Rademacher Complexity and Generalization Performance of Multi-category Margin Classifiers

Khadija Musayeva, Fabien Lauer, Yann Guermeur

Rademacher Complexity and Generalization Performance of Multi-category Margin Classifiers

Khadija Musayeva, Fabien Lauer and Yann Guermeur

December 3, 2018

Abstract

One of the main open problems in the theory of multi-category margin classification is the form of the optimal dependency of a guaranteed risk on the number C of categories, the sample size m and the margin parameter γ. From a practical point of view, the theoretical analysis of generalization performance contributes to the development of new learning algorithms. In this paper, we focus only on the theoretical aspect of the question posed. More precisely, under minimal learnability assumptions, we derive a new risk bound for multi-category margin classifiers. We improve the dependency on C over the state of the art when the margin loss function considered satisfies the Lipschitz condition. We start with the basic supremum inequality that involves a Rademacher complexity as a capacity measure. This capacity measure is then linked to the metric entropy through the chaining method. In this context, our improvement is based on the introduction of a new combinatorial metric entropy bound.

1 Introduction

Although the theory of binary pattern classification is well established [1, 2], the theory of multi-category classification is far from being complete. The research in this case addresses problems such as the sample-complexity analysis of empirical risk minimization algorithms [3], or consistency analysis of multi-class loss functions and of specific families of classifiers [4]. Another open question is the optimal dependency of guaranteed risks of multi-category classifiers on the number C of categories and the sample size m. It is all the more the case for the problems that involve a large number of classes. When the considered classifiers are margin ones that take decision based on a score per category, the dependency on the margin parameter γ also becomes relevant to the characterization of their generalization performance. If this question has been mainly studied for specific families of classifiers, be it k-nearest neighbors [5], kernel methods [6, 7] and decision trees [8], tackling it under minimal learnability assumptions remains a challenging task. This paper focuses on obtaining guaranteed risks under such assumptions.

The first step in the derivation of risk bounds is the choice of the margin loss function. Two families of margin loss functions can be distinguished: indicator margin loss functions and those that satisfy the Lipschitz condition. Deriving guaranteed risks with the optimal dependency on the parameters of interest is relatively straightforward in the first case [9]. The family of Lipschitz continuous loss functions, on the other hand, offers a richer setting to this task. In this case, one can obtain a guaranteed risk whose control term involves a Rademacher complexity [10]. Then a sequence of transitions between capacity measures is performed. More precisely, using the chaining method one can control the Rademacher complexity of a function class through the sum of its metric entropies [11]. A combinatorial bound is then used to estimate the metric entropy of the class in terms of its combinatorial dimension. In this sequence of transitions, one can choose the capacity measure at the level of which to reduce the multi-class problem to an ensemble of bi-class ones, that is, to perform a *decomposition*. Performing a decomposition for Rademacher complexity, a linear dependency on C was obtained in [8]. This dependency has been improved to a sublinear one in [9] by postponing the decomposition to the level of metric entropy.

In this paper, we exactly follow the pathway of [9]. Our contribution is based on the following line of reasoning. Theorem 7 of [9] provides a sublinear (but still close to linear) dependency on C using a decomposition result for metric entropies (Lemma 1 of [9]) in L_p-norm with $p = 2$ and the combinatorial metric entropy bound of [12]. On the other hand, using the decomposition result with $p = \infty$ and the L_∞-norm metric entropy bound of [13], one can obtain a radical dependency on C, this, however, at the expense of a degraded dependency on m. Hence, we consider the values of p in between these two extreme ones, and extend the L_2-norm bound of [12] to L_p-norms with integer $p > 2$. When applied in the chaining, it results in an improved dependency on C over that of Theorem 7 of [9]. Specifically, we obtain a radical dependency on C (up to logarithmic factors) without worsening the dependencies on m and γ.

The organization of the paper is as follows. In the next section, we introduce the theoretical framework and describe the transitions between the capacity measures. Then, Section 3 gives the new combinatorial metric entropy bound, whose proof can be found in A. In Section 4, we demonstrate how this result can be applied in the chaining to derive an improved upper bound on the Rademacher complexity. Conclusions and ongoing research are highlighted in Section 5. All intermediate results used in the proofs are collected in B.

Notation We denote the set of strictly positive reals by $\mathbb{R}_+$, and let $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$. $[\![i, j]\!]$ stands for the set of integers from i to j. $\mathbb{1}_A$ stands for the indicator function for the event A such that $\mathbb{1}_A = 1$ if A occurs, and 0 otherwise. $\lfloor x \rfloor$ is the greatest integer less than or equal to x, $\lceil x \rceil$ is the smallest integer greater than or equal to x.

2 Theoretical Framework

We consider C-category pattern classification problems with $C \geqslant 3$. Each object is represented by its description $x \in \mathcal{X}$ and the categories y belong to $\mathcal{Y} = [\![1, C]\!]$. We assume that $(\mathcal{X}, \mathcal{A}_{\mathcal{X}})$ and $(\mathcal{Y}, \mathcal{A}_{\mathcal{Y}})$ are measurable spaces. Denote by $\mathcal{A}_{\mathcal{X}} \otimes \mathcal{A}_{\mathcal{Y}}$ the product sigma-algebra on $\mathcal{X} \times \mathcal{Y}$. We assume that the link between descriptions and categories can be characterized by an unknown probability measure P 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 with values in $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, distributed according to P. The available information on P is limited to an m-sample $\mathbf{Z}_m = (Z_i)_{1 \leqslant i \leqslant m} = ((X_i, Y_i))_{1 \leqslant i \leqslant m}$ distributed according to P^m. In the following, we distinguish the sample size m from the generic notation n which stands for a number of points in a set that needs not be a realization of a random sample.

We consider multi-category margin classifiers that take their decisions based on a score per category and focus on those that implement classes of functions with values in a hypercube of $\mathbb{R}^C$ (thus, in contrast to [7], no correlation assumption is made on the component functions). Most well-known classifiers, such as neural networks [14], support vector machines [4], and nearest neighbors [5] are margin classifiers.

Definition 1 (Multi-category margin classifiers)**.** *Let $\mathcal{G} = \prod_{k=1}^{C} \mathcal{G}_k$ be a class of functions from $\mathcal{X}$ into $[-M_{\mathcal{G}}, M_{\mathcal{G}}]^C$ with $M_{\mathcal{G}} \in [1, +\infty)$. For each $g = (g_k)_{1 \leqslant k \leqslant C} \in \mathcal{G}$, dr_g is a* multi-category margin classifier *such that for all $x \in \mathcal{X}$, $dr_g(x) = \operatorname{argmax}_{1 \leqslant k \leqslant C} g_k(x)$, breaking ties with a dummy category $*$.*

To sidestep the complications that might arise from the measurability of a supremum of an uncountable set, we assume that the classes $\mathcal{G}_k$, and in general, all sets of functions considered in the sequel satisfy the "image admissibility Suslin" condition [15, page 101].

The classification performance of margin classifiers can be characterized based on the following functions.

Definition 2 (Class $\mathcal{F}_{\mathcal{G}}$ of margin functions)**.** *Let $\mathcal{G}$ be as in Definition 1. For any $g \in \mathcal{G}$, the margin function $f_g : \mathcal{Z} \to [-M_{\mathcal{G}}, M_{\mathcal{G}}]$ is*

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

Then, we define $\mathcal{F}_{\mathcal{G}} = \{f_g : g \in \mathcal{G}\}$.

Given $g \in \mathcal{G}$, dr_g misclassifies (x, y) if $dr_g(x) \neq y$, or equivalently, if $f_g(x, y) \leqslant 0$. The goal of the learning process is to minimize the probability of error or *risk* over $\mathcal{G}$.

Definition 3 (Risk L)**.** *Let $\mathcal{G}$ be as in Definition 1. Let ϕ be the standard indicator loss function defined as*

$$\forall t \in \mathbb{R}, \ \phi(t) = \mathbb{1}_{\{t \leqslant 0\}}.$$

For any $g \in \mathcal{G}$, its risk $L(g)$ is

$$L(g) = \mathbb{E}_Z \left[\phi\left(f_g\left(Z\right)\right)\right] = P\left(dr_g(X) \neq Y\right).$$

To make use of the values of functions f_g (and not just of their signs) in the assessment of the classification performance, we appeal to the following margin loss function.

Definition 4 (Parameterized truncated hinge loss function ϕ_γ)**.** *For any $\gamma \in (0, 1]$, the parameterized truncated hinge loss function ϕ_γ is defined as*

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

It is clear from the definition that ϕ_γ dominates the standard indicator loss function given in Definition 3 and that it is Lipschitz continuous. Observe that when this loss function is applied to f_g, the values of the latter strictly above γ and below zero become irrelevant to the estimation of the classification accuracy. Taking benefit from this fact, we introduce functions $f_{g,\gamma}$ by restricting the codomain of f_g to $[0, \gamma]$ for all $g \in \mathcal{G}$. In [8], a partial restriction is the main source of improvement upon the result of [10] in terms of the dependency on C. The use of the set of functions $f_{g,\gamma}$ leads to even a finer bound, this time in terms of the diameter of the function class as we switch from $2M_{\mathcal{G}}$ to γ.

Definition 5 (Class $\mathcal{F}_{\mathcal{G},\gamma}$ of truncated margin functions)**.** *Let $\mathcal{F}_{\mathcal{G}}$ be a class of functions satisfying Definition 2. Fix $\gamma \in (0, 1]$. For any $f_g \in \mathcal{F}_{\mathcal{G}}$, we define $f_{g,\gamma} : \mathcal{Z} \to [0, \gamma]$ as*

$$\forall (x, k) \in \mathcal{Z},\ f_{g,\gamma}\left(x, k\right) = \max(0, \min\left(\gamma, f_g\left(x, k\right)\right)),$$

and $\mathcal{F}_{\mathcal{G},\gamma} = \{f_{g,\gamma} : g \in \mathcal{G}\}$.

For any $g \in \mathcal{G}$, its risk, $L(g)$ can be upper bounded by the margin risk $L_\gamma(g)$ obtained on the basis of the loss function ϕ_γ. It is the m-sample $\mathbf{Z}_m$ based estimate of L_γ that appears in our guaranteed risk.

