\in[\![1;C]\!]$ such that $\rho_g\left(x,k^*\right)>0$. Then

$$\sum_{k=1}^{C}\phi_{2,\gamma}\circ\rho_g\left(x,k\right)P\left(Y=k\mid X=x\right) = 1+\left(\phi_{2,\gamma}\circ\rho_g\left(x,k^*\right)-1\right)P\left(Y=k^*\mid X=x\right)\tag{7}$$

$$< 1.\tag{8}$$

By definition of $\mathcal{G}$ and g^*, it springs from Eqs. 6 and 8 that g^* satisfies:

$$\forall x\in\mathcal{X},\exists!\,k\left(x\right)\in[\![1;C]\!]:\ \rho_{g^*}\left(x,k\left(x\right)\right)>0.$$

Furthermore, due to Eq. 7,

$$\forall x\in\mathcal{X},\ \begin{cases}k\left(x\right)\in\operatorname{argmax}_{1\leqslant k\leqslant C}P\left(Y=k\mid X=x\right)\\ \rho_{g^*}\left(x,k\left(x\right)\right)\geqslant\gamma\end{cases},$$

so that

$$\sum_{k=1}^{C} \phi_{2,\gamma} \circ \rho_{g^*}(x,k)\, P(Y=k \mid X=x) = 1 - \max_{1 \leqslant k \leqslant C} P(Y=k \mid X=x).$$

Appendix B: Technical Lemmas

We have seen in Section 3.2 that in order to prove the guaranteed risk, it is enough to prove Lemma 1. This proof makes use of technical lemmas with are gathered in this appendix. They involve concepts which are defined first.

The concept of covering number (ϵ-entropy), as well as the underlying concepts of ϵ-cover and ϵ-net, can be traced back to Kolmogorov and Tihomirov (1961).

Definition 8 (ϵ-cover, ϵ-net, covering numbers, and ϵ-entropy) Let $(\mathcal{E}, \rho)$ be a pseudo-metric space, $\mathcal{E}' \subset \mathcal{E}$ and $\epsilon \in \mathbb{R}_+^*$. An *$\epsilon$-cover* of $\mathcal{E}'$ is a coverage of $\mathcal{E}'$ with open balls of radius ϵ the centers of which belong to $\mathcal{E}$. These centers form an *ϵ-net* of $\mathcal{E}'$. An *internal/proper ϵ-net* of $\mathcal{E}'$ is an *ϵ-net* of $\mathcal{E}'$ included in $\mathcal{E}'$. If $\mathcal{E}'$ has an ϵ-net of finite cardinality, then its *covering number* $\mathcal{N}(\epsilon, \mathcal{E}', \rho)$ is the smallest cardinality of its ϵ-nets:

$$\mathcal{N}(\epsilon, \mathcal{E}', \rho) = \min\left\{ \left|\mathcal{E}''\right| : \ \left(\mathcal{E}'' \subset \mathcal{E}\right) \wedge \left(\forall e \in \mathcal{E}', \ \rho\left(e, \mathcal{E}''\right) < \epsilon\right) \right\}.$$

If there is no such finite net, then the covering number is defined to be infinite. The corresponding binary logarithm, $\log_2(\mathcal{N}(\epsilon, \mathcal{E}', \rho))$, is called the *minimal ϵ-entropy* of $\mathcal{E}'$, or simply the *ϵ-entropy* of $\mathcal{E}'$. $\mathcal{N}^{\text{int}}(\epsilon, \mathcal{E}', \rho)$ will designate a covering number of $\mathcal{E}'$ obtained by considering internal ϵ-nets only. In the finite case, we have thus:

$$\mathcal{N}^{\text{int}}(\epsilon, \mathcal{E}', \rho) = \min\left\{ \left|\mathcal{E}''\right| : \ \left(\mathcal{E}'' \subset \mathcal{E}'\right) \wedge \left(\forall e \in \mathcal{E}', \ \rho\left(e, \mathcal{E}''\right) < \epsilon\right) \right\}.$$

Definition 9 (Pseudo-distance $d_{p,\mathbf{t}_n}$) Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{T}$. For $n \in \mathbb{N}^*$, let $\mathbf{t}_n = (t_i)_{1 \leqslant i \leqslant n} \in \mathcal{T}^n$. Then,

$$\forall p \in [1, +\infty), \forall (f, f') \in \mathcal{F}^2, \ \ d_{p,\mathbf{t}_n}(f, f') = \left\|f - f'\right\|_{L_p(\mu_{\mathbf{t}_n})} = \left(\frac{1}{n} \sum_{i=1}^{n} \left|f(t_i) - f'(t_i)\right|^p \right)^{\frac{1}{p}}$$

and

$$\forall (f, f') \in \mathcal{F}^2, \ \ d_{\infty,\mathbf{t}_n}(f, f') = \left\|f - f'\right\|_{L_\infty(\mu_{\mathbf{t}_n})} = \max_{1 \leqslant i \leqslant n} \left|f(t_i) - f'(t_i)\right|,$$

where $\mu_{\mathbf{t}_n}$ denotes the uniform (counting) probability measure on $\{t_i : 1 \leqslant i \leqslant n\}$.

