Combinatorial and Structural Results for γ-Ψ-dimensions

Yann Guermeur
LORIA-CNRS
Campus Scientifique, BP 239
54506 Vandœuvre-lès-Nancy Cedex, France
(e-mail: Yann.Guermeur@loria.fr)

May 23, 2022

Running Title: Combinatorial and Structural Results for γ-Ψ-dimensions
Keywords: margin multi-category classifiers, guaranteed risks, scale-sensitive combinatorial dimensions, γ-Ψ-dimensions
Mathematics Subject Classification: 68Q32, 62H30

Abstract

This article deals with the generalization performance of margin multi-category classifiers, when minimal learnability hypotheses are made. In that context, the derivation of a guaranteed risk is based on the handling of capacity measures belonging to three main families: Rademacher/Gaussian complexities, metric entropies and scale-sensitive combinatorial dimensions. The usefulness of the scale-sensitive combinatorial dimensions rests on the availability of two types of results. Combinatorial results connect them to metric entropies. Structural results perform the transition from the multi-class case to the bi-class one. The results currently available for the standard scale-sensitive combinatorial dimension, the fat-shattering dimension, make it useless. We establish the advantages springing from replacing it with two γ-Ψ-dimensions: the margin Graph dimension and the margin Natarajan dimension. Two major conclusions can be drawn:

1. involving the margin Graph dimension always improves the combinatorial results;
2. the margin Natarajan dimension can be used to exploit basic features of the classifier so as to bypass the main weakness of the fat-shattering dimension: its structural result.

1 Introduction

One of the main open problems of the theory of margin multi-category pattern classification is the characterization of the way the confidence interval of an upper bound on the probability of error should vary as a function of the three basic parameters which are the sample size m, the number C of categories and the margin parameter γ (see Kontorovich and Weiss, 2014, for a survey). When working under minimal learnability hypotheses, the derivation of such a *guaranteed risk* is based on the handling of capacity measures belonging to three main families: Rademacher/Gaussian complexities (Bartlett and Mendelson, 2002), metric entropies (Kolmogorov and Tihomirov, 1961) and scale-sensitive combinatorial dimensions (Kearns and Schapire, 1994; Guermeur, 2007). The usefulness of the scale-sensitive combinatorial dimensions to derive guaranteed risks rests on the availability of two types of results. *Combinatorial results* (Alon et al., 1997; Mendelson and Vershynin, 2003; Rudelson and Vershynin, 2006; Musayeva et al., 2019) connect them to metric entropies. *Structural results* (Duan, 2012; Maurer, 2016; Guermeur, 2017; Mohri et al., 2018) perform the transition from the multi-class case to the bi-class one. The structural result dedicated to the main scale-sensitive combinatorial dimension, the fat-shattering dimension (Kearns and

Schapire, 1994), is of no use to derive a bound. This motivates the exploration of an alternative option: replacing this dimension with the two main γ-Ψ-dimensions (Guermeur, 2007), i.e., the margin Graph dimension and the margin Natarajan dimension. This article introduces the corresponding combinatorial and structural results. The dependence on m, C and γ of the resulting guaranteed risks is then characterized. This establishes that involving the margin Graph dimension always improves the combinatorial results. Furthermore, the margin Natarajan dimension appears very promising to take into account basic features of the classifier so as to produce efficient structural results. When this happens, then the improvement of the confidence interval primarily regards the dependence on γ.

The organization of the paper is as follows. Section 2 introduces the theoretical framework. Section 3 highlights the need for new structural results to improve the multi-class bounds. Section 4 characterizes the connections between the three combinatorial dimensions considered. Sections 5 and 6 introduce and discuss the new combinatorial and structural results dedicated to the two γ-Ψ-dimensions. The corresponding bounds on the metric entropies and guaranteed risks are derived in Section 7. At last, we draw conclusions in Section 8. To make reading easier, all technical lemmas and proofs have been gathered in appendix.

2 Margin Multi-category Classifiers

We work under minimal assumptions on the data and the classifiers, which exhibit one important feature: for each description, they return one score per category.

2.1 Theoretical Framework

Let $[\![n_-; n_+]\!]$ denote the set of integers ranging from n_- to n_+. We consider the case of C-category pattern classification problems with $C \in \mathbb{N} \setminus [\![0; 2]\!]$. $\mathcal{X}$ is the description space and $\mathcal{Y} = [\![1; C]\!]$ the set of categories. Their connection is utterly characterized by an unknown probability measure P. Let $Z = (X, Y)$ be a random pair with values in $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, distributed according to P. We are given an 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 $\mathbf{Z}_m \sim P^m$). The classifiers are based on classes of vector-valued functions with one component function per category. We add a basic learnability hypothesis: the classes of component functions are *uniform Glivenko-Cantelli* (uGC) (Dudley et al., 1991). Those classes must be uniformly bounded up to additive constants. We replace this property by a slightly stronger one: the vector-valued

functions take their values in a hypercube of $\mathbb{R}^C$. To sum up, we make minimal hypotheses to ensure that all capacity measures met in the sequel are finite (none of the bounds formulated is trivial).