Definition 6 (Margin risk L_γ and empirical margin risk $L_{\gamma,m}$)**.** *Let $\mathcal{G}$ be a class of functions satisfying Definition 1. Let ϕ_γ be as in Definition 4. Then, for $\gamma \in (0, 1]$, the margin risk L_γ associated with any $g \in \mathcal{G}$ is*

$$L_\gamma(g) = \mathbb{E}_Z \left[\phi_\gamma\left(f_{g,\gamma}\left(Z\right)\right)\right].$$

Its m-sample $\mathbf{Z}_m$ based estimate is the empirical margin risk defined as

$$L_{\gamma,m}(g) = \frac{1}{m} \sum_{i=1}^{m} \phi_\gamma\left(f_{g,\gamma}\left(Z_i\right)\right).$$

In what follows, we give the definitions of the capacity measures we use and outline the transitions between them, which are at the basis of the derivation of our result. We use $\mathcal{F}$ to denote a uniformly bounded class of functions on a generic measurable space $(\mathcal{T}, \mathcal{A}_{\mathcal{T}})$. First, we recall the definition of the Rademacher complexity.

Definition 7 (Rademacher complexity)**.** *Let $P_{\mathcal{T}}$ be a probability measure on $(\mathcal{T}, \mathcal{A}_{\mathcal{T}})$ and $\mathbf{T}_n = (T_i)_{1 \leqslant i \leqslant n}$ a sequence of independently distributed according to $P_{\mathcal{T}}$ random variables with values in $\mathcal{T}$. Let $\boldsymbol{\sigma}_n = (\sigma_i)_{1 \leqslant i \leqslant n}$ be a Rademacher sequence, i.e.,a sequence of independent random variables uniformly distributed in $\{-1, +1\}$. Then, the empirical Rademacher complexity of $\mathcal{F}$ given $\mathbf{T}_n$ is defined as*

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

and its Rademacher complexity is $R_n(\mathcal{F}) = \mathbb{E}_{\mathbf{T}_n}\left[\hat{R}_n(\mathcal{F})\right]$.

The capacity measures central in the derivation of our result are covering/packing numbers. Their definitions require the introduction of the following empirical pseudo-metrics: for any $f, f' \in \mathcal{F}$ and $\mathbf{t}_n = (t_i)_{1 \leqslant i \leqslant n} \in \mathcal{T}^n$,

$$d_{p,\mathbf{t}_n}(f, f') = \begin{cases} \left(\frac{1}{n} \sum_{i=1}^{n} |f(t_i) - f'(t_i)|^p\right)^{\frac{1}{p}}, \text{if } p \in [1, +\infty) \\ \max_{1 \leqslant i \leqslant n} |f(t_i) - f'(t_i)|, \text{if } p = +\infty. \end{cases}$$

Definition 8 (Covering numbers, metric entropy, packing numbers)**.** *The L_p-norm* ϵ-covering number *of $\mathcal{F}$, $\mathcal{N}(\epsilon, \mathcal{F}, d_{p,\mathbf{t}_n})$, is the smallest cardinality of the ϵ-nets of $\mathcal{F}$, i.e., subsets $\bar{\mathcal{F}} \subseteq \mathcal{F}$ such that $\forall f \in \mathcal{F}$ there exists $\bar{f} \in \bar{\mathcal{F}}$ such that $d_{p,\mathbf{t}_n}(f, \bar{f}) < \epsilon$. The logarithm of $\mathcal{N}(\epsilon, \mathcal{F}, d_{p,\mathbf{t}_n})$ is the* metric entropy *of $\mathcal{F}$. A subset $\bar{\mathcal{F}}$ of $\mathcal{F}$ is* ϵ-separated *with respect to $d_{p,\mathbf{t}_n}$ if, for any two distinct elements $f, f' \in \bar{\mathcal{F}}$, $d_{p,\mathbf{t}_n}(f, f') \geqslant \epsilon$. The ϵ-packing number of $\mathcal{F}$, $\mathcal{M}(\epsilon, \mathcal{F}, d_{p,\mathbf{t}_n})$, is the maximal cardinality of its ϵ-separated subsets. The uniform covering and packing numbers are*

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

and

$$\mathcal{M}_p(\epsilon, \mathcal{F}, n) = \sup_{\mathbf{t}_n \in \mathcal{T}^n} \mathcal{M}(\epsilon, \mathcal{F}, d_{p,\mathbf{t}_n}),$$

respectively.

The capacity measures appearing last in our bounds are combinatorial dimensions. They provide useful information about whether the class of interest uniformly satisfies the classical limit theorems [16].

Definition 9 (Fat-shattering dimension [17], strong dimension [13])**.** *For $\gamma \in \mathbb{R}_+$, a subset $S = \{t_i : 1 \leqslant i \leqslant n\}$ of $\mathcal{T}$ is said to be* γ-shattered *by $\mathcal{F}$ if there is a function $v : S \to \mathbb{R}$ 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*

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

The fat-shattering dimension *of $\mathcal{F}$ at scale γ, γ-dim$(\mathcal{F})$, is the maximal cardinality of a subset of $\mathcal{T}$ γ-shattered by $\mathcal{F}$, if such a maximum exists. Otherwise, γ-dim$(\mathcal{F}) = \infty$. For a class $\mathcal{F}$ of integer valued functions, the notion of* strong dimension*, S-dim$(\mathcal{F})$, is obtained from the definition of the fat-shattering dimension by setting $\gamma = 1$ and restricting the co-domain of v to $\mathbb{Z}$.*

As in [9, 18, 19], we make the hypothesis that the fat-shattering dimensions of the classes $\mathcal{G}_k$, γ-dim $(\mathcal{G}_k)$, grow no faster than polynomially with γ^{-1}.

Hypothesis 1. *Let $\mathcal{G}$ be a class of functions satisfying Definition 1. We assume that there exists a pair $(K_{\mathcal{G}}, d_{\mathcal{G}}) \in \mathbb{R}_+^2$ such that*

$$\forall \epsilon \in (0, M_{\mathcal{G}}], \quad \max_{1 \leqslant k \leqslant C} \epsilon\text{-}dim\,(\mathcal{G}_k) \leqslant K_{\mathcal{G}} \epsilon^{-d_{\mathcal{G}}}.$$

Among the well-known examples of classifiers that satisfy such an assumption are support vector machines with $d_{\mathcal{G}} = 2$ (Theorem 4.6 in [20]) and feedforward neural networks with $d_{\mathcal{G}} = 2l$ for l layers (Corollary 27 in [2]). It should be noted that Lipschitz classifiers, such as nearest neighbours also satisfy this assumption as demonstrated by Corollary 4 in [21]. Depending on the growth rate $d_{\mathcal{G}}$, our assumptions regarding the data are summarized in Table 1.

Table 1: Assumptions made on the sample size m and the number of categories C with respect to the growth rate $d_{\mathcal{G}}$ of the fat-shattering dimensions in Hypothesis 1.

Growth rate	Assumptions
$d_{\mathcal{G}} \leqslant 2$	$m > C > 4$
$d_{\mathcal{G}} > 2$	$m \geqslant C^{1.2}$, $C > 4$

Our starting point is the following basic supremum inequality that bounds the risk by the empirical margin risk plus a control term based on a Rademacher complexity.

Theorem 1 (Theorem 5 in [9])**.** *Let $\mathcal{G}$ be a class of functions satisfying Definition 1. For $\gamma \in (0, 1]$, let $\mathcal{F}_{\mathcal{G},\gamma}$ be the class of functions deduced from $\mathcal{G}$ according to Definition 2. For fixed $\gamma \in (0, 1]$ and $\delta \in (0, 1)$, with P^m probability at least $1 - \delta$,*

$$\forall g \in \mathcal{G}, \quad L(g) \leqslant L_{\gamma,m}(g) + \frac{2}{\gamma} R_m\left(\mathcal{F}_{g,\gamma}\right) + \sqrt{\frac{\ln(\frac{1}{\delta})}{2m}}.$$

We perform the following sequence of transitions between the capacity measures to derive our result. First, we relate the empirical Rademacher complexity of $\mathcal{F}_{\mathcal{G},\gamma}$ to its metric entropy through the chaining method (see [11]). More precisely, we use the following formulation of the chaining bound due to [9]:

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant h(N) + 2\sum_{j=1}^{N} \left(h(j) + h(j-1)\right) \sqrt{\frac{\ln \mathcal{N}\left(h(j), \mathcal{F}_{\mathcal{G},\gamma}, d_{2,\mathbf{z}_m}\right)}{m}}, \tag{1}$$

where $N \in \mathbb{N}^*$ and $h : \mathbb{N} \to \mathbb{R}_+$ is a decreasing function satifying $h(0) \geqslant \gamma$. Next, using Lemma 1 in [9], we decompose the metric entropy of $\mathcal{F}_{\mathcal{G},\gamma}$ in terms of the ones of the classes $\mathcal{G}_k$:

$$\forall p \in [1, +\infty], \quad \ln \mathcal{N}\left(\epsilon, \mathcal{F}_{\mathcal{G},\gamma}, d_{p,\mathbf{z}_m}\right) \leqslant \sum_{k=1}^{C} \ln \mathcal{N}\left(\frac{\epsilon}{C^{1/p}}, \mathcal{G}_k, d_{p,\mathbf{x}_m}\right), \tag{2}$$

where $\mathbf{x}_m = (x_i)_{1 \leqslant i \leqslant m} \in \mathcal{X}^m$. Finally, our combinatorial bound derived below gives an estimate on the metric entropies of the classes $\mathcal{G}_k$ in terms of their fat-shattering dimensions.

3 L_p-norm Combinatorial Metric Entropy Bound

We extend the L_2-norm metric entropy bound of [12] to L_p-norms with $p \in \mathbb{N}^* \setminus \{1,2\}$. The bound of [12] does not depend on the sample size thanks to the use of the probabilistic extraction principle. In our extension we derive two bounds. In one of them, we keep the dependency on the sample size, and in the other, we remove it using the L_p-norm generalization of the aforementioned principle. Under Hypothesis 1, depending on the value of $d_{\mathcal{G}}$, the application of one or the other bound in the chaining allows us to optimize the dependency on C while not degrading the ones on m and γ, as will be seen in Section 4.

Specifically, we have the following L_p-norm metric entropy bounds, whose proof is given in A.

Theorem 2. *Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ into $[-M_{\mathcal{F}}, M_{\mathcal{F}}]$ with $M_{\mathcal{F}} \in [1, +\infty)$. For $\epsilon \in (0, M_{\mathcal{F}}]$, let $d(\epsilon) = \epsilon\text{-}dim(\mathcal{F})$. For all values of $p \in \mathbb{N}^* \setminus \{1,2\}$ and $\epsilon \in (0, M_{\mathcal{F}}]$,*

(a) if $n \geqslant d\left(\frac{\epsilon}{15p}\right)$, then

$$\ln \mathcal{N}_p(\epsilon, \mathcal{F}, n) \leqslant 2d\left(\frac{\epsilon}{15p}\right) \ln\left(\frac{15epnM_{\mathcal{F}}}{d\left(\frac{\epsilon}{15p}\right)\epsilon}\right);$$

(b) if $n \geqslant d\left(\frac{\epsilon}{37p}\right)$, then

$$\ln \mathcal{N}_p(\epsilon, \mathcal{F}, n) \leqslant 10p\, d\left(\frac{\epsilon}{36p}\right) \ln\left(\frac{7p^{\frac{1}{7}} M_{\mathcal{F}}}{\epsilon}\right).$$

From (2) one can see that, based on $C^{\frac{1}{p}} = 2^{\left(\frac{1}{p}\log_2(C)\right)}$, the dependency on C in the scale of covering numbers can be eliminated for all $p \geqslant \log_2(C)$. The combination of the decomposition formula (2) with Theorem 2 using $p = \lceil \log_2(C) \rceil$ for $C > 4$ yields the following result.