Definition 10 (Uniform covering numbers, Williamson et al., 2001) Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{T}$ and $\bar{\mathcal{F}} \subset \mathcal{F}$. For $p \in [1, +\infty]$, $\epsilon \in \mathbb{R}_+^*$, and $n \in \mathbb{N}^*$, the *uniform covering numbers* $\mathcal{N}_p\left(\epsilon, \bar{\mathcal{F}}, n\right)$ are defined as follows:

$$\mathcal{N}_p\left(\epsilon, \bar{\mathcal{F}}, n\right) = \sup_{\mathbf{t}_n \in \mathcal{T}^n} \mathcal{N}\left(\epsilon, \bar{\mathcal{F}}, d_{p,\mathbf{t}_n}\right).$$

We define accordingly the *uniform covering numbers* $\mathcal{N}_p^{\text{int}}\left(\epsilon, \bar{\mathcal{F}}, n\right)$ as:

$$\mathcal{N}_p^{\text{int}}\left(\epsilon, \bar{\mathcal{F}}, n\right) = \sup_{\mathbf{t}_n \in \mathcal{T}^n} \mathcal{N}^{\text{int}}\left(\epsilon, \bar{\mathcal{F}}, d_{p,\mathbf{t}_n}\right).$$

Definition 11 (Fat-shattering dimension, Kearns and Schapire, 1994) Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{T}$. For $\gamma \in \mathbb{R}_+^*$, a subset $s_{\mathcal{T}^n} = \{t_i : 1 \leqslant i \leqslant n\}$ of $\mathcal{T}$ is said to be *γ-shattered* by $\mathcal{F}$ if there is a vector $\mathbf{b}_n = (b_i)_{1 \leqslant i \leqslant n} \in \mathbb{R}^n$ such that, for every vector $\mathbf{s}_n = (s_i)_{1 \leqslant i \leqslant n} \in \{-1, 1\}^n$, there is a function $f_{\mathbf{s}_n} \in \mathcal{F}$ satisfying

$$\forall i \in [\![1, n]\!], \quad s_i\left(f_{\mathbf{s}_n}\left(t_i\right) - b_i\right) \geqslant \gamma.$$

The vector $\mathbf{b}_n$ is called a *witness* to the γ-shattering. The *fat-shattering dimension at scale* γ of the class $\mathcal{F}$, $\gamma\text{-dim}\left(\mathcal{F}\right)$, is the maximal cardinality of a subset of $\mathcal{T}$ γ-shattered by $\mathcal{F}$, if such maximum exists. Otherwise, $\mathcal{F}$ is said to have infinite fat-shattering dimension at scale γ.

Definition 12 For $p \in [1, 2]$, the function class $\mathcal{F}_p$ is the set $\left\{f_{O,p} : x \mapsto \|O - \kappa_x\|_{\mathbf{H}_\kappa}^p\right\}$ satisfying Hypothesis 1.

Lemma 2 *Let $\mathcal{F}$ be the class of constant functions on $\mathcal{T}$ whose values range from 0 to $M_{\mathcal{F}}$. Then*

$$\forall n \in \mathbb{N}^*, \; R_n\left(\mathcal{F}\right) \leqslant \frac{M_{\mathcal{F}}}{2\sqrt{n}}.$$

Proof For every $\mathbf{t}_n = (t_i)_{1 \leqslant i \leqslant n} \in \mathcal{T}^n$,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f\left(t_i\right)\right] &= \frac{1}{n} \mathbb{E}_{\boldsymbol{\sigma}_n}\left[M_{\mathcal{F}} \mathbb{1}_{\{\sum_{i=1}^n \sigma_i > 0\}} \sum_{i=1}^n \sigma_i\right] \\
&= \frac{M_{\mathcal{F}}}{2n} \mathbb{E}_{\boldsymbol{\sigma}_n}\left[\left|\sum_{i=1}^n \sigma_i\right|\right].
\end{aligned}$$

Jensen's inequality gives

$$\mathbb{E}_{\boldsymbol{\sigma}_n}\left[\left|\sum_{i=1}^{n}\sigma_i\right|\right] \leqslant \sqrt{n},$$

thus concluding the proof.