Definition 1 (Margin classifier)**.** *Let* $\{\mathcal{G}_k : 1 \leqslant k \leqslant C\}$ *be a set of classes of functions from* $\mathcal{X}$ *into* $[-M_{\mathcal{G}}, M_{\mathcal{G}}]$ *with* $M_{\mathcal{G}} \in [1, +\infty)$ $(\forall k \in [\![1; C]\!],\ \mathcal{G}_k \subset [-M_{\mathcal{G}}, M_{\mathcal{G}}]^{\mathcal{X}})$*. These classes are supposed to be uGC. Let* $\mathcal{G}$ *be a subset of* $\prod_{k=1}^{C} \mathcal{G}_k$ *(*$\mathcal{G}$ *is a class of functions from* $\mathcal{X}$ *into* $[-M_{\mathcal{G}}, M_{\mathcal{G}}]^C$*). An operator dr named* decision rule *maps every function* $g = (g_k)_{1 \leqslant k \leqslant C} \in \mathcal{G}$ *to a* margin multi-category classifier dr_g *in* $(\mathcal{Y} \bigcup \{*\})^{\mathcal{X}}$*. For every pair* $(g, x) \in \mathcal{G} \times \mathcal{X}$*,* $dr_g(x)$ *is either the index of the component function of* g *taking the highest value at* x*, or the dummy category* $*$ *in case of ex æquo.*

The generalization capabilities of such classifiers can be characterized by means of the values taken by the differences of the component functions. This calls for the definition of standard concepts of the theory of margin multi-category pattern classification.

Definition 2 (Margin operator ρ)**.** *Let* $\mathcal{G}$ *be a function class defined as in Definition 1. Define* ρ *as an operator on* $\mathcal{G}$ *such that:*

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

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

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

The risk of $g \in \mathcal{G}$ is given by: $L(g) = \mathbb{E}_{(X,Y)\sim P}\left[\mathbb{1}_{\{\rho_g(X,Y)\leqslant 0\}}\right] = P(\mathrm{dr}_g(X) \neq Y)$.

Definition 3 (Margin loss functions)**.** *A class of* margin loss functions ϕ_γ *parameterized by* $\gamma \in (0, 1]$ *is a class of nonincreasing functions from* $\mathbb{R}$ *into* $[0, 1]$ *satisfying:*

$$\begin{cases} \forall \gamma \in (0, 1],\ \phi_\gamma(0) = 1 \text{ and } \phi_\gamma(\gamma) = 0 \\ \forall (\gamma, \gamma') \in (0, 1]^2,\ \gamma < \gamma' \Longrightarrow \phi_{\gamma'} \text{ majorizes } \phi_\gamma \end{cases}.$$

Given ϕ_γ, the risk with margin γ of g, $L_\gamma(g)$, is defined as: $L_\gamma(g) = \mathbb{E}_{Z\sim P}[\phi_\gamma \circ \rho_g(Z)]$. $L_{\gamma,m}(g)$ designates the corresponding empirical risk, measured on $\mathbf{Z}_m$. When using ϕ_γ, the behaviour of the margin functions outside the interval $[0, \gamma]$ is irrelevant to characterize the generalization performance. The idea to exploit this property by means of a squashing operator can be traced back to Bartlett (1998). The present study uses the operator π_γ.

Definition 4 (Squashing operator π_γ)**.** *Let $\mathcal{F} \subset \mathbb{R}^{\mathcal{T}}$. For $\gamma \in (0,1]$, define the* piecewise-linear squashing operator π_γ *as:*

$$\begin{array}{rccc} \pi_\gamma : & \mathcal{F} & \longrightarrow & \mathcal{F}_\gamma \\ & f & \mapsto & f_\gamma \end{array}$$

$$\forall t \in \mathbb{R}, \quad f_\gamma(t) = f(t)\,\mathbb{1}_{\{f(t)\in(0,\gamma]\}} + \gamma \mathbb{1}_{\{f(t)>\gamma\}}.$$

When deriving a guaranteed risk, replacing the class of margin functions $\rho_{\mathcal{G}}$ with its image by π_γ, the class $\rho_{\mathcal{G},\gamma}$, has only advantages. On the one hand, it induces a decrease of the capacity (see Section 3.1) which can narrow the confidence interval. On the other hand, it does not affect the data-fit term (since $\forall \gamma \in (0,1]\,, \;\; \phi_\gamma \circ \pi_\gamma = \phi_\gamma$). Thus, making the best of it is a major challenge.

2.2 Guaranteed Risks

In the theoretical framework of interest, the starting point of the derivation of a guaranteed risk is a supremum inequality taking the form:

$$P^m \left\{ \sup_{g\in\mathcal{G}} \left(L_*(g) - L_{\gamma,m}(g) \right) > F_i\left(m, \gamma, \delta, \operatorname{cap}(\rho_{\mathcal{G},\gamma})\right) \right\} \leqslant \delta, \tag{1}$$

where L_* is either L or L_γ and the capacity measure $\operatorname{cap}(\rho_{\mathcal{G},\gamma})$ involved in the expression of the function F_i depends on the choice of ϕ_γ. Then, the problem consists in upper bounding $\operatorname{cap}(\rho_{\mathcal{G},\gamma})$ as a function of the basic parameters m, C and γ, so that eventually, with probability $1-\delta$, the supremum of the empirical process of interest is bounded from above by a function F_f of m, C, γ and δ only, i.e.,

$$\sup_{g\in\mathcal{G}} \left(L_*(g) - L_{\gamma,m}(g) \right) \leqslant F_f(m, C, \gamma, \delta).$$

We now introduce the three types of capacity measures considered in this study, in their order of appearance in the derivation of the bounds.