Corollary 1. *Let $\mathcal{G}$ be a class of functions as in Definition 1. For $\gamma \in (0,1]$, let $\mathcal{F}_{\mathcal{G},\gamma}$ be the class of functions deduced from $\mathcal{G}$ according to Definition 5. For $\epsilon \in (0, M_{\mathcal{G}}]$, let $d(\epsilon) = \max_{1 \leqslant k \leqslant C} \epsilon\text{-}dim(\mathcal{G}_k)$. Then, for $\epsilon \in (0, \gamma]$ and $C > 4$,*

$$\ln \mathcal{N}_p(\epsilon, \mathcal{F}_{\mathcal{G},\gamma}, m) \leqslant 2Cd\left(\frac{\epsilon}{30\log_2(2C)}\right) \ln\left(\frac{30en\log_2(2C)\, M_{\mathcal{G}}}{\epsilon}\right), \tag{3}$$

and

$$\ln \mathcal{N}_p(\epsilon, \mathcal{F}_{\mathcal{G},\gamma}, m) \leqslant 10C\log_2(2C)\, d\left(\frac{\epsilon}{72\log_2(2C)}\right) \ln\left(\frac{14\log_2^{\frac{1}{7}}(2C)\, M_{\mathcal{G}}}{\epsilon}\right). \tag{4}$$

Proof. Inequality (3) follows from the application of (2) and part (a) of Theorem 2 (where we drop $d(\epsilon)$ from the denominator inside the logarithm as it is greater than one), along with the fact that $C^{1/\lceil \log_2(C) \rceil} < 2$ and $\lceil \log_2(C) \rceil < \log_2(2C)$. We obtain Inequality (4) in a similar way using part (b) of Theorem 2 instead. □

4 Bound on the Rademacher complexity

As it was noted in [18], under Hypothesis 1, the growth rate of the fat-shattering dimension has a dramatic effect on the behavior of the Rademacher complexity of the function class. The availability of two kinds of metric entropy bounds allows us to adapt to this impact in the chaining so as to optimize the dependency on C without worsening those on m and γ. Under the aforementioned hypothesis, two cases can be distinguished. For $d_{\mathcal{G}} \in (0,2)$, the formula (1) can be upper bounded by an integral and the use of the dimension-free bound (4) leads to the optimized result. For $d_{\mathcal{G}} \geqslant 2$, such a result is obtained from the application of (3) in (1). The second case can also be characterized by the fact that there is a freedom in the choice of the number N of steps to construct the chaining. To optimize this construction when $d_{\mathcal{G}} > 2$, we make the non-restrictive assumption that m is greater than a small power of C.

Theorem 3. *Let $\mathcal{G}$ be a class of functions as in Definition 1. For $\gamma \in (0,1]$, let $\mathcal{F}_{\mathcal{G},\gamma}$ be the class of functions deduced from $\mathcal{G}$ according to Definition 2. Then, under Hypothesis 1, there is a function $K(\gamma, d_{\mathcal{G}}, K_{\mathcal{G}})$ such that for all $C > 4$,*

$$R_m(\mathcal{F}_{\mathcal{G},\gamma}) \leqslant K(\gamma, d_{\mathcal{G}}, K_{\mathcal{G}}) \sqrt{\frac{C}{m}} \times \begin{cases} (\ln(C))^{\frac{d_{\mathcal{G}}}{2}+\frac{1}{2}}, & \text{if } 0 < d_{\mathcal{G}} < 2, \\ \ln(C) \ln\left(\frac{m}{C}\right) \ln^{\frac{1}{2}}\left(\frac{m \ln^{\frac{2}{3}}(C)}{C^{\frac{1}{3}}}\right), & \text{if } d_{\mathcal{G}} = 2, \\ m^{\frac{1}{2}-\frac{1}{d_{\mathcal{G}}}} (\ln(C))^{2-\frac{d_{\mathcal{G}}}{2}} \ln^{\frac{1}{2}}\left(\frac{m^{1+\frac{1}{d_{\mathcal{G}}}}}{\ln(C)}\right), & \text{if } d_{\mathcal{G}} > 2 \text{ and } m \geqslant C^{1.2}. \end{cases}$$

Compared to Theorem 7 of [9], one can see that in all three cases, the dependency on C is improved: the powers of C are replaced by powers of $\ln(C)$ without losing in the dependencies on m and γ. It is interesting to note that, in the third case, when $d_{\mathcal{G}} \geqslant 4$, which is true for instance for feedforward neural networks (see Corollary 27 in [2]), the dependency on C is slightly better than radical. This is, however, at the cost of the constant factor $d_{\mathcal{G}}^{d_{\mathcal{G}}}$.

Proof of Theorem 3. For all $j \in \mathbb{N}$, we set $h(j) = \gamma 2^{-\alpha(d_{\mathcal{G}})j}$ with $\alpha(d_{\mathcal{G}}) > 0$ for all $d_{\mathcal{G}} \in \mathcal{R}_+^*$ in (1). In the following, we use the relation

$$\forall r > q > 0, \quad \mathcal{N}(\epsilon, \mathcal{F}, d_{q,\mathbf{t}_n}) \leqslant \mathcal{N}(\epsilon, \mathcal{F}, d_{r,\mathbf{t}_n}) \tag{5}$$

which follows directly from the fact that

$$\forall f, f' \in \mathcal{F}, \quad d_{q,\mathbf{t}_n}(f, f') \leqslant d_{r,\mathbf{t}_n}(f, f').$$

First case: $d_{\mathcal{G}} \in (0,2)$ This is the only case where Pollard's entropy condition [16] is satisfied. For this case we could directly use Dudley's integral formula (Formula 33 in [9]), however, to

optimize with respect to constants, we start from (1) and upper bound it by an integral in the following way.

Apply (5) and (4) in sequence to the right-hand side of (1) and use Hypothesis 1 to get

$$\begin{aligned}\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant & \gamma 2^{-\alpha(d_{\mathcal{G}})N}+2\sqrt{\frac{10C\log_2(2C)}{m}}\sum_{j=1}^{N}\left(\gamma 2^{-\alpha(d_{\mathcal{G}})j}+\gamma 2^{-\alpha(d_{\mathcal{G}})(j-1)}\right)\\ & \times\left[d\left(\frac{\gamma 2^{-\alpha(d_{\mathcal{G}})j}}{72\log_2(2C)}\right)\ln\left(\frac{14M_{\mathcal{G}}\log_2^{\frac{1}{7}}(2C)}{\gamma 2^{-\alpha(d_{\mathcal{G}})j}}\right)\right]^{1/2}\\ \leqslant & \gamma 2^{-\alpha(d_{\mathcal{G}})N}+2\sqrt{\frac{10C\log_2(2C)K_{\mathcal{G}}}{m}}\left(72\log_2(2C)\right)^{\frac{d_{\mathcal{G}}}{2}}\gamma^{1-\frac{d_{\mathcal{G}}}{2}}\left(1+2^{\alpha(d_{\mathcal{G}})}\right)\\ & \times\sum_{j=1}^{N}2^{-\alpha(d_{\mathcal{G}})\left(1-\frac{d_{\mathcal{G}}}{2}\right)j}\ln^{\frac{1}{2}}\left(\frac{14M_{\mathcal{G}}\log_2^{\frac{1}{7}}(2C)}{\gamma 2^{-\alpha(d_{\mathcal{G}})j}}\right).\end{aligned}$$

Letting $\alpha(d_{\mathcal{G}})=\dfrac{2}{2-d_{\mathcal{G}}}$, we obtain

$$\begin{aligned}\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant & \gamma 2^{-\frac{2}{2-d_{\mathcal{G}}}N}+2\sqrt{\frac{10C\log_2(2C)K_{\mathcal{G}}}{m}}\left(72\log_2(2C)\right)^{\frac{d_{\mathcal{G}}}{2}}\gamma^{1-\frac{d_{\mathcal{G}}}{2}}\left(1+2^{\frac{2}{2-d_{\mathcal{G}}}}\right)\\ & \times\sum_{j=1}^{N}2^{-j}\ln^{\frac{1}{2}}\left(\frac{14M_{\mathcal{G}}\log_2^{\frac{1}{7}}(2C)}{\gamma 2^{-\frac{2}{2-d_{\mathcal{G}}}j}}\right)\\ = & \gamma 2^{-\frac{2}{2-d_{\mathcal{G}}}N}+4\sqrt{\frac{10C\log_2(2C)K_{\mathcal{G}}}{m}}\left(72\log_2(2C)\right)^{\frac{d_{\mathcal{G}}}{2}}\gamma^{1-\frac{d_{\mathcal{G}}}{2}}\left(1+2^{\frac{2}{2-d_{\mathcal{G}}}}\right)\\ & \times\sum_{j=1}^{N}\left(2^{-j}-2^{-j-1}\right)\ln^{\frac{1}{2}}\left(\frac{14M_{\mathcal{G}}\log_2^{\frac{1}{7}}(2C)}{\gamma 2^{-\frac{2}{2-d_{\mathcal{G}}}j}}\right).\end{aligned}$$

Taking $N\to\infty$, we can upper bound the last expression as

$$\begin{aligned}\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant & 4\sqrt{\frac{10C\log_2(2C)K_{\mathcal{G}}}{m}}\left(72\log_2(2C)\right)^{\frac{d_{\mathcal{G}}}{2}}\gamma^{1-\frac{d_{\mathcal{G}}}{2}}\left(1+2^{\frac{2}{2-d_{\mathcal{G}}}}\right)\\ & \times\int_0^{1/2}\ln^{\frac{1}{2}}\left(\frac{14M_{\mathcal{G}}\log_2^{\frac{1}{7}}(2C)}{\gamma\epsilon^{\frac{2}{2-d_{\mathcal{G}}}}}\right)d\epsilon.\end{aligned}$$