Lemma 3 *Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{T}$ including the null function and $\lambda \in \mathbb{R}_+^*$. Let $\mathcal{F}_\lambda = \{\alpha f : \alpha \in (0,\lambda], f \in \mathcal{F}\}$. Then,*

$$\forall n \in \mathbb{N}^*,\ R_n(\mathcal{F}_\lambda) \leqslant 2\lambda R_n(\mathcal{F}).$$

Proof For every $\mathbf{t}_n = (t_i)_{1\leqslant i\leqslant n} \in \mathcal{T}^n$,

$$\mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{f\in\mathcal{F},\alpha\in(0,\lambda]}\sum_{i=1}^{n}\sigma_i\alpha f(t_i)\right] \leqslant \lambda\mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i f(t_i)\right|\right].$$

Let $(\cdot)_+$ and $(\cdot)_-$ denote respectively the positive and negative part functions. Due to the subadditivity of the supremum,

$$\mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i f(t_i)\right|\right] \leqslant \mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{n}\sigma_i f(t_i)\right)_+\right] + \mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{n}\sigma_i f(t_i)\right)_-\right].$$

By symmetry, $-\boldsymbol{\sigma}_n$ has the same distribution as $\boldsymbol{\sigma}_n$ and $(-\cdot)_- = (\cdot)_+$. Thus, the two expectations of the right-hand side are equal, with the consequence that

$$\mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i f(t_i)\right|\right] \leqslant 2\mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{n}\sigma_i f(t_i)\right)_+\right].$$

To conclude the proof, it suffices to notice that since the class $\mathcal{F}$ includes the null function, $\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{n}\sigma_i f(t_i)\right)_+ = \sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i f(t_i)$ almost surely.

Lemma 4 *Let the function classes $\mathcal{F}_p$ be those of Definition 12. For $\epsilon \in (0,1]$, let ϕ_ϵ be the function mapping $k \in [\![1; \lceil \frac{1}{\epsilon} \rceil]\!]$ to $1 + \frac{2k-1}{2}\lceil \frac{1}{\epsilon} \rceil^{-1}$. Then for $\epsilon \in (0,1]$, $n \in \mathbb{N}^*$ and $\mathbf{x}_n = (x_i)_{1\leqslant i\leqslant n} \in \mathcal{X}^n$,*

$$\mathcal{N}^{int}\left(\epsilon, \bigcup_{p\in[1,2]}\mathcal{F}_p, d_{2,\mathbf{x}_n}\right) \leqslant \sum_{k=1}^{\lceil \frac{1}{\epsilon} \rceil}\mathcal{N}^{int}\left(\frac{\epsilon}{2}, \mathcal{F}_{\phi_\epsilon(k)}, d_{2,\mathbf{x}_n}\right). \tag{9}$$

Proof Let us consider any value $p_0 \in [1,2]$. Let $k_0 \in \operatorname{argmin}_{1 \leqslant k \leqslant \lceil \frac{1}{\epsilon} \rceil} |p_0 - \phi_\epsilon(k)|$ and $\delta = |p_0 - \phi_\epsilon(k_0)|$. Note that by construction, $\delta \leqslant \frac{\epsilon}{2}$. Let $\mathcal{N}_{\phi_\epsilon(k_0)}$ be a proper $\frac{\epsilon}{2}$-net of $\mathcal{F}_{\phi_\epsilon(k_0)}$ (with respect to the pseudo-metric $d_{2,\mathbf{x}_n}$). By definition, for every $O \in \mathbf{H}_\kappa$, there exists $\bar{O} \in \mathbf{H}_\kappa$ such that the function $f_{\bar{O},\phi_\epsilon(k_0)}$ belongs to $\mathcal{N}_{\phi_\epsilon(k_0)}$ and

$$d_{2,\mathbf{x}_n}\left(f_{O,\phi_\epsilon(k_0)}, f_{\bar{O},\phi_\epsilon(k_0)}\right) < \frac{\epsilon}{2}.$$

Then

$$\begin{aligned} d_{2,\mathbf{x}_n}\left(f_{O,p_0}, f_{\bar{O},\phi_\epsilon(k_0)}\right) &\leqslant d_{2,\mathbf{x}_n}\left(f_{O,p_0}, f_{O,\phi_\epsilon(k_0)}\right) + d_{2,\mathbf{x}_n}\left(f_{O,\phi_\epsilon(k_0)}, f_{\bar{O},\phi_\epsilon(k_0)}\right) \\ &< d_{\infty,\mathbf{x}_n}\left(f_{O,p_0}, f_{O,\phi_\epsilon(k_0)}\right) + \frac{\epsilon}{2}. \end{aligned} \tag{10}$$