Definition 5 (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*

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

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 classes $\mathcal{F}$ considered here are endowed with empirical (pseudo-)metrics derived from the L_p-norms. For $n \in \mathbb{N}^*$, let $\mathbf{t}_n = (t_i)_{1 \leqslant i \leqslant n} \in \mathcal{T}^n$. Then,

$$\forall \left(f, f'\right) \in \mathcal{F}^2, \begin{cases} \forall p \in [1, +\infty), \ d_{p,\mathbf{t}_n}(f, f') = \left(\frac{1}{n}\sum_{i=1}^n |f(t_i) - f'(t_i)|^p\right)^{\frac{1}{p}} \\ d_{\infty,\mathbf{t}_n}(f, f') = \max_{1 \leqslant i \leqslant n} |f(t_i) - f'(t_i)| \end{cases}.$$

Definition 6 (Covering numbers and metric entropy)**.** *Let* $\mathcal{F}$ *be a class of real-valued functions on* $\mathcal{T}$ *endowed with the pseudo-metric* $d_{p,\mathbf{t}_n}$*. Let* $\tilde{\mathcal{F}}$ *be a totally bounded subset of* $(\mathcal{F}, d_{p,\mathbf{t}_n})$*. Then for every* $\epsilon \in \mathbb{R}_+^*$*, the* ϵ-covering number $\mathcal{N}\left(\epsilon, \tilde{\mathcal{F}}, d_{p,\mathbf{t}_n}\right)$ *of* $\tilde{\mathcal{F}}$ *is the minimal cardinality of a subset* $\bar{\mathcal{F}}$ *of* $\mathcal{F}$ *satisfying:*

$$\forall \tilde{f} \in \tilde{\mathcal{F}}, \ \exists \bar{f} \in \bar{\mathcal{F}} : \ \ d_{p,\mathbf{t}_n}\left(\tilde{f}, \bar{f}\right) < \epsilon.$$

The definition of the proper/internal covering number $\mathcal{N}^{int}\left(\epsilon, \tilde{\mathcal{F}}, d_{p,\mathbf{t}_n}\right)$ *results from the restriction* $\bar{\mathcal{F}} \subset \tilde{\mathcal{F}}$*. The* uniform covering numbers $\mathcal{N}_p\left(\epsilon, \tilde{\mathcal{F}}, n\right)$ *and* $\mathcal{N}_p^{int}\left(\epsilon, \tilde{\mathcal{F}}, n\right)$ *are given by:*

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

The function mapping ϵ *to the binary logarithm of the* ϵ*-covering number of a set is called the* metric entropy *of this set.*

There is a close connection between covering and packing properties of bounded subsets in pseudo-metric spaces.

Definition 7 (Packing numbers)**.** *Let the pseudo-metric space* $(\mathcal{F}, d_{p,\mathbf{t}_n})$ *be defined as in Definition 6. Then for every* $\epsilon \in \mathbb{R}_+^*$*, its subset* $\tilde{\mathcal{F}}$ *is* ϵ*-separated with respect to* $d_{p,\mathbf{t}_n}$ *if and only if:*

$$\forall \left\{f, f'\right\} \subset \tilde{\mathcal{F}}, \ d_{p,\mathbf{t}_n}\left(f, f'\right) \geqslant \epsilon.$$

If $\tilde{\mathcal{F}}$ *is totally bounded, then its* ϵ*-packing number* $\mathcal{M}\left(\epsilon, \tilde{\mathcal{F}}, d_{p,\mathbf{t}_n}\right)$ *is the maximal cardinality of its* ϵ*-separated subsets.* $\mathcal{M}_p\left(\epsilon, \tilde{\mathcal{F}}, n\right) = \sup_{\mathbf{t}_n \in \mathcal{T}^n} \mathcal{M}\left(\epsilon, \tilde{\mathcal{F}}, d_{p,\mathbf{t}_n}\right)$ *designates the corresponding* uniform packing number.

The scale-sensitive combinatorial dimensions evaluated here are γ-Ψ-dimensions, i.e., scale-sensitive extensions of the Ψ-dimensions (Ben-David et al., 1995).