Denote $K=14M_{\mathcal{G}}\log_2^{\frac{1}{7}}(2C)/\gamma$ and let us now compute the integral

$$L=\int_0^{1/2}\ln^{\frac{1}{2}}\left(K/\epsilon^{\frac{2}{2-d_{\mathcal{G}}}}\right)d\epsilon=\sqrt{\frac{2}{2-d_{\mathcal{G}}}}\int_0^{1/2}\ln^{\frac{1}{2}}\left(\frac{K^{\frac{2-d_{\mathcal{G}}}{2}}}{\epsilon}\right)d\epsilon.$$

Set $\epsilon=K^{\frac{2-d_{\mathcal{G}}}{2}}e^{-t^2}$. Then,

$$L=\sqrt{\frac{2}{2-d_{\mathcal{G}}}}K^{\frac{2-d_{\mathcal{G}}}{2}}\int_{\ln^{\frac{1}{2}}\left(2K^{\frac{2-d_{\mathcal{G}}}{2}}\right)}^{\infty}t\cdot(2te^{-t^2})dt.$$

Applying the integration by parts formula, we obtain

$$\begin{aligned}L= & \sqrt{\frac{2}{2-d_{\mathcal{G}}}}K^{\frac{2-d_{\mathcal{G}}}{2}}\left(\frac{\ln^{\frac{1}{2}}\left(2K^{\frac{2-d_{\mathcal{G}}}{2}}\right)}{2K^{\frac{2-d_{\mathcal{G}}}{2}}}+\int_{\ln^{\frac{1}{2}}\left(2K^{\frac{2-d_{\mathcal{G}}}{2}}\right)}^{\infty}e^{-t^2}dt\right)\\ \leqslant & \frac{1}{\sqrt{2(2-d_{\mathcal{G}})}}\left(\ln^{\frac{1}{2}}\left(2K^{\frac{2-d_{\mathcal{G}}}{2}}\right)+\frac{1}{2\ln^{\frac{1}{2}}\left(2K^{\frac{2-d_{\mathcal{G}}}{2}}\right)}\right).\end{aligned}$$

Consequently,

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant 4\sqrt{\frac{10\cdot 72^{d_{\mathcal{G}}}\cdot K_{\mathcal{G}}}{2(2-d_{\mathcal{G}})}}\cdot\frac{\sqrt{C}(\log_2(2C))^{1/2+d_{\mathcal{G}}/2}}{\sqrt{m}}\gamma^{1-\frac{d_{\mathcal{G}}}{2}}\left(1+2^{\frac{2}{2-d_{\mathcal{G}}}}\right)$$
$$\times\left(\ln^{\frac{1}{2}}\left(2K^{\frac{2-d_{\mathcal{G}}}{2}}\right)+\frac{1}{2\ln^{\frac{1}{2}}\left(2K^{\frac{2-d_{\mathcal{G}}}{2}}\right)}\right).$$

Second case: $d_{\mathcal{G}}\geqslant 2$ In this case, we apply (5) and (3) to (1) and use Hypothesis 1 to get

$$\begin{aligned}\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant & \gamma 2^{-\alpha(d_{\mathcal{G}})N}+2\sqrt{\frac{2C}{m}}\sum_{j=1}^{N}\left(\gamma 2^{-\alpha(d_{\mathcal{G}})j}+\gamma 2^{-\alpha(d_{\mathcal{G}})(j-1)}\right)\\ &\times\left[d\left(\frac{\gamma 2^{-\alpha(d_{\mathcal{G}})j}}{30\log_2(2C)}\right)\ln\left(\frac{30emM_{\mathcal{G}}\log_2(2C)}{\gamma 2^{-\alpha(d_{\mathcal{G}})j}}\right)\right]^{1/2}\\ \leqslant & \gamma 2^{-\alpha(d_{\mathcal{G}})N}+2\sqrt{\frac{2CK_{\mathcal{G}}}{m}}\left(30\log_2(2C)\right)^{d_{\mathcal{G}}/2}\gamma^{1-\frac{d_{\mathcal{G}}}{2}}\left(1+2^{\alpha(d_{\mathcal{G}})}\right)\\ &\times\sum_{j=1}^{N}2^{\alpha(d_{\mathcal{G}})\left(\frac{d_{\mathcal{G}}-2}{2}\right)j}\ln^{\frac{1}{2}}\left(\frac{30emM_{\mathcal{G}}\log_2(2C)\cdot 2^{\alpha(d_{\mathcal{G}})j}}{\gamma}\right).\end{aligned}\tag{6}$$

Unlike the first case, we now control the number of steps N in (6) through the parameters of interest, C and m. The aim is to optimize the dependencies with respect to them while making sure that (i) N is a strictly positive integer and (ii) as $m\to\infty$, $N\to\infty$.

Now, if $d_{\mathcal{G}}=2$, set $\alpha(d_{\mathcal{G}})=1$. Thus, from (6), we have

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right)\leqslant\gamma 2^{-N}+180\sqrt{\frac{2CK_{\mathcal{G}}}{m}}\log_2(2C)\sum_{j=1}^{N}\ln^{\frac{1}{2}}\left(\frac{30emM_{\mathcal{G}}\log_2(2C)\cdot 2^j}{\gamma}\right).$$

Setting $N=\left\lceil\log_2\left(\sqrt{\frac{m}{C}}\right)\right\rceil$ and bounding the series, we obtain

$$\begin{aligned}\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) &\leqslant \gamma\sqrt{\frac{C}{m}}+180\sqrt{\frac{2CK_{\mathcal{G}}}{m}}\log_2(2C)\sum_{j=1}^{N}\ln^{\frac{1}{2}}\left(\frac{30emM_{\mathcal{G}}\log_2(2C)\cdot 2^j}{\gamma}\right)\\ &< \gamma\sqrt{\frac{C}{m}}\\ &\quad+180\sqrt{\frac{2CK_{\mathcal{G}}}{m}}\log_2(2C)\left\lceil\log_2\left(\sqrt{\frac{m}{C}}\right)\right\rceil\ln^{\frac{1}{2}}\left(\frac{60em^{3/2}\log_2(2C)M_{\mathcal{G}}}{\gamma\sqrt{C}}\right).\end{aligned}$$

For the final case, $d_{\mathcal{G}}>2$, we set $\alpha(d_{\mathcal{G}})=\dfrac{2}{d_{\mathcal{G}}-2}$ in (6) and bound the geometric series:

$$\begin{aligned}\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant & \gamma 2^{-\frac{2}{d_{\mathcal{G}}-2}N}+2\sqrt{\frac{2CK_{\mathcal{G}}}{m}}\left(30\log_2(2C)\right)^{d_{\mathcal{G}}/2}\gamma^{1-\frac{d_{\mathcal{G}}}{2}}(1+2^{\frac{2}{d_{\mathcal{G}}-2}})\\ &\times\sum_{j=1}^{N}2^j\ln^{\frac{1}{2}}\left(\frac{30emM_{\mathcal{G}}\log_2(2C)\cdot 2^{\frac{2}{d_{\mathcal{G}}-2}j}}{\gamma}\right)\\ \leqslant & \gamma 2^{-\frac{2}{d_{\mathcal{G}}-2}N}+4\cdot 2^N\sqrt{\frac{2CK_{\mathcal{G}}}{m}}\left(30\log_2(2C)\right)^{d_{\mathcal{G}}/2}\gamma^{1-\frac{d_{\mathcal{G}}}{2}}(1+2^{\frac{2}{d_{\mathcal{G}}-2}})\\ &\times\ln^{\frac{1}{2}}\left(\frac{30emM_{\mathcal{G}}\log_2(2C)\cdot 2^{\frac{2}{d_{\mathcal{G}}-2}N}}{\gamma}\right).\end{aligned}\tag{7}$$

Now, let $N = \left\lceil \frac{d_{\mathcal{G}}-2}{2d_{\mathcal{G}}} \log_2 \left(\frac{m}{\log_2^{2d_{\mathcal{G}}}(2C)^{\frac{1}{d_{\mathcal{G}}}}} \right) \right\rceil$. Note that, with the assumption $m \geqslant C^{1.2}$, $m > \log_2^{2d_{\mathcal{G}}}(2C)^{\frac{1}{d_{\mathcal{G}}}}$ for all $d_{\mathcal{G}} > 2$ and thus, N is a strictly positive integer. Applying it to (7), we get

$$\begin{aligned}\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant & \frac{\gamma \log_2^2(2C)^{\frac{1}{d_{\mathcal{G}}}}}{m^{\frac{1}{d_{\mathcal{G}}}}} + 8\sqrt{2K_{\mathcal{G}}} \cdot 30^{d_{\mathcal{G}}/2} d_{\mathcal{G}}^{d_{\mathcal{G}}-2} \gamma^{1-\frac{d_{\mathcal{G}}}{2}} \left(1 + 2^{\frac{2}{d_{\mathcal{G}}-2}}\right) \\ & \times \frac{\sqrt{C}\left(\log_2(2C)\right)^{2-d_{\mathcal{G}}/2}}{m^{\frac{1}{d_{\mathcal{G}}}}} \ln^{\frac{1}{2}}\left(\frac{60 e d_{\mathcal{G}}^2 m^{1+\frac{1}{d_{\mathcal{G}}}} M_{\mathcal{G}}}{\gamma \log_2(2C)}\right).\end{aligned}$$

□

5 Conclusions

We derived a sharper risk bound for multi-category margin classifiers following the pathway of [9]. In this pathway, the first capacity measure that appears in the control term of the guaranteed risk is a Rademacher complexity. It is then related to the metric entropy through the chaining method. Using a decomposition for metric entropy, we transition from the multi-class setting to the bi-class one. Finally, a combinatorial bound gives an estimate on the metric entropy in terms of the combinatorial dimension. The metric entropy bound used in [9] is the L_2-norm one of [12], which in this paper we generalized to L_p-norms with integer $p > 2$. This generalization resulted in an improved dependency on the number C of categories compared to [9] without worsening the dependency on the sample size m nor the one on the margin parameter γ.

So far, to get an explicit dependency on C under minimal learnability assumptions, a transition from the multi-class case to the bi-class one has been been performed at the level of one of two capacity measures. Realizing it at the level of a Rademacher complexity, a linear dependency on C was obtained in [8]. In this paper, as in [9], we showed that postponing it to the level of metric entropy, this dependency can be improved to a sublinear one. The case that remains to be studied is a decomposition at the level of a combinatorial dimension, more precisely, at that of the fat-shattering dimension. The goal is to complete the picture of the impact that performing a decomposition at the level of one of three different capacity measures has on the dependencies on C, m and γ.