Now, since Hypothesis 1 implies that $\max_{1 \leqslant i \leqslant n} \|O - \kappa_{x_i}\|_{\mathbf{H}_\kappa} \leqslant 1$,

$$\begin{aligned} d_{\infty,\mathbf{x}_n}\left(f_{O,p_0}, f_{O,\phi_\epsilon(k_0)}\right) &= \max_{1 \leqslant i \leqslant n} \left| \|O - \kappa_{x_i}\|_{\mathbf{H}_\kappa}^{p_0} - \|O - \kappa_{x_i}\|_{\mathbf{H}_\kappa}^{\phi_\epsilon(k_0)} \right| \\ &\leqslant \max_{t \in [0,1]} \left| t^{p_0} - t^{\phi_\epsilon(k_0)} \right|. \end{aligned}$$

Let F be the function mapping $t \in [0,1]$ to $\left|t^{p_0} - t^{\phi_\epsilon(k_0)}\right|$. If $\delta > 0$ (F is not the null function), then its maximum is reached at $t_* = \left(\frac{\min\{p_0, \phi_\epsilon(k_0)\}}{\max\{p_0, \phi_\epsilon(k_0)\}}\right)^{\frac{1}{\delta}}$. Thus,

$$\begin{aligned} d_{\infty,\mathbf{x}_n}\left(f_{O,p_0}, f_{O,\phi_\epsilon(k_0)}\right) &\leqslant F(t_*) \\ &= t_*^{p_0} \frac{\delta}{\phi_\epsilon(k_0)} = t_*^{\phi_\epsilon(k_0)} \frac{\delta}{p_0} \\ &< \delta \\ &\leqslant \frac{\epsilon}{2}. \end{aligned} \tag{11}$$

By substitution of Eqs. 11 into 10, $d_{2,\mathbf{x}_n}\left(f_{O,p_0}, f_{\bar{O},\phi_\epsilon(k_0)}\right) < \epsilon$, which implies that $\mathcal{N}_{\phi_\epsilon(k_0)}$ is also an ϵ-net of $\mathcal{F}_{p_0}$. Taking the union over the domain of p_0, i.e., $[1,2]$, establishes that $\bigcup_{k \in [\![1; \lceil \frac{1}{\epsilon} \rceil]\!]} \mathcal{N}_{\phi_\epsilon(k)}$ is a proper ϵ-net of $\bigcup_{p \in [1,2]} \mathcal{F}_p$, which concludes the proof.

The following lemma is implicit in the discussion following Definition 4.2 in Mendelson (2002).

Lemma 5 *Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ into $[-M_{\mathcal{F}}, M_{\mathcal{F}}]$ with $M_{\mathcal{F}} \in \mathbb{R}_+^*$. For every $\epsilon \in (0, M_{\mathcal{F}}]$, if $\sup_{\mathbf{t}_n \in \mathcal{T}^n} \hat{R}_n(\mathcal{F}) < \epsilon$ for some $n \in \mathbb{N}^*$, then ϵ-dim$(\mathcal{F}) < n$.*

Lemma 6 *For every $\beta \in \left[\frac{1}{2}, 1\right]$ and every triplet $(u, v, w) \in (\mathbb{R}_+)^3$ such that $u \leqslant v$ and $w > 0$,*

$$v^\beta - u^\beta \leqslant w + \beta w^{1-\frac{1}{\beta}} (v - u).$$

Proof The result is obvious for $\beta = 1$. So, we prove it for $\beta \in \left[\frac{1}{2}, 1\right)$ only. The value of w minimizing the right-hand side is $[(1-\beta)(v-u)]^\beta$. Thus, proving the lemma boils down to establishing that the function f mapping u to $\left(\frac{1}{1-\beta}\right)^{1-\beta} (v-u)^\beta - v^\beta + u^\beta$ is nonnegative on $[0, v]$. Taking the derivative, it is easy to establish that f is increasing on $\left[0, \frac{1-\beta}{2-\beta} v\right]$ and decreasing on $\left[\frac{1-\beta}{2-\beta} v, v\right]$. Thus, $\min_{u \in [0,v]} f(u) = \min\{f(0), f(v)\} = f(v) = 0$.

Lemma 7 *Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{T}$. If $\Phi : \mathbb{R} \longrightarrow \mathbb{R}$ is such that there exist $L_\Phi \in \mathbb{R}_+^*$ and $c \in \mathbb{R}$ satisfying:*

$$\forall (u, v) \in \mathbb{R}^2, \ \ |\Phi(u) - \Phi(v)| \leqslant L_\Phi |u - v| + c,$$

then

$$\forall n \in \mathbb{N}^*, \ R_n(\Phi \circ \mathcal{F}) \leqslant L_\Phi R_n(\mathcal{F}) + \frac{c}{2}.$$