Definition 8 (γ-Ψ-dimensions, Definition 28 in Guermeur, 2007)**.** *Let $\mathcal{F} \subset \mathbb{R}^{\mathcal{Z}}$ be such that:*

$$\forall f \in \mathcal{F},\ \forall x \in \mathcal{X},\ \max_{1 \leqslant k < l \leqslant C} \{f(x,k) + f(x,l)\} = 0. \tag{2}$$

Let Ψ be a family of mappings from $\mathcal{Y}$ into $\{-1,0,1\}$. For $\gamma \in \mathbb{R}_+^*$, a subset $s_{\mathcal{Z}^n} = \{z_i = (x_i, y_i) :\ 1 \leqslant i \leqslant n\}$ of $\mathcal{Z}$ is said to be γ-Ψ-shattered *by $\mathcal{F}$ if there is a vector $\boldsymbol{\psi}_n = \left(\psi^{(i)}\right)_{1 \leqslant i \leqslant n} \in \Psi^n$ satisfying $\left(\psi^{(i)}(y_i)\right)_{1 \leqslant i \leqslant n} = \mathbf{1}_n$, and 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*

$$\forall i \in [\![1;n]\!],\ \ s_i \left(s_i \max_{\left\{k:\ \psi^{(i)}(k) = s_i\right\}} f_{\mathbf{s}_n}(x_i,k) - b_i \right) \geqslant \gamma. \tag{3}$$

The γ-Ψ-dimension *of $\mathcal{F}$, denoted by γ-Ψ-$dim\,(\mathcal{F})$, is the maximal cardinality of a subset of $\mathcal{Z}$ γ-Ψ-shattered by $\mathcal{F}$, if such maximum exists. Otherwise, $\mathcal{F}$ is said to have infinite γ-Ψ-dimension.*

Remark 1. *Let us consider the degenerate case $C = 2$. Then, $s_i \max_{\left\{k:\ \psi^{(i)}(k) = s_i\right\}} f_{\mathbf{s}_n}(x_i,k) = f_{\mathbf{s}_n}(z_i)$, so that Formula* (3) *reduces to*

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

and thus Definition 8 reduces to the definition of the main scale-sensitive combinatorial dimension, the fat-shattering *or* γ-dimension *γ-dim (Kearns and Schapire, 1994; Bartlett and Long, 1998), with a restricted domain for vector $\mathbf{b}_n$. Furthermore, if we define the function class $\mathcal{F}_{(1)}$ on $\mathcal{X}$ as follows: $\mathcal{F}_{(1)} = \{f(\cdot,1) :\ f \in \mathcal{F}\}$, then*

$$\forall \gamma \in \mathbb{R}_+^*,\ \ \gamma\text{-}dim\left(\mathcal{F}_{(1)}\right) = \gamma\text{-}dim\left(\mathcal{F}\right),$$

provided that the definition of γ-$dim\left(\mathcal{F}_{(1)}\right)$ is the standard one (requiring only that $\mathbf{b}_n \in \mathbb{R}^n$). Thus, in the bi-class case, the constraint $\mathbf{b}_n \in \mathbb{R}_+^n$ of Definition 8 establishes the equivalence of the two definitions of the fat-shattering dimension, for the function class on $\mathcal{Z}$ (classifier with two outputs) and the one on $\mathcal{X}$ (classifier with one single output).

Definition 8 and Remark 1 suggest to adopt the following convention. The definition of any scale-sensitive combinatorial dimension of a class of functions with domain $\mathcal{Z}$ includes the restriction $\mathbf{b}_n \in \mathbb{R}_+^n$. On the contrary, when the domain is $\mathcal{X}$, then the standard hypothesis $\mathbf{b}_n \in \mathbb{R}^n$ applies. The relevance of this choice will appear gradually (sometimes implicitly) in the sequel.

Definition 9 (Margin Graph dimension and margin Natarajan dimension)**.** *Let $\mathcal{F}$ be a function class defined as in Definition 8 and let $\gamma \in \mathbb{R}_+^*$. The* Graph dimension with margin *γ of $\mathcal{F}$, denoted by γ-G-dim$(\mathcal{F})$, is the γ-Ψ-dimension of $\mathcal{F}$ corresponding to the following choice for Ψ:*

$$\Psi_G = \left\{ \left(\psi_k : y \mapsto \mathbb{1}_{\{y=k\}} - \mathbb{1}_{\{y \neq k\}} \right) : \ k \in \mathcal{Y} \right\}.$$

The Natarajan dimension with margin *γ of $\mathcal{F}$, denoted by γ-N-dim$(\mathcal{F})$, is the γ-Ψ-dimension of $\mathcal{F}$ corresponding to the following choice for Ψ:*

$$\Psi_N = \left\{ \left(\psi_{k,l} : y \mapsto \mathbb{1}_{\{y=k\}} - \mathbb{1}_{\{y=l\}} \right) : \ \{k,l\} \subset \mathcal{Y} \right\}.$$

Remark 2. *The instance of* (3) *associated with the margin Graph dimension is obtained by setting $\boldsymbol{\psi}_n = (\psi_{y_i})_{1 \leqslant i \leqslant n}$ so that*

$$\forall i \in [\![1;n]\!], \ \begin{cases} \text{if } s_i = 1, \ f_{\mathbf{s}_n}(x_i, y_i) - b_i \geqslant \gamma \\ \text{if } s_i = -1, \ \max_{k \neq y_i} f_{\mathbf{s}_n}(x_i, k) + b_i \geqslant \gamma \end{cases}.$$