A Proof of Theorem 2

Let $\mathcal{T}_n = \{t_i : 1 \leqslant i \leqslant n\} \subset \mathcal{T}$ and $\boldsymbol{t}_n = (t_i)_{1 \leqslant i \leqslant n}$. Let $\mathcal{F}_\epsilon$ be an ϵ-separated with respect to the pseudo-metric $d_{p,\boldsymbol{t}_n}$ subset of $\mathcal{F}$ of maximal cardinality. By definition, $|\mathcal{F}_\epsilon| = \mathcal{M}\left(\epsilon, \mathcal{F}, d_{p,\boldsymbol{t}_n}\right) = \mathcal{M}\left(\epsilon, \mathcal{F}_\epsilon|_{\mathcal{T}_n}, d_{p,\boldsymbol{t}_n}\right) = |\mathcal{F}_\epsilon|_{\mathcal{T}_n}|$, where $\mathcal{F}_\epsilon|_{\mathcal{T}_n}$ denotes the class $\mathcal{F}_\epsilon$ whose domain is restricted to $\mathcal{T}_n$. We distinguish three major steps in the proof: i) discretize functions in the set $\mathcal{F}_\epsilon|_{\mathcal{T}_n}$, ii) demonstrate that the set of discretized functions is separated, and iii) upper bound the cardinality of the discretized set. The purpose of discretizing the set of real-valued functions is to reduce the

original problem into the one that can be addressed by combinatorial means: we upper bound the packing number of the discretized set which is then related to that of the original set via the step (ii).

(a) Let $\epsilon' = 4(4K_p)^{1/p}$, $\eta = \dfrac{\epsilon}{\epsilon'+2}$ and $N = \lfloor 2M_{\mathcal{F}}/\eta \rfloor$. Define the class $\tilde{\mathcal{F}}^{\eta}$ of functions from $\mathcal{T}_n$ into $[\![0, N]\!]$ obtained by the discretization of functions in $\mathcal{F}_\epsilon$ in the following way:

$$\tilde{\mathcal{F}}^{\eta} = \left\{ \tilde{f} : \tilde{f}(t_i) = \left\lfloor \frac{f(t_i) + M_{\mathcal{F}}}{\eta} \right\rfloor, i \in [\![1, n]\!],\ f \in \mathcal{F}_\epsilon|_{\mathcal{T}_n} \right\}.$$

We claim that with such a discretization, for any $\tilde{f}_1, \tilde{f}_2 \in \tilde{\mathcal{F}}^{\eta}$, $d_{p,\boldsymbol{t}_n}\left(\tilde{f}_1, \tilde{f}_2\right) \geqslant \epsilon'$. Using $\left|\lfloor a \rfloor - \lfloor b \rfloor\right|^p \geqslant (\max(0, |a-b|-1))^p$ for all $a, b \in \mathbb{R}_+$,

$$\begin{aligned} d_{p,\boldsymbol{t}_n}\left(\tilde{f}_1, \tilde{f}_2\right) &= \left(\frac{1}{n} \sum_{i=1}^{n} \left| \left\lfloor \frac{f_1(t_i) + M_{\mathcal{F}}}{\eta} \right\rfloor - \left\lfloor \frac{f_2(t_i) + M_{\mathcal{F}}}{\eta} \right\rfloor \right|^p \right)^{\frac{1}{p}} \\ &\geqslant \left(\frac{1}{n} \sum_{i \in I} \left(\frac{1}{\eta} \left| f_1(t_i) - f_2(t_i) \right| - 1 \right)^p \right)^{\frac{1}{p}}, \end{aligned}$$

where I denotes the set of indices such that $\dfrac{1}{\eta} \left| f_1(t_i) - f_2(t_i) \right| \geqslant 1$, for all $i \in I$. Next, by the inverse triangle inequality, $d_{p,\boldsymbol{t}_n}(f_1, f_2) \geqslant d_{p,\boldsymbol{t}_n}(f_1, 0) - d_{p,\boldsymbol{t}_n}(f_2, 0)$ for all $f_1, f_2 \in \mathcal{F}$, the right-hand side of the above inequality can be bounded as

$$\begin{aligned} d_{p,\boldsymbol{t}_n}\left(\tilde{f}_1, \tilde{f}_2\right) &\geqslant \frac{1}{\eta} \left(\frac{1}{n} \sum_{i \in I} \left| f_1(t_i) - f_2(t_i) \right|^p \right)^{\frac{1}{p}} - \left(\frac{|I|}{n} \right)^{\frac{1}{p}} \\ &\geqslant \frac{1}{\eta} \left(\frac{1}{n} \sum_{i \in I} \left| f_1(t_i) - f_2(t_i) \right|^p \right)^{\frac{1}{p}} - 1. \end{aligned} \tag{8}$$

Let I^c denote the complement of I. Now, by definition of $\mathcal{F}_\epsilon$,

$$\frac{1}{n} \sum_{i \in I} \left| f_1(t_i) - f_2(t_i) \right|^p + \frac{1}{n} \sum_{i \in I^c} \left| f_1(t_i) - f_2(t_i) \right|^p \geqslant \epsilon^p.$$

It follows that

$$\begin{aligned} \epsilon^p &\leqslant \frac{1}{n} \sum_{i \in I} \left| f_1(t_i) - f_2(t_i) \right|^p + \frac{|I^c| \eta^p}{n} \leqslant \frac{1}{n} \sum_{i \in I} \left| f_1(t_i) - f_2(t_i) \right|^p + \eta^p \\ \implies (\epsilon^p - \eta^p)^{1/p} &\leqslant \left(\frac{1}{n} \sum_{i \in I} \left| f_1(t_i) - f_2(t_i) \right|^p \right)^{1/p}. \end{aligned}$$

Applying the last inequality to (8) and using $((a-b)+b) \leqslant ((a-b)^{1/p} + b^{1/p})^p$ with $a, b \in \mathbb{R}_+$ and $a \geqslant b$ (where we set $a = (\epsilon'+2)^p$ and $b = 1$), we get

$$d_{p,\boldsymbol{t}_n}\left(\tilde{f}_1, \tilde{f}_2\right) \geqslant \frac{1}{\eta} (\epsilon^p - \eta^p)^{1/p} - 1 = \left((\epsilon'+2)^p - 1\right)^{1/p} - 1 \geqslant \epsilon'.$$

This proves our claim. Then, it follows that

$$\mathcal{M}\left(\epsilon, \mathcal{F}_\epsilon, d_{p,\boldsymbol{t}_n}\right) \leqslant \mathcal{M}(\epsilon', \tilde{\mathcal{F}}^{\eta}, d_{p,\boldsymbol{t}_n}) = |\tilde{\mathcal{F}}^{\eta}|. \tag{9}$$

The major step that remains to perform to arrive at the claimed bound is to upper bound the right-hand side of (9). To this end, we appeal to Proposition 3. Let d_s be the strong dimension of $\tilde{\mathcal{F}}^{\eta}$. By part (1) of Lemma 3.2 in [13],

$$d_s \leqslant \left(\frac{\eta}{2}\right)\text{-dim}(\mathcal{F}_{\epsilon}|_{\mathcal{T}_n}) = \left(\frac{\epsilon}{8(4K_p)^{1/p}+4}\right)\text{-dim}(\mathcal{F}_{\epsilon}|_{\mathcal{T}_n}).$$

By Lemma 1 and the fact that $p \geqslant 3$, on the other hand, we have

$$8(4K_p)^{1/p} + 4 < 8 \cdot 4^{1/p}p + 4 < 15p.$$

We can plug this result in the upper bound on d_s based on the fact that the fat-shattering dimension decreases with the scale:

$$\begin{aligned} d_s &\leqslant \left(\frac{\epsilon}{15p}\right)\text{-dim}(\mathcal{F}_{\epsilon}|_{\mathcal{T}_n}) \\ &\leqslant \left(\frac{\epsilon}{15p}\right)\text{-dim}(\mathcal{F}) = d\left(\frac{\epsilon}{15p}\right). \end{aligned}$$

Now, according to Proposition 3,

$$\begin{aligned} |\tilde{\mathcal{F}}^{\eta}| &\leqslant \left(\frac{eNn}{d\left(\frac{\epsilon}{15p}\right)}\right)^{2d\left(\frac{\epsilon}{15p}\right)} \\ &\leqslant \left(\frac{en}{d\left(\frac{\epsilon}{15p}\right)}\left\lfloor\frac{2M_{\mathcal{F}}}{\eta}\right\rfloor\right)^{2d\left(\frac{\epsilon}{15p}\right)} \\ &\leqslant \left(\frac{en}{d\left(\frac{\epsilon}{15p}\right)}\left(\frac{8M_{\mathcal{F}}(4K_p)^{1/p}+4M_{\mathcal{F}}}{\epsilon}\right)\right)^{2d\left(\frac{\epsilon}{15p}\right)}. \end{aligned} \tag{10}$$

Applying Lemma 1 to the right-hand side of (10) and simplifying it we get

$$|\tilde{\mathcal{F}}^{\eta}| \leqslant \left(\frac{15enM_{\mathcal{F}}p}{\epsilon d\left(\frac{\epsilon}{15p}\right)}\right)^{2d\left(\frac{\epsilon}{15p}\right)}. \tag{11}$$

We apply the relation (9) and the following well-known inequality [23]

$$\mathcal{N}(\epsilon, \mathcal{F}, d_{p,\boldsymbol{t}_n}) \leqslant \mathcal{M}(\epsilon, \mathcal{F}, d_{p,\boldsymbol{t}_n}) \tag{12}$$

in sequence to the left-hand side of (11). Finally, to obtain the claimed result, we take supremum over $\boldsymbol{t}_n \in \mathcal{T}^n$ of both sides of the obtained bound.