Proof The proof is basically the one of Talagrand's contraction lemma. For every $\mathbf{t}_n = (t_i)_{1 \leqslant i \leqslant n} \in \mathcal{T}^n$,

$$\begin{aligned} \hat{R}_n(\Phi \circ \mathcal{F}) = & \ \mathbb{E}_{\boldsymbol{\sigma}_n} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \Phi \circ f(t_i) \right] \\ = & \ \frac{1}{n} \mathbb{E}_{\boldsymbol{\sigma}_{n-1}} \left[\mathbb{E}_{\sigma_n} \left[\sup_{f \in \mathcal{F}} \{u_{n-1}(f) + \sigma_n \Phi \circ f(t_n)\} \right] \right] \end{aligned}$$

where $u_{n-1}(f) = \sum_{i=1}^{n-1} \sigma_i \Phi \circ f(t_i)$. By definition of the supremum, for any $\epsilon > 0$, there exists $(f_+, f_-) \in \mathcal{F}^2$ such that

$$\begin{cases} u_{n-1}(f_+) + \Phi \circ f_+(t_n) \geqslant (1-\epsilon) \left[\sup_{f \in \mathcal{F}} \{u_{n-1}(f) + \Phi \circ f(t_n)\}\right] \\ u_{n-1}(f_-) - \Phi \circ f_-(t_n) \geqslant (1-\epsilon) \left[\sup_{f \in \mathcal{F}} \{u_{n-1}(f) - \Phi \circ f(t_n)\}\right] \end{cases}.$$

Thus, for any $\epsilon > 0$, by definition of $\mathbb{E}_{\sigma_n}$,

$$
\begin{aligned}
&(1-\epsilon)\,\mathbb{E}_{\sigma_n}\left[\sup_{f\in\mathcal{F}}\left\{u_{n-1}(f)+\sigma_n\Phi\circ f(t_n)\right\}\right]\\
&=(1-\epsilon)\left[\frac{1}{2}\sup_{f\in\mathcal{F}}\left\{u_{n-1}(f)+\Phi\circ f(t_n)\right\}+\frac{1}{2}\sup_{f\in\mathcal{F}}\left\{u_{n-1}(f)-\Phi\circ f(t_n)\right\}\right]\\
&\leqslant\frac{1}{2}\left(u_{n-1}(f_+)+\Phi\circ f_+(t_n)+u_{n-1}(f_-)-\Phi\circ f_-(t_n)\right).
\end{aligned}
$$

Let $s=\operatorname{sign}\left(f_+(t_n)-f_-(t_n)\right)$. Then, the previous inequality implies

$$
\begin{aligned}
&(1-\epsilon)\,\mathbb{E}_{\sigma_n}\left[\sup_{f\in\mathcal{F}}\left\{u_{n-1}(f)+\sigma_n\Phi\circ f(t_n)\right\}\right]\\
&\leqslant\frac{1}{2}\left(u_{n-1}(f_+)+\Phi\circ f_+(t_n)+u_{n-1}(f_-)-\Phi\circ f_-(t_n)\right)\\
&\leqslant\frac{1}{2}\left(u_{n-1}(f_+)+u_{n-1}(f_-)+sL_\Phi\left(f_+(t_n)-f_-(t_n)\right)\right)+\frac{c}{2}\quad\text{(by hypothesis)}\\
&=\frac{1}{2}\left(u_{n-1}(f_+)+sL_\Phi f_+(t_n)\right)+\frac{1}{2}\left(u_{n-1}(f_-)-sL_\Phi f_-(t_n)\right)+\frac{c}{2}\\
&\leqslant\frac{1}{2}\sup_{f\in\mathcal{F}}\left\{u_{n-1}(f)+sL_\Phi f(t_n)\right\}+\frac{1}{2}\sup_{f\in\mathcal{F}}\left\{u_{n-1}(f)-sL_\Phi f(t_n)\right\}+\frac{c}{2}\\
&=\mathbb{E}_{\sigma_n}\left[\sup_{f\in\mathcal{F}}\left\{u_{n-1}(f)+\sigma_nL_\Phi f(t_n)\right\}\right]+\frac{c}{2}.
\end{aligned}
$$

Taking the limit for ϵ going to 0 produces:

$$
\mathbb{E}_{\sigma_n}\left[\sup_{f\in\mathcal{F}}\left\{u_{n-1}(f)+\sigma_n\Phi\circ f(t_n)\right\}\right]\leqslant\mathbb{E}_{\sigma_n}\left[\sup_{f\in\mathcal{F}}\left\{u_{n-1}(f)+\sigma_nL_\Phi f(t_n)\right\}\right]+\frac{c}{2}.
$$

Iterating the process for $i\in[\![1;n-1]\!]$ concludes the proof.