In the case of the Natarajan dimension with margin γ, choosing $\boldsymbol{\psi}_n$ is equivalent to choosing a vector $\mathbf{c}_n = (c_i)_{1 \leqslant i \leqslant n} \in \mathcal{Y}^n$ satisfying for every $i \in [\![1;n]\!]$, $c_i \neq y_i$. Then, $\boldsymbol{\psi}_n$ is set equal to $(\psi_{y_i,c_i})_{1 \leqslant i \leqslant n}$, so that (3) *becomes*

$$\forall i \in [\![1;n]\!], \ \begin{cases} \text{if } s_i = 1, \ f_{\mathbf{s}_n}(x_i, y_i) - b_i \geqslant \gamma \\ \text{if } s_i = -1, \ f_{\mathbf{s}_n}(x_i, c_i) + b_i \geqslant \gamma \end{cases}.$$

2.3 Scheme of Derivation of the Guaranteed Risks

For all known instances of Formula (1), the scheme of derivation of function F_f involving the families of capacity measures considered in this study is standard. It corresponds to the directed graph depicted in Figure 1.

Here, the function class $\mathcal{G}_0$ is equal to $\bigcup_{k=1}^{C} \mathcal{G}_k$. The value of m' is either m or $2m$, when the derivation of Inequality (1) involves a ghost sample (Vapnik and Chervonenkis, 1971; Pollard, 1984). When following a path from the source to the target, two types of transitions are met. A first group, the horizontal arrows, corresponds to a change of capacity measure. The standard sequence (from left to right) consists in the chaining method (Dudley, 1967; Talagrand, 2014), to connect the Rademacher complexity to covering numbers, a transition through the corresponding packing numbers, and then a combinatorial result, to switch to a combinatorial dimension. The second group, the layer of vertical arrows, is that of the structural results, performing the transition from the capacity of $\rho_{\mathcal{G},\gamma}$ to that of $\mathcal{G}_0$

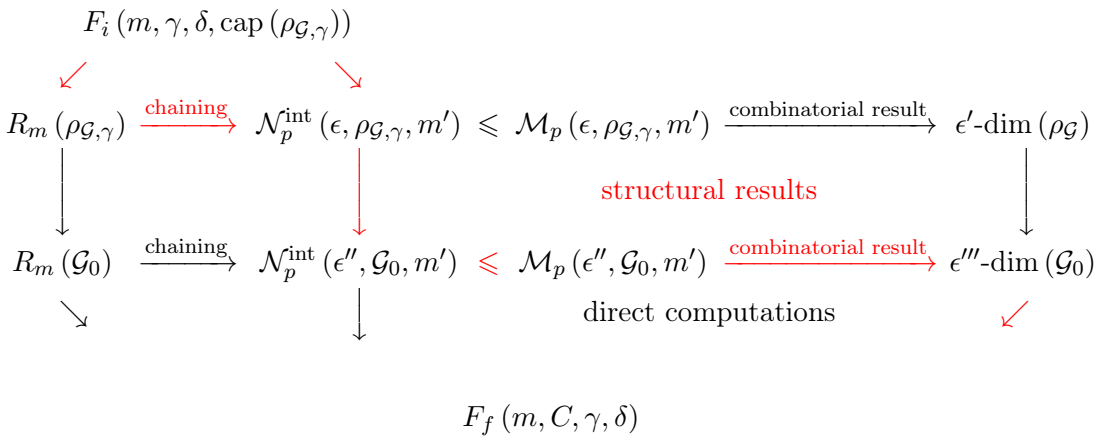

Figure 1: Graph of the transitions from a function F_i to a function F_f.

(roughly speaking from the multi-class case to the bi-class one). As an example, the paths in red are the ones explored in Guermeur (2017).

3 State-of-the-Art Structural Results

The literature provides us with structural results for all three types of capacity measures considered.

3.1 Major Lemmas

The sharpest structural result for the Rademacher complexity of the class $\rho_{\mathcal{G},\gamma}$ is obtained by combining the proof of Theorem 9.2 in Mohri et al. (2018) with Talagrand's contraction lemma (see for instance Lemma 5.7 in Mohri et al., 2018).

Lemma 1. *Let $\mathcal{G}$ be a function class defined as in Definition 1. Then,*

$$\forall \gamma \in (0,1]\,,\ \forall n \in \mathbb{N}^*,\ R_n\left(\rho_{\mathcal{G},\gamma}\right) \leqslant \min\left\{R_n\left(\rho_{\mathcal{G}}\right), CR_n\left(\mathcal{G}_0\right)\right\}.$$

The counterpart of Lemma 1 dealing with covering numbers is the following structural result.