(b) To derive a dimension-free combinatorial bound we use the L_p-norm generalization of probabilistic extraction principle: Lemma 8 of [9]. According to this lemma, there exists a subset $\mathcal{T}_q = \{t_{i_k} : 1 \leqslant k \leqslant q\}$ of $\mathcal{T}_n$ of cardinality

$$q \leqslant \frac{112\,(2M_{\mathcal{F}})^{2p}\ln(|\mathcal{F}_{\epsilon}|)}{3\epsilon^{2p}}, \tag{13}$$

such that $\mathcal{F}_\epsilon$ is $\epsilon_1 = \epsilon/2^{\frac{p+1}{p}}$-separated with respect to $d_{p,\boldsymbol{t}_q}$, with $\boldsymbol{t}_q = (t_{i_k})_{1\leqslant k\leqslant q}$. Let $\mathcal{F}_\epsilon|_{\mathcal{T}_q}$ denote the class $\mathcal{F}_\epsilon$ whose domain is restricted to $\mathcal{T}_q$. We have

$$|\mathcal{F}_\epsilon| = \mathcal{M}\left(\epsilon_1, \mathcal{F}_\epsilon, d_{p,\boldsymbol{t}_q}\right) = \mathcal{M}\left(\epsilon_1, \mathcal{F}_\epsilon|_{\mathcal{T}_q}, d_{p,\boldsymbol{t}_q}\right) = |\mathcal{F}_\epsilon|_{\mathcal{T}_q}|. \tag{14}$$

We let $\eta = \dfrac{\epsilon_1}{\epsilon' + 2}$ and discretize the functions in the set $\mathcal{F}_\epsilon|_{\mathcal{T}_q}$ in a similar way as in part (a):

$$\tilde{\mathcal{F}}^\eta = \left\{\tilde{f} : \tilde{f}(t_{i_k}) = \left\lfloor \frac{f(t_{i_k}) + M_{\mathcal{F}}}{\eta} \right\rfloor, k \in [\![1, q]\!],\ f \in \mathcal{F}_\epsilon|_{\mathcal{T}_q}\right\}.$$

Applying the same procedure as in the proof of part (a), we obtain that for any $\tilde{f}_1, \tilde{f}_2 \in \tilde{\mathcal{F}}^\eta$, $d_{p,\boldsymbol{t}_q}\left(\tilde{f}_1, \tilde{f}_2\right) \geqslant \epsilon'$, and hence

$$\mathcal{M}\left(\epsilon_1, \mathcal{F}_\epsilon, d_{p,\boldsymbol{t}_q}\right) \leqslant \mathcal{M}(\epsilon', \tilde{\mathcal{F}}^\eta, d_{p,\boldsymbol{t}_q}) = |\tilde{\mathcal{F}}^\eta|. \tag{15}$$

By Proposition 3,

$$|\tilde{\mathcal{F}}^\eta| \leqslant \left(\frac{eNq}{d_s}\right)^{2d_s},$$

where d_s is the strong dimension of $\tilde{\mathcal{F}}^\eta$. Plugging the value of N and performing similar computations as in Inequalties (10)-(11) of part (a), we get

$$|\tilde{\mathcal{F}}^\eta| \leqslant \left(\frac{23eqM_{\mathcal{F}}(4K_p)^{1/p}}{\epsilon d_s}\right)^{2d_s}. \tag{16}$$

Now, we go back from the discretized set $\tilde{\mathcal{F}}^\eta$ to $\mathcal{F}_\epsilon$ using the relations (14) and (15) which yield: $|\mathcal{F}_\epsilon| \leqslant |\tilde{\mathcal{F}}^\eta|$. Using it and Inequality (13) in (16) give:

$$\ln\left(|\mathcal{F}_\epsilon|\right) \leqslant 2d_s \ln\left(\frac{2576 \cdot 2^{2p} e M_{\mathcal{F}}^{2p+1}(4K_p)^{1/p} \ln\left(|\mathcal{F}_\epsilon|\right)}{3\epsilon^{2p+1} d_s}\right).$$

Now, based on $\ln(u) < \sqrt{u}$ and by a straightforward computation,

$$\ln\left(|\mathcal{F}_\epsilon|\right) \leqslant 4d_s \ln\left(\frac{2576 \cdot 2^{2p+1} e M_{\mathcal{F}}^{2p+1}(4K_p)^{1/p}}{3\epsilon^{2p+1}}\right). \tag{17}$$

Next, we bound d_s using part (1) of Lemma 3.2 in [13] and Lemma 1:

$$\begin{aligned}
d_s &\leqslant \left(\frac{\eta}{2}\right)\text{-dim}(\mathcal{F}_\epsilon|_{\mathcal{T}_q}) \\
&= \left(\frac{\epsilon}{2^{\frac{4p+1}{p}}(4K_p)^{1/p} + 2^{\frac{3p+1}{p}}}\right)\text{-dim}(\mathcal{F}_\epsilon|_{\mathcal{T}_q}) \\
&\leqslant \left(\frac{\epsilon}{16 \cdot 2^{\frac{3}{p}} p + 8 \cdot 2^{\frac{1}{p}}}\right)\text{-dim}(\mathcal{F}_\epsilon|_{\mathcal{T}_q}) \\
&\leqslant \left(\frac{\epsilon}{36p}\right)\text{-dim}(\mathcal{F}_\epsilon|_{\mathcal{T}_q}).
\end{aligned}$$

Plugging this into (17) and applying Lemma 1 to K_p, we obtain

$$\begin{aligned}
\ln\left(|\mathcal{F}_\epsilon|\right) &\leqslant 4d\left(\frac{\epsilon}{36p}\right)\ln\left(\frac{2576 \cdot 2^{2p+1} e M_{\mathcal{F}}^{2p+1} 4^{1/p} p}{3\epsilon^{2p+1}}\right) \\
&\leqslant 10p\, d\left(\frac{\epsilon}{36p}\right)\ln\left(\frac{7p^{\frac{1}{7}} M_{\mathcal{F}}}{\epsilon}\right).
\end{aligned}$$

The claim follows from the application of $|\mathcal{F}_\epsilon| = \mathcal{M}(\epsilon, \mathcal{F}, d_{p,\boldsymbol{t}_n})$, Inequality (12) and taking supremum over $\boldsymbol{t}_n \in \mathcal{T}^n$ of both sides of the obtained bound.

B Technical Results

Lemma 1. *For all $p \in \mathbb{N}^* \setminus \{1, 2\}$,*

$$\sum_{k=1}^{\infty} \frac{k^p}{2^k} < p^p.$$

Proof. By Formula (8.5) in [24, page 119],

$$\sum_{k=1}^{\infty} \frac{k^p}{u^k} = \frac{u\psi_p(-u)}{(u-1)^{(p+1)}},$$

where $\psi_p(u) = \sum_{j=0}^{p-1} (-1)^j \binom{p}{j+1} (u+1)^j \psi_{(p-1)-j}(u)$ is an Eulerian polynomial in u of degree $p-1$ with $\psi_0(u) = \psi_1(u) = 1$ (see page 116 in [24] for explicit form of this polynomial for smaller values of p). Thus for $u = 2$,

$$\sum_{k=1}^{\infty} \frac{k^p}{2^k} = 2\psi_p(-2).$$

We now show by induction that for all $p > 2$, $\psi_p(-2) < \frac{p^p}{2}$. By definition,

$$\psi_p(-2) = \sum_{j=0}^{p-1} \binom{p}{j+1} \psi_{(p-1)-j}(-2).$$

For the base case, $p = 3$, it is easily seen that $\psi_3(-2) < 3^3/2$. Now, assume for $k > 3$, $\psi_k(-2) < k^k/2$. Then,

$$\begin{aligned}
\psi_{k+1}(-2) &= \sum_{j=0}^{k} \binom{k+1}{j+1} \psi_{k-j}(-2) \\
&= (k+1)\psi_k(-2) + \sum_{j=1}^{k} \binom{k+1}{j+1} \psi_{k-j}(-2) \\
&< (k+1)k^k/2 + \sum_{j=0}^{k-1} \binom{k+1}{j+2} \psi_{(k-1)-j}(-2) \\
&= (k+1)k^k/2 + \sum_{j=0}^{k-1} \left(\binom{k}{j+1} + \binom{k}{j+2} \right) \psi_{(k-1)-j}(-2). \qquad (18)
\end{aligned}$$

We have that

$$\begin{aligned}
\binom{k}{j+2} &= \frac{k!}{(j+2)!(k-(j+2))!} \\
&= \frac{k!}{(j+1)!(k-(j+2))!} \cdot \frac{k-(j+1)}{(k-(j+1))(j+2)} \\
&= \frac{k!}{(j+1)!(k-(j+1))!} \cdot \frac{k-(j+1)}{j+2} \\
&< k \binom{k}{j+1}.
\end{aligned}$$

Applying it in (18), we obtain

$$\begin{aligned}\psi_{k+1}(-2) &< (k+1)k^k/2 + \sum_{j=0}^{k-1}(k+1)\binom{k}{j+1}\psi_{(k-1)-j}(-2)\\ &< (k+1)k^k/2 + (k+1)\psi_k(-2)\\ &< (k+1)k^k.\end{aligned}$$

Now, by the binomial theorem, for all $k > 1$,

$$\begin{aligned}(k+1)^k &= \binom{k}{0}k^0 + \cdots + \binom{k}{k-1}k^{k-1} + \binom{k}{k}k^k\\ &= 1 + \cdots + k\cdot k^{k-1} + k^k\\ &> 2k^k.\end{aligned}$$

Consequently,

$$\psi_{k+1}(-2) < (k+1)\cdot(k+1)^k/2 = (k+1)^{k+1}/2,$$

where we used the convention that $\forall k > n$, $\binom{n}{k} = 0$. □

The results demonstrated hereafter are the generalizations of those in [12]. In the following, we denote $K_p = \sum_{k=1}^{\infty}\frac{k^p}{2^k}$ with $p \in \mathbb{N}^* \setminus \{1,2\}$.

Lemma 2 (After Lemma 5 of [12])**.** *Let X be a bounded random variable. Let $M_p(X) = (\mathbb{E}|X|^p])^{1/p}$. Then, there exist numbers $a \in \mathbb{R}$ and $\beta \in (0, 1/2]$, such that*

$$\mathbb{P}\left\{X > a + \frac{M_p(X)}{4(2K_p)^{1/p}}\right\} \geqslant \frac{\beta}{2} \text{ and } \mathbb{P}\left\{X < a - \frac{M_p(X)}{4(2K_p)^{1/p}}\right\} \geqslant 1-\beta,$$

or vice versa.

Proof. The proof closely follows that of Lemma 5 of [12] where the variance of X is replaced by its higher moments.