Lemma 8 *For any $p\in[1,2]$, let $\mathcal{F}_p$ be the function class given by Definition 12. Then,*

$$
\sup_{\mathbf{x}_n\in\mathcal{X}^n}\mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{f\in\mathcal{F}_p}\frac{1}{n}\sum_{i=1}^n\sigma_if(x_i)\right]\leqslant\left(\frac{5}{8}p\right)^{\frac{p}{2}}\frac{1}{n^{\frac{p}{4}}}. \tag{12}
$$

Proof We first resort to Lemma 7. Due to Lemma 6, this can be done with Φ being the function mapping t to $t^{\frac{p}{2}}$ and $L_\Phi=\frac{p}{2}K^{\frac{p-2}{p}}n^{\frac{2-p}{4}}$ (so that $c=$

$Kn^{-\frac{p}{4}}$). Then,

$$\frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{f\in\mathcal{F}_p}\sum_{i=1}^{n}\sigma_i\left\|O-\kappa_{x_i}\right\|_{\mathbf{H}_\kappa}^{p}\right]\leqslant\frac{K}{2}\frac{1}{n^{\frac{p}{4}}}+\frac{p}{2}K^{\frac{p-2}{p}}\frac{1}{n^{\frac{p+2}{4}}}\mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{f\in\mathcal{F}_2}\sum_{i=1}^{n}\sigma_i\left\|O-\kappa_{x_i}\right\|_{\mathbf{H}_\kappa}^{2}\right]. \tag{13}$$

We now bound the expectation in the right-hand side of Eq. 13.

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{f\in\mathcal{F}_2}\sum_{i=1}^{n}\sigma_i\left\|O-\kappa_{x_i}\right\|_{\mathbf{H}_\kappa}^{2}\right]\\
&\leqslant\mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{f\in\mathcal{F}_2}\sum_{i=1}^{n}\sigma_i\left\|O\right\|_{\mathbf{H}_\kappa}^{2}+\sum_{i=1}^{n}\sigma_i\left\|\kappa_{x_i}\right\|_{\mathbf{H}_\kappa}^{2}+2\sup_{f\in\mathcal{F}_2}\sum_{i=1}^{n}\sigma_i\langle O,\kappa_{x_i}\rangle_{\mathbf{H}_\kappa}\right] &(14)\\
&\leqslant\frac{1}{8}\sqrt{n}+2\mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{f\in\mathcal{F}_2}\sum_{i=1}^{n}\sigma_i\langle O,\kappa_{x_i}\rangle_{\mathbf{H}_\kappa}\right], &(15)
\end{aligned}$$

where the first expectation of Eq. 14 has been upper bounded by means of Lemma 2. The remaining expectation (in Eq. 15), associated with a class of linear functions, can be bounded by means of the Cauchy-Schwarz inequality and Jensen's inequality.

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{f\in\mathcal{F}_2}\sum_{i=1}^{n}\sigma_i\langle O,\kappa_{x_i}\rangle_{\mathbf{H}_\kappa}\right] &= \mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{f\in\mathcal{F}_2}\left\langle O,\sum_{i=1}^{n}\sigma_i\kappa_{x_i}\right\rangle_{\mathbf{H}_\kappa}\right]\\
&\leqslant\frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}_n}\left[\left\|\sum_{i=1}^{n}\sigma_i\kappa_{x_i}\right\|_{\mathbf{H}_\kappa}\right]\\
&\leqslant\frac{1}{4}\sqrt{n}.
\end{aligned}$$

Putting things together gives:

$$\mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{f\in\mathcal{F}_2}\sum_{i=1}^{n}\sigma_i\left\|O-\kappa_{x_i}\right\|_{\mathbf{H}_\kappa}^{2}\right]\leqslant\frac{5}{8}\sqrt{n}.$$

Substituting this upper bound into Eq. 13 and setting $K=\left(\frac{5}{8}p\right)^{\frac{p}{2}}$ gives (12), thus concluding the proof.

Appendix C: Proof of Lemma 1

With the technical lemmas of Appendix B at hand, the proof of Lemma 1 is the following one.

Proof For every $\mathbf{x}_n = (x_i)_{1 \leqslant i \leqslant n} \in \mathcal{X}^n$,

$$\begin{aligned}
\hat{R}_n\left(\bigcup_{p\in[1,2]} \mathcal{H}_p\right) &= \mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{h\in\bigcup_{p\in[1,2]}\mathcal{H}_p} \frac{1}{n}\sum_{i=1}^n \sigma_i h\left(x_i\right)\right] \\
&\leqslant \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{h\in\bigcup_{p\in[1,2]}\mathcal{H}_p} \sum_{i=1}^n \sigma_i R + \sup_{h\in\bigcup_{p\in[1,2]}\mathcal{H}_p} \sum_{i=1}^n \sigma_i a \left\|O-\kappa_{x_i}\right\|_{\mathbf{H}_\kappa}^p\right] \\
&= \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{R\in[0,1]} \sum_{i=1}^n \sigma_i R\right] + \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{h\in\bigcup_{p\in[1,2]}\mathcal{H}_p} \sum_{i=1}^n \sigma_i a \left\|O-\kappa_{x_i}\right\|_{\mathbf{H}_\kappa}^p\right] \\
&= \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{R\in[0,1]} \sum_{i=1}^n \sigma_i R\right] + \hat{R}_n\left(\left\{af:\ a\in(0,\Lambda], f\in\bigcup_{p\in[1,2]}\mathcal{F}_p\right\}\right).
\end{aligned} \tag{16}$$