Lemma 2 (Lemma 1 in Guermeur, 2017)**.** *Let $\mathcal{G}$ be a function class defined as in Definition 1. For every $\gamma \in (0,1]$, $\epsilon \in \mathbb{R}_+^*$, $n \in \mathbb{N}^*$, $p \in [1,+\infty]$, and $\mathbf{z}_n = ((x_i,y_i))_{1\leqslant i\leqslant n} \in \mathcal{Z}^n$,*

$$\mathcal{N}^{int}\left(\epsilon,\rho_{\mathcal{G},\gamma},d_{p,\mathbf{z}_n}\right) \leqslant \mathcal{N}^{int}\left(\epsilon,\rho_{\mathcal{G}},d_{p,\mathbf{z}_n}\right) \leqslant \prod_{k=1}^{C}\mathcal{N}^{int}\left(C^{-\frac{1}{p}}\epsilon,\mathcal{G}_k,d_{p,\mathbf{x}_n}\right) \leqslant \left(\mathcal{N}^{int}\left(C^{-\frac{1}{p}}\epsilon,\mathcal{G}_0,d_{p,\mathbf{x}_n}\right)\right)^C,$$

where $\mathbf{x}_n = (x_i)_{1 \leqslant i \leqslant n}$.

The main method available to derive structural results for the γ-dimension (see for instance the proof of Lemma 6.2 in Duan, 2012) consists in three main steps: upper bounding the dimension of interest in terms of a metric entropy of the same class, applying a decomposition (similar to Lemma 2), and applying a combinatorial result. For the class $\rho_{\mathcal{G},\gamma}$, it gives birth to the following Lemma.

Lemma 3. *Let $\mathcal{G}$ be a function class defined as in Definition 1. For every $\gamma \in (0,1]$ and $\epsilon \in \left(0, \frac{\gamma}{2}\right]$,*

$$\begin{aligned}
\epsilon\text{-}dim\left(\rho_{\mathcal{G},\gamma}\right) &\leqslant \epsilon\text{-}dim\left(\rho_{\mathcal{G}}\right) \\
&\leqslant 320 \log_2\left(\frac{24 M_{\mathcal{G}} \sqrt{C}}{\epsilon}\right) \sum_{k=1}^{C} \left(\frac{\epsilon}{96\sqrt{C}}\right) \text{-}dim\left(\mathcal{G}_k\right) \\
&\leqslant 320 \log_2\left(\frac{24 M_{\mathcal{G}} \sqrt{C}}{\epsilon}\right) C \left(\frac{\epsilon}{96\sqrt{C}}\right) \text{-}dim\left(\mathcal{G}_0\right). \qquad (4)
\end{aligned}$$

3.2 Shortcomings of the State-of-the-Art Structural Results

We reviewed the state-of-the-art decomposition results associated with the three families of capacity measures involved in this study. None is utterly satisfactory. The decomposition involving Rademacher complexities (Lemma 1) can produce a function F_f depending at least linearly on C. This behaviour is to be compared with the one of the decomposition involving covering numbers (Lemma 2), that always ensures a sublinear dependence (see for instance Theorem 3 in Musayeva et al., 2019). Furthermore, the upper bounds on the metric entropies resulting from applying a combinatorial result to $\rho_{\mathcal{G},\gamma}$ followed by Lemma 3 are always worse than those obtained by application of Lemma 2 followed by the same combinatorial result (applied to $\mathcal{G}_0$). The reason is simple: the two computations are similar, except for an extra application of a combinatorial result in the first case. This supplementary step introduces a multiplicative factor $\ln\left(\frac{1}{\epsilon}\right)$ in the bounds. We illustrate the phenomenon with the two state-of-the-art combinatorial results connecting the packing numbers of a class of real-valued functions to its fat-shattering dimension: Lemma 3.5 in Alon et al. (1997), for the L_∞-norm, and Theorem 1 in Mendelson and Vershynin (2003), for the L_2-norm. The use of Lemma 2 produces for $n \geqslant \left(\frac{\epsilon}{4}\right)\text{-dim}\left(\mathcal{G}_0\right)$,

$$\log_2\left(\mathcal{N}_\infty^{\text{int}}\left(\epsilon, \rho_{\mathcal{G},\gamma}, n\right)\right) \leqslant C \left\{ \left\lceil \left(\frac{\epsilon}{4}\right)\text{-dim}\left(\mathcal{G}_0\right) \log_2\left(\frac{4 M_{\mathcal{G}} e n}{\epsilon}\right) \right\rceil \log_2\left(\frac{16 M_{\mathcal{G}}^2 n}{\epsilon^2}\right) + 1 \right\} \qquad (5)$$

and for $n \in \mathbb{N}^*$,

$$\log_2\left(\mathcal{N}_2^{\text{int}}\left(\epsilon, \rho_{\mathcal{G},\gamma}, n\right)\right) \leqslant 20C\left(\frac{\epsilon}{48\sqrt{C}}\right)\text{-dim}\left(\mathcal{G}_0\right)\log_2\left(\frac{12M_{\mathcal{G}}\sqrt{C}}{\epsilon}\right). \tag{6}$$