Divide $\mathbb{R}_+$ into the intervals I_k of length $cM_p(X)$ with

$$\frac{1}{2(2K_p)^{1/p}} < c < \frac{1}{(2K_p)^{1/p}}$$

by setting

$$I_k = (cM_p(X)k, cM_p(X)(k+1)], \quad k \geqslant 0.$$

Assume the lemma does not hold and let $(\beta_i)_{i\geqslant 0}$ be a non-increasing sequence of non-negative numbers such that

$$\mathbb{P}\{X > 0\} = \beta_0 \leqslant 1/2$$

and

$$\mathbb{P}\{X \in I_k\} = \beta_k - \beta_{k+1}, \quad k \geqslant 0.$$

For the conclusion of the lemma to fail it should hold that

$$\forall k \geqslant 0, \quad \beta_{k+1} \leqslant \beta_k/2. \tag{19}$$

Now, assume that for some k, $\beta_{k+1} > \beta_k/2$ and consider intervals

$$J_1 = (-\infty, 0] \cup (0, cM_p(X)k] = (-\infty, 0] \cup \left(\bigcup_{0 \leqslant j \leqslant k-1} I_j \right)$$

and $J_2 = (cM_p(X)(k+1), \infty)$. Then,

$$\mathbb{P}\{X \in J_1\} = (1 - \beta_0) + \sum_{0 \leqslant j \leqslant k-1} (\beta_j - \beta_{j+1}) = 1 - \beta_k$$

and

$$\mathbb{P}\{X \in J_2\} = \sum_{j \geqslant k+1} (\beta_j - \beta_{j+1}) = \beta_{k+1}.$$

By definition of $(\beta_i)_{i \geqslant 0}$ and by our assumption, $1/2 \geqslant \beta_0 \geqslant \beta_k \geqslant \beta_{k+1} > \beta_k/2 \geqslant 0$, which means that $\beta_k \in (0, 1/2]$. Now, let a be the middle point between the intervals J_1 and J_2 and let $\beta = \beta_k$. We have that

$$cM_p(X)k = a - \frac{cM_p(X)}{2} < a - \frac{M_p(X)}{4(2K_p)^{1/p}} \implies 1 - \beta \leqslant \mathbb{P}\left\{ X < a - \frac{M_p(X)}{4(2K_p)^{1/p}} \right\}$$

and

$$cM_p(X)(k+1) = a + \frac{cM_p(X)}{2} > a + \frac{M_p(X)}{4(2K_p)^{1/p}} \implies \frac{\beta}{2} \leqslant \mathbb{P}\left\{ X > a + \frac{M_p(X)}{4(2K_p)^{1/p}} \right\}.$$

Thus, the lemma holds. This proves (19). Now, by induction from (19) we get that

$$\beta_k \leqslant 1/2^{k+1}.$$

We use it in the computation of $M_p^p(X)$. By definition,

$$M_p^p(X) = \int_0^\infty \mathbb{P}\{|X| > t\} dt^p = \int_0^\infty \mathbb{P}\{X > t\} dt^p + \int_0^\infty \mathbb{P}\{X < -t\} dt^p.$$

By construction, whenever $t \in I_k$, $\mathbb{P}\{X > t\} \leqslant \mathbb{P}\{X > cM_p(X)k\} = \mathbb{P}\{X \in \bigcup_{l \geqslant k} I_l\} = \sum_{l \geqslant k} (\beta_l - \beta_{l+1}) = \beta_k$. Thus,

$$\begin{aligned}
\int_0^\infty \mathbb{P}\{X > t\} dt^p &\leqslant \sum_{k \geqslant 0} \int_{I_k} \beta_k p t^{p-1} dt \\
&\leqslant (cM_p(X))^p \sum_{k \geqslant 0} \frac{(k+1)^p - k^p}{2^{k+1}} \\
&\leqslant (cM_p(X))^p \sum_{k \geqslant 1} \frac{k^p}{2^k} \\
&= (cM_p(X))^p K_p \\
&< M_p^p(X)/2.
\end{aligned}$$

By a similar procedure, it can be proved that

$$\int_0^\infty \mathbb{P}\{X < -t\} dt^p < M_p^p(X)/2.$$

This produces a contradiction $M_p^p(X) < M_p^p(X)/2 + M_p^p(X)/2 = M_p^p(X)$ proving the lemma. □

In the following, $\mathcal{T} = \{t_i : 1 \leqslant i \leqslant n\}$ is a finite set and $\mathbf{t}_n = (t_i)_{1 \leqslant i \leqslant n}$.

Lemma 3 (After Lemma 6 of [12]). *Let $\mathcal{F}$ be a finite class of functions from $\mathcal{T}$ into $[0, M_{\mathcal{F}}]$ with $M_{\mathcal{F}} \in \mathbb{R}_+$ and $|\mathcal{F}| > 1$. Assume that for some $\epsilon \in (0, M_{\mathcal{F}}]$, $\mathcal{F}$ is ϵ-separated in the pseudo-metric $d_{p,\mathbf{t}_n}$. Then there exist $i \in [\![1, n]\!]$, $a \in \mathbb{R}$ and $\beta \in (0, 1/2]$ such that*

$$\left| \left\{ f \in \mathcal{F} : f(t_i) > a + \frac{\epsilon}{8(4K_p)^{1/p}} \right\} \right| \geqslant p_1 |\mathcal{F}|$$
$$\left| \left\{ f \in \mathcal{F} : f(t_i) < a - \frac{\epsilon}{8(4K_p)^{1/p}} \right\} \right| \geqslant p_2 |\mathcal{F}|,$$

with $p_1 \geqslant \frac{\beta}{2}$ and $p_2 \geqslant 1 - \beta$ or vice versa.

Proof. $\mathcal{F}$ can be viewed as a finite probability space $(\mathcal{F}, \mathcal{A}, P_{\mathcal{F}})$ with a uniform probability measure $P_{\mathcal{F}}(A) = |A|/|\mathcal{F}|$ for any $A \in \mathcal{A}$. Then, for any two random elements $f, f' \in \mathcal{F}$ selected independently according to $P_{\mathcal{F}}$,

$$\begin{aligned}\mathbb{E}_{f,f' \sim P_{\mathcal{F}}} \left(d_{p,\mathbf{t}_n}(f, f')\right)^p &= \mathbb{E}_{f,f' \sim P_{\mathcal{F}}} \left[\frac{1}{n} \sum_{i=1}^n |f(t_i) - f'(t_i)|^p \right] \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{f,f' \sim P_{\mathcal{F}}} |f(t_i) - f'(t_i)|^p .\end{aligned}$$

By the Minkowski inequality, for any $i \in [\![1, n]\!]$,

$$\begin{aligned}\mathbb{E}_{f,f' \sim P_{\mathcal{F}}} |f(t_i) - f'(t_i)|^p &\leqslant \left(\left(\mathbb{E}_{f \sim P_{\mathcal{F}}} |f(t_i)|^p\right)^{1/p} + \left(\mathbb{E}_{f' \sim P_{\mathcal{F}}} |-f'(t_i)|^p\right)^{1/p} \right)^p \\ &= \left(\left(\mathbb{E}_{f \sim P_{\mathcal{F}}} |f(t_i)|^p\right)^{1/p} + \left(\mathbb{E}_{f' \sim P_{\mathcal{F}}} |f'(t_i)|^p\right)^{1/p} \right)^p \\ &= 2^p \mathbb{E}_{f \sim P_{\mathcal{F}}} |f(t_i)|^p .\end{aligned}$$

Taking it into account in the formula above, we obtain,

$$\mathbb{E}_{f,f' \sim P_{\mathcal{F}}} \left(d_{p,\mathbf{t}_n}(f, f')\right)^p \leqslant \frac{2^p}{n} \sum_{i=1}^n \mathbb{E}_{f \sim P_{\mathcal{F}}} |f(t_i)|^p .$$

Now, the event that the realizations of f and f' are different elements in $\mathcal{F}$ happens with probability $1 - 1/|\mathcal{F}|$. Then, by the separation assumption on $\mathcal{F}$ we have

$$\mathbb{E}_{f,f' \sim P_{\mathcal{F}}} \left(d_{p,\mathbf{t}_n}(f, f')\right)^p \geqslant (1 - 1/|\mathcal{F}|)\, \epsilon^p \geqslant (1 - 1/2)\, \epsilon^p = \epsilon^p/2.$$

Thus,

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}_{f \sim P_{\mathcal{F}}} |f(t_i)|^p \geqslant \frac{\epsilon^p}{2^{p+1}}.$$

It means that there exists $i \in [\![1, n]\!]$, such that

$$\mathbb{E}_{f \sim P_{\mathcal{F}}} |f(t_i)|^p \geqslant \frac{\epsilon^p}{2^{p+1}}.$$

Next, we apply Lemma 2 to the random element f and take into account that

$$M_p(f(t_i)) \geqslant \frac{\epsilon}{2^{1+1/p}}$$

and that

$$\frac{M_p(f(t_i))}{4(2K_p)^{1/p}} \geqslant \frac{\epsilon}{8 \times 2^{1/p}(2K_p)^{1/p}} = \frac{\epsilon}{8(4K_p)^{1/p}}.$$

Then, it follows that

$$\frac{\beta}{2} \leqslant P_{\mathcal{F}}\left\{f(t_i) > a + \frac{M_p(f(t_i))}{4(2K_p)^{1/p}}\right\} \leqslant P_{\mathcal{F}}\left\{f(t_i) > a + \frac{\epsilon}{8(4K_p)^{1/p}}\right\}$$

and, similarly,

$$1 - \beta \leqslant P_{\mathcal{F}}\left\{f(t_i) < a - \frac{M_p(f(t_i))}{4(2K_p)^{1/p}}\right\} \leqslant P_{\mathcal{F}}\left\{f(t_i) < a - \frac{\epsilon}{8(4K_p)^{1/p}}\right\}.$$

Finally, the claim follows from the definition of $P_{\mathcal{F}}$. □

The results given in the sequel call for the introduction of the definition of the ϵ-separating tree.

Definition 10. *Let $\mathcal{F}$ be a class of functions on $\mathcal{T}$. A tree $T(\mathcal{F})$ is a finite collection of subsets of $\mathcal{F}$, such that its any two elements are either disjoint or one of them contains the other. A son of $\bar{\mathcal{F}} \in T(\mathcal{F})$ is its maximal (with respect to inclusion) proper subset. An element of $T(\mathcal{F})$ with no sons is called a leaf. Let $\epsilon > 0$. If every $\bar{\mathcal{F}} \in T(\mathcal{F})$ which is not a leaf has exactly two sons $\bar{\mathcal{F}}_+, \bar{\mathcal{F}}_-$ and*

$$\exists i \in [\![1, n]\!],\ \forall (f_+, f_-) \in \left(\bar{\mathcal{F}}_+, \bar{\mathcal{F}}_-\right), \quad f_+(t_i) > f_-(t_i) + \epsilon,$$

then $T(\mathcal{F})$ is an ϵ-separating tree.