The first Rademacher complexity in the right-hand side of Eq. 16 can be upper bounded thanks to Lemma 2, which gives:

$$\frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{R\in[0,1]} \sum_{i=1}^n \sigma_i R\right] \leqslant \frac{1}{2\sqrt{n}}.$$

For the second Rademacher complexity, we make use of the monotonicity of the Rademacher complexity with respect to the inclusion and Lemma 3. Let $\mathcal{F}_0$ be the union of the function class $\bigcup_{p\in[1,2]}\mathcal{F}_p$ and the null function.

$$\begin{aligned}
\hat{R}_n\left(\left\{af:\ a\in(0,\Lambda], f\in\bigcup_{p\in[1,2]}\mathcal{F}_p\right\}\right) &\leqslant \hat{R}_n\left(\{af:\ a\in(0,\Lambda], f\in\mathcal{F}_0\}\right) \\
&\leqslant 2\Lambda\hat{R}_n\left(\mathcal{F}_0\right).
\end{aligned}$$

The union over all the values of p involved in the definition of the class $\mathcal{F}_0$ prevents us from upper bounding directly its Rademacher complexity. To get around this difficulty, we resort to the standard approach, the use of Dudley's chaining method (Dudley 1967), precisely Theorem 9 in Guermeur (2017). It states that if δ is a positive and decreasing function on $\mathbb{N}$ such

that $\delta(0) \geqslant 1$, then

$$\forall N \in \mathbb{N}^*,\ \hat{R}_n(\mathcal{F}_0) \leqslant \delta(N) + 2\sum_{j=1}^{N} \left(\delta(j) + \delta(j-1)\right) \sqrt{\frac{\ln\left(\mathcal{N}^{\mathrm{int}}\left(\delta(j), \mathcal{F}_0, d_{2,\mathbf{x}_n}\right)\right)}{n}}. \tag{17}$$

Obviously, the covering numbers of $\mathcal{F}_0$ can be upper bounded by the covering numbers of $\bigcup_{p\in[1,2]} \mathcal{F}_p$ plus one. To obtain an initial upper bound on these latter quantities, a structural result is available: Lemma 4. In that way, the problem of handling the union of all the classes $\mathcal{F}_p$ can be taken care of, and the remaining problem boils down to upper bounding the covering numbers of the classes $\mathcal{F}_p$ (considered independently). To that end, we apply a combinatorial result: Theorem 1 in Mendelson and Vershynin (2003). This gives

$$\forall p \in [1,2],\ \forall \epsilon \in (0,1],\ \ \mathcal{N}_2^{\mathrm{int}}(\epsilon, \mathcal{F}_p, n) \leqslant \left(\frac{5}{\epsilon}\right)^{12\cdot\left(\frac{\epsilon}{24}\right)\text{-dim}(\mathcal{F}_p)}. \tag{18}$$

The fat-shattering dimensions ϵ-dim$(\mathcal{F}_p)$ can be upper bounded in terms of the corresponding Rademacher complexities by means of Lemma 5. Here appears the point of introducing the metric entropies (by means of Dudley's chaining method) and applying a structural result on covering numbers (Lemma 4): even though we cannot upper bound $\hat{R}_n(\mathcal{F}_0)$ directly, an upper bound on the Rademacher complexities of the classes $\mathcal{F}_p$ is available: Lemma 8. Thus, we obtain:

$$\begin{aligned}\forall p \in [1,2],\ \forall \epsilon \in (0,1],\ \ \epsilon\text{-dim}(\mathcal{F}_p) &\leqslant \left(\frac{5}{8}p\right)^2 \left(\frac{1}{\epsilon}\right)^{\frac{4}{p}} \\ &\leqslant 2\left(\frac{1}{\epsilon}\right)^4. \end{aligned} \tag{19}$$