Using instead Lemma 3 gives for $n \geqslant \left(\frac{\epsilon}{4}\right)\text{-dim}\left(\rho_{\mathcal{G}}\right)$,

$$\begin{aligned}&\log_2\left(\mathcal{N}_\infty^{\text{int}}\left(\epsilon, \rho_{\mathcal{G},\gamma}, n\right)\right)\\ &\leqslant \left\lceil 320\log_2\left(\frac{96M_{\mathcal{G}}\sqrt{C}}{\epsilon}\right)C\left(\frac{\epsilon}{384\sqrt{C}}\right)\text{-dim}\left(\mathcal{G}_0\right)\log_2\left(\frac{2\gamma en}{\epsilon}\right)\right\rceil\log_2\left(\frac{4\gamma^2 n}{\epsilon^2}\right)+1\end{aligned} \tag{7}$$

and for $n \in \mathbb{N}^*$,

$$\log_2\left(\mathcal{N}_2^{\text{int}}\left(\epsilon, \rho_{\mathcal{G},\gamma}, n\right)\right) \leqslant 6400\log_2\left(\frac{1152M_{\mathcal{G}}\sqrt{C}}{\epsilon}\right)C\left(\frac{\epsilon}{4608\sqrt{C}}\right)\text{-dim}\left(\mathcal{G}_0\right)\log_2\left(\frac{6\gamma}{\epsilon}\right). \tag{8}$$

A comparison of Inequalities (5) and (7) on the one hand, and Inequalities (6) and (8) on the other, makes it possible to identify the extra logarithmic factors. With these observations at hand, one could think that when the assumptions regarding the classifier are minimal, then the best structural result is Lemma 2. Put in a different way, the paths in the graph of Figure 1 generating the best functions F_f could be those marked in red. However, Lemma 2 makes no use of the capacity reduction induced by the squashing (operator π_γ). When delaying the decomposition at this level, this squashing is only exploited upstream, i.e., by the chaining formula. Those limitations raise a question.

Question 1. *Can the introduction in the graph of transitions of γ-Ψ-dimensions of $\rho_{\mathcal{G}}$ improve the dependence of function F_f on the basic parameters?*

The answers to Question 1 should spring from replacing the classical graph of transitions (Figure 1) with the one of Figure 2 and exploring in the new graph the paths highlighted in blue.

This exploration takes the form of the derivation of new combinatorial and structural results. The evaluation of this contribution rests on connections between the scale-sensitive combinatorial dimensions which are established in the following section.

4 Connections Between the Scale-Sensitive Combinatorial Dimensions

We first consider the connections between the fat-shattering dimension and the margin Graph dimension. It takes the form of a kind of "equivalence".

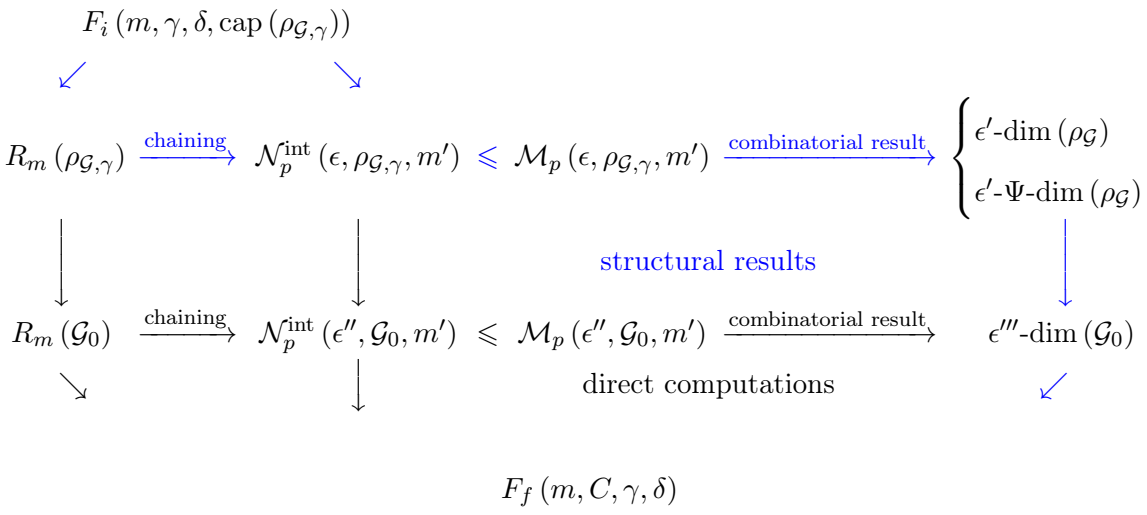

Figure 2: Paths from F_i to F_f involving combinatorial dimensions of the class $\rho_{\mathcal{G}}$.