Proposition 1 (After Proposition 8 in [12])**.** *Let $\mathcal{F}$ be a finite class of functions from $\mathcal{T}$ into $[0, M_{\mathcal{F}}]$ with $M_{\mathcal{F}} \in \mathbb{R}_+$. Assume that for some $\epsilon \in (0, M_{\mathcal{F}}]$, $\mathcal{F}$ is ϵ-separated in the pseudo-metric $d_{p,\mathbf{t}_n}$. Then, there is a $\epsilon/4(4K_p)^{1/p}$-separating tree of $\mathcal{F}$ with at least $|\mathcal{F}|^{1/2}$ leaves.*

Proof. By Lemma 3, $\mathcal{F}$ has two subsets $\mathcal{F}_+$ and $\mathcal{F}_-$ such that

$$\exists i \in [\![1, n]\!],\ \exists a \in \mathbb{R},\ \forall (f_+, f_-) \in \mathcal{F}_+ \times \mathcal{F}_-, \quad \begin{cases} f_+(t_i) > a + \epsilon/8(4K_p)^{1/p} \\ f_-(t_i) < a - \epsilon/8(4K_p)^{1/p}, \end{cases}$$

which implies

$$f_+(t_i) < f_-(t_i) + \epsilon/4(4K_p)^{1/p}.$$

The rest of the proof is based on induction on the cardinality of $\mathcal{F}$ and is exactly as in [12], except that the tree is now $\epsilon/4(4K_p)^{1/p}$-separated. □

Proposition 2 (After Proposition 10 in [12]). *Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ into a finite set B of integers. Let $S \subseteq \mathcal{T}$ and let $v : S \to B$. The number of pairs (S, v) strongly shattered by $\mathcal{F}$ is at least the number of leaves in any 1-separating tree of $\mathcal{F}$.*

Proof. The proof follows exactly the one of Proposition 10 in [12], with a few minor technical changes. Let $\bar{\mathcal{F}}$ be a node in a 1-separating tree of $\mathcal{F}$. Let $N(A)$ denote the number of pairs strongly shattered by a set A. For the proof it suffices to show that if $\bar{\mathcal{F}}_+$ and $\bar{\mathcal{F}}_-$ are two sons of $\bar{\mathcal{F}}$, then

$$N(\bar{\mathcal{F}}) \geqslant N(\bar{\mathcal{F}}_+) + N(\bar{\mathcal{F}}_-). \tag{20}$$

By definition of the 1-separating tree, there exists $i_0 \in [\![1, n]\!]$ such that

$$\forall (f_+, f_-) \in \left(\bar{\mathcal{F}}_+, \bar{\mathcal{F}}_-\right), \quad f_+(t_{i_0}) > f_-(t_{i_0}) + 1.$$

It follows that

$$\exists b \in B, \forall (f_+, f_-) \in \left(\bar{\mathcal{F}}_+, \bar{\mathcal{F}}_-\right), \quad \begin{cases} f_+(t_{i_0}) > b \\ f_-(t_{i_0}) < b. \end{cases} \tag{21}$$

If a pair is strongly shattered either by $\bar{\mathcal{F}}_+$ or $\bar{\mathcal{F}}_-$, then it is also strongly shattered by $\bar{\mathcal{F}}$. On the other hand, if a pair (S, v) is strongly shattered both by $\bar{\mathcal{F}}_+$ and $\bar{\mathcal{F}}_-$, then $t_{i_0} \notin S$. Otherwise, there would exist $(f'_+, f'_-) \in \left(\bar{\mathcal{F}}_+, \bar{\mathcal{F}}_-\right)$ satisfying $f'_+(t_{i_0}) \leqslant v(t_{i_0}) - 1$ and $f'_-(t_{i_0}) \geqslant v(t_{i_0}) + 1$. Combining it with (21) yields a contradiction:

$$b + 1 < v(t_{i_0}) < b - 1.$$

Now, consider a pair $(S \cup \{t_{i_0}\}, v')$, where $v'(t_i) = v(t_i)$ for all $t_i \in S$ and $v'(t_{i_0}) = b$. This pair is shattered by $\bar{\mathcal{F}}$, but neither by $\bar{\mathcal{F}}_+$ or $\bar{\mathcal{F}}_-$. As S is shattered both by $\bar{\mathcal{F}}_+$ and $\bar{\mathcal{F}}_-$, then from (21) it follows that,

$$\forall (s_i)_{1 \leqslant i \leqslant n} \in \{-1, 1\}^n, \exists f_+ \in \bar{\mathcal{F}}_+, \quad \begin{cases} \forall i \in [\![1, n]\!], s_i \left(f_+(t_i) - v(t_i)\right) \geqslant 1, \\ f_+(t_{i_0}) \geqslant b + 1, \end{cases}$$

similarly,

$$\forall (s_i)_{1 \leqslant i \leqslant n} \in \{-1, 1\}^n, \exists f_- \in \bar{\mathcal{F}}_-, \quad \begin{cases} \forall i \in [\![1, n]\!], s_i \left(f_-(t_i) - v(t_i)\right) \geqslant 1, \\ f_-(t_{i_0}) \leqslant b - 1. \end{cases}$$

It proves the claim that $\bar{\mathcal{F}}$ shatters the pair $(S \cup \{t_{i_0}\}, v')$. Therefore, in both cases we get (20). □

The next result is obtained by combining Propositions 1 and 2.

Corollary 2 (After Corollary 11 in [12]). *Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ into a finite set B of integers. Let $S \subseteq \mathcal{T}$ and let $v : S \to B$. If $\mathcal{F}$ is $4(4K_p)^{1/p}$-separated in the pseudo-metric $d_{p,\mathbf{t}_n}$, then it strongly shatters at least $|\mathcal{F}|^{1/2}$ pairs (S, v) .*

Proposition 3 (After Proposition 12 in [12]). *Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ into $[\![0, b]\!]$. Let $d_s = S\text{-}dim(\mathcal{F})$. Assume $\mathcal{F}$ is $4(4K_p)^{1/p}$-separated in the pseudo-metric $d_{p,\mathbf{t}_n}$. Then for any $d \geqslant d_s$,*

$$|\mathcal{F}| \leqslant \left(\frac{ebn}{d}\right)^{2d}.$$

Proof. By Corollary 2, $\mathcal{F}$ strongly shatters at least $|\mathcal{F}|^{1/2}$ pairs (S, v). On the other hand, the total number of such pairs for which the cardinality of S is at most d_s is bounded above by

$$\sum_{k=0}^{d_s} \binom{n}{k} b^k.$$

To see this, note that there are at most $\binom{n}{k}$ number of sets S of size k and for each such S the number of functions h is bounded above by b^k. Therefore,

$$|\mathcal{F}|^{1/2} \leqslant \sum_{k=0}^{d_s} \binom{n}{k} b^k.$$

The proof is completed by bounding the right-hand side of the above inequality in a standard way as follows:

$$\begin{aligned}\sum_{k=0}^{d_s} \binom{n}{k} b^k \leqslant \sum_{k=0}^{d} \binom{n}{k} b^k \leqslant b^d \sum_{k=0}^{d} \frac{n^k}{k!} \leqslant b^d \sum_{k=0}^{d} \frac{d^k}{k!} \cdot \left(\frac{n}{d}\right)^k \\ \leqslant \left(\frac{bn}{d}\right)^d \sum_{k=0}^{d} \frac{d^k}{k!} \leqslant \left(\frac{enb}{d}\right)^d,\end{aligned}$$

where we used the convention that for all $k > n$, $\binom{n}{k} = 0$. □

References

[1] V. Vapnik, Statistical Learning Theory, John Wiley & Sons, Inc., New York, 1998.

[2] P. Bartlett, The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network, IEEE Transactions on Information Theory 44 (2) (1998) 525–536.

[3] A. Daniely, S. Sabato, S. Ben-David, S. Shalev-Shwartz, Multiclass learnability and the ERM principle, in: COLT'11, 2011, pp. 207–232.

[4] Ü. Doğan, T. Glasmachers, C. Igel, A unified view on multi-class support vector classification, Journal of Machine Learning Research 17(45) (2016) 1–32.

[5] A. Kontorovich, R. Weiss, Maximum margin muliclass nearest neighbors, in: ICML'14, 2014.

[6] T. Zhang, Statistical analysis of some multi-category large margin classification methods, Journal of Machine Learning Research 5 (2004) 1225–1251.

[7] Y. Lei, Ü. Doğan, A. Binder, M. Kloft, Multi-class SVMs: From tighter data-dependent generalization bounds to novel algorithms, in: NIPS 28, 2015, pp. 2026–2034.

[8] V. Kuznetsov, M. Mohri, U. Syed, Multi-class deep boosting, in: NIPS 27, 2014, pp. 2501–2509.

[9] Y. Guermeur, L_p-norm Sauer–Shelah lemma for margin multi-category classifiers, Journal of Computer and System Sciences 89 (2017) 450–473.

[10] V. Koltchinskii, D. Panchenko, Empirical margin distributions and bounding the generalization error of combined classifiers, The Annals of Statistics 30 (1) (2002) 1–50.

[11] M. Talagrand, Upper and Lower Bounds for Stochastic Processes: Modern Methods and Classical Problems, Springer-Verlag, Berlin Heidelberg, 2014.

[12] S. Mendelson, R. Vershynin, Entropy and the combinatorial dimension, Inventiones Mathematicae 152 (2003) 37–55.

[13] N. Alon, S. Ben-David, N. Cesa-Bianchi, D. Haussler, Scale-sensitive dimensions, uniform convergence, and learnability, Journal of the ACM 44 (4) (1997) 615–631.

[14] M. Anthony, P. Bartlett, Neural Network Learning: Theoretical Foundations, Cambridge University Press, Cambridge, 1999.

[15] R. Dudley, A course on empirical processes, in: Ecole d'été de Probabilités de Saint-Flour XII-1982, Springer, 1984, pp. 1–142.

[16] R. Dudley, Uniform central limit theorems, Cambridge University Press, 1999.

[17] M. Kearns, R. Schapire, Efficient distribution-free learning of probabilistic concepts, Journal of Computer and System Sciences 48 (3) (1994) 464–497.

[18] S. Mendelson, Rademacher averages and phase transitions in Glivenko-Cantelli classes, IEEE Transactions on Information Theory 48 (1) (2002) 251–263.

[19] S. Mendelson, G. Schechtman, The shattering dimension of sets of linear functionals, The Annals of Probability 32 (3) (2004) 1746–1770.

[20] P. Bartlett, J. Shawe-Taylor, Generalization performance of support vector machines and other pattern classifiers, in: B. Schölkopf, C. Burges, A. Smola (Eds.), Advances in Kernel Methods - Support Vector Learning, The MIT Press, Cambridge, MA, 1999, Ch. 4, pp. 43–54.

[21] L. A. Gottlieb, A. Kontorovich, R. Krauthgamer, Efficient classification for metric data, IEEE Transactions on Information Theory 60 (9) (2014) 5750–5759.

[22] Y. Guermeur, VC theory of large margin multi-category classifiers, Journal of Machine Learning Research 8 (2007) 2551–2594.

[23] A. Kolmogorov, V. Tihomirov, ϵ-entropy and ϵ-capacity of sets in functional spaces, American Mathematical Society Translations, series 2 17 (1961) 277–364.

[24] B. C. Berndt, Ramanujan's Notebooks, Part I, Springer-Verlag New York, 1985.