Note that according to Mendelson's terminology, the classes $\mathcal{F}_p$ have a polynomial fat-shattering dimension with degree 4. This enables us to apply Theorem 18 in Mendelson (2003) so as to express in advance the dependence on n of our bound on $R_n(\mathcal{F}_0)$: a $O\left(\left(\frac{\ln(n)}{n}\right)^{\frac{1}{4}}\right)$. Applying in sequence Inequalities Eqs. 9, 18 and 19 produces the bound on the metric entropies of interest:

$$\forall \epsilon \in (0,1],\ \ \log_2\left(\mathcal{N}^{\mathrm{int}}(\epsilon, \mathcal{F}_0, d_{2,\mathbf{x}_n})\right) \leqslant \log_2\left(\mathcal{N}^{\mathrm{int}}\left(\epsilon, \bigcup_{p\in[1,2]} \mathcal{F}_p, d_{2,\mathbf{x}_n}\right) + 1\right)$$

$$\begin{aligned}
&\leqslant \log_2\left(\left\lceil\frac{1}{\epsilon}\right\rceil \max_{p\in[1,2]} \mathcal{N}^{\text{int}}\left(\frac{\epsilon}{2}, \mathcal{F}_p, d_{2,\mathbf{x}_n}\right)+1\right)\\
&\leqslant \log_2\left(2\left\lceil\frac{1}{\epsilon}\right\rceil\right) + 12 \max_{p\in[1,2]} \left(\frac{\epsilon}{48}\right)\text{-dim}\left(\mathcal{F}_p\right)\log_2\left(\frac{10}{\epsilon}\right)\\
&\leqslant \log_2\left(2\left\lceil\frac{1}{\epsilon}\right\rceil\right) + 24\left(\frac{48}{\epsilon}\right)^4\log_2\left(\frac{10}{\epsilon}\right)\\
&\leqslant 25\left(\frac{48}{\epsilon}\right)^4\log_2\left(\frac{10}{\epsilon}\right). \qquad (20)
\end{aligned}$$

This implies that

$$\forall\epsilon\in(0,1]\,,\quad \sqrt{\ln\left(\mathcal{N}^{\text{int}}\left(\epsilon,\mathcal{F}_0,d_{2,\mathbf{x}_n}\right)\right)} \leqslant K\frac{1}{\epsilon^2}\sqrt{\ln\left(\frac{10}{\epsilon}\right)},$$

where $K = 11520$. A substitution of this bound into Eq. 17 gives:

$$\forall N\in\mathbb{N}^*,\ \hat{R}_n\left(\mathcal{F}_0\right) \leqslant \delta\left(N\right) + \frac{2K}{\sqrt{n}}\sum_{j=1}^{N}\frac{\delta\left(j\right)+\delta\left(j-1\right)}{\delta^2\left(j\right)}\sqrt{\ln\left(\frac{10}{\delta\left(j\right)}\right)}.$$

Since the right-hand side does not depend on $\mathbf{x}_n$, the empirical Rademacher complexity can be replaced with the Rademacher complexity. For $N = \left\lceil\frac{1}{4}\log_2\left(\frac{n}{\log_2(n)}\right)\right\rceil$, let us set $\delta\left(j\right) = \left(\frac{\log_2(n)}{n}\right)^{\frac{1}{4}}2^{N-j}$. We then get

$$\begin{aligned}
R_n\left(\mathcal{F}_0\right) &\leqslant \left(\frac{\log_2\left(n\right)}{n}\right)^{\frac{1}{4}} + \frac{6K}{\sqrt{n}}\sum_{j=1}^{N}\frac{1}{\delta\left(j\right)}\sqrt{\ln\left(\frac{10}{\delta\left(j\right)}\right)}\\
&\leqslant \left(\frac{\log_2\left(n\right)}{n}\right)^{\frac{1}{4}}\left[1+\frac{6K}{\sqrt{\log_2\left(n\right)}}\sqrt{\ln\left(10\left(\frac{n}{\log_2\left(n\right)}\right)^{\frac{1}{4}}\right)}\sum_{j=1}^{N}2^{j-N}\right]\\
&\leqslant \left(\frac{\log_2\left(n\right)}{n}\right)^{\frac{1}{4}}\left[1+\frac{12K}{\sqrt{\log_2\left(n\right)}}\sqrt{\ln\left(10\left(\frac{n}{\log_2\left(n\right)}\right)^{\frac{1}{4}}\right)}\right].
\end{aligned}$$

Collecting all terms gives:

$$R_n\left(\bigcup_{p\in[1,2]}\mathcal{H}_p\right) \leqslant \frac{1}{2\sqrt{n}} + 2\Lambda\left(\frac{\log_2\left(n\right)}{n}\right)^{\frac{1}{4}}\left[1+\frac{12K}{\sqrt{\log_2\left(n\right)}}\sqrt{\ln\left(10\left(\frac{n}{\log_2\left(n\right)}\right)^{\frac{1}{4}}\right)}\right],$$

i.e., Eq. 5, thus concluding the proof.

Isotropic Kernel Machine

Yann Guermeur
LORIA-CNRS, Campus Scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy Cedex, France
E-mail: yann.guermeur@loria.fr

Nicolas Wicker
Department of Mathematics, University of Lille, Cité Scientifique, 59655 Villeneuve d'Ascq, France
E-mail: nicolas.wicker@univ-lille.fr

Paper received: 30 December 2022