Lemma 4. *Let $\mathcal{F}$ be a function class defined as in Definition 8. Then,*

$$\forall\gamma\in(0,1]\,,\ \forall\epsilon\in\left(0,\frac{\gamma}{2}\right]\,,\ \ \epsilon\text{-}dim\,(\mathcal{F}_\gamma)\leqslant\epsilon\text{-}G\text{-}dim\,(\mathcal{F})\leqslant\epsilon\text{-}dim\,(\mathcal{F})\,. \tag{9}$$

To assess the scope of Lemma 4, one must consider the strategy implemented to establish combinatorial results for function classes $\mathcal{F}_\gamma$ under the sole hypothesis that the fat-shattering dimension of $\mathcal{F}$ is defined. This strategy, used for instance in Bartlett (1998), simply consists in upper bounding the packing numbers of $\mathcal{F}_\gamma$ as a function of its fat-shattering dimension, and then upper bounding this dimension by the fat-shattering dimension of $\mathcal{F}$. Lemma 4 tells us that if the class $\mathcal{F}$ also has γ-Ψ-dimensions, then its fat-shattering dimension can be replaced with its (smaller) margin Graph dimension, thus improving the combinatorial result (whatever the L_p-norm used). The gain is conditioned by the availability of non trivial upper bounds on the margin Graph dimension of $\mathcal{F}$, i.e., bounds that no longer hold true if this dimension is replaced with the fat-shattering dimension of $\mathcal{F}$. Lemma 5, a scale-sensitive counterpart of Theorem 10 in Ben-David et al. (1995), provides a bound of this kind.

Lemma 5. *Let $\mathcal{F}$ be a function class defined as in Definition 8. Suppose that $\gamma\in\mathbb{R}_+^*$ is such that γ-G-dim $(\mathcal{F})$ is finite. Then,*

$$\gamma\text{-}N\text{-}dim\,(\mathcal{F})\leqslant\gamma\text{-}G\text{-}dim\,(\mathcal{F})\leqslant 32\log_2^2\left(e\,(C-1)\right)\gamma\text{-}N\text{-}dim\,(\mathcal{F})^{\alpha(C)}\,, \tag{10}$$

where $\alpha\,(C)=1+\frac{1}{4\ln(C-1)+2}$.

It is easy to exhibit examples of function classes $\mathcal{F}$ with γ-Ψ-dimensions, and for which the right-hand side inequality of Formula (10) no longer holds true with γ-G-dim $(\mathcal{F})$ replaced with γ-dim $(\mathcal{F})$. Example 1 is of this kind.

Example 1. *Let $\mathcal{G}$ be a set of two functions $g^{(1)}$ and $g^{(2)}$ on $\mathcal{X} = \{x\}$ given by $g^{(1)}(x) = \left(\frac{3}{4}, \frac{1}{4}, 0\right)^T$ and $g^{(2)}(x) = \left(0, \frac{1}{2}, \frac{1}{2}\right)^T$. Let us set $\gamma = \frac{1}{4}$. Then, γ-G-dim $(\rho_{\mathcal{G}}) = 0$ and γ-dim $(\rho_{\mathcal{G}}) = 1$.*

Indeed, by definition, $\left(\rho_{g^{(1)}}(x,k)\right)_{1 \leqslant k \leqslant 3} = \left(\frac{1}{4}, -\frac{1}{4}, -\frac{3}{8}\right)^T$, and $\left(\rho_{g^{(2)}}(x,k)\right)_{1 \leqslant k \leqslant 3} = \left(-\frac{1}{4}, 0, 0\right)^T$. Consequently, the formula $\max_{k \neq l} \left\{\rho_{g^{(1)}}(x,k) + \rho_{g^{(2)}}(x,l)\right\} = \gamma < 2\gamma$ implies that none of the three singletons $\{(x,k)\}$ is γ-G-shattered by $\rho_{\mathcal{G}}$. On the contrary,

$$\begin{cases} \rho_{g^{(1)}}(x,1) \geqslant \gamma \\ -\rho_{g^{(2)}}(x,1) \geqslant \gamma \end{cases},$$

i.e., the class $\rho_{\mathcal{G}}$ γ-shatters $\{(x,1)\}$ for $b = 0$. To sum up, γ-dim $(\rho_{\mathcal{G}}) = 1$ but γ-N-dim $(\rho_{\mathcal{G}}) = \gamma$-G-dim $(\rho_{\mathcal{G}}) = 0$ (so that the last term of Formula (10) is also equal to 0).

5 Combinatorial Results

The new results exposed in this section and the next one provide the building blocks needed to derive upper bounds on the metric entropies of $\rho_{\mathcal{G},\gamma}$, for $p \geqslant 2$, following the blue paths of Figure 2. However, the combinatorial results are more general, since they apply to any pair $(\mathcal{F}_\gamma, \mathcal{F})$ where $\mathcal{F}$ is a function class for which the γ-Ψ-dimensions are defined, and not only the pairs $(\rho_{\mathcal{G},\gamma}, \rho_{\mathcal{G}})$. To keep the comparison with the literature simple, we focus on the two most popular options: $p = \infty$ and $p = 2$, but the generalization is straightforward using the ideas developed in the proof of Theorem 2 in Musayeva et al. (2019).

5.1 Margin Graph Dimension

In the case of the margin Graph dimension, the availability of Lemma 4 has a major consequence: the new combinatorial results should be compared to those of the literature applied to $\mathcal{F}_\gamma$, with the fat-shattering dimension of $\mathcal{F}_\gamma$ replaced with the margin Graph dimension of $\mathcal{F}$ (instead of its fat-shattering dimension, as is usually done). Our first result deals with the case $p = \infty$.

Lemma 6. *Let $\mathcal{F}$ be a function class defined as in Definition 8. For $\epsilon \in \mathbb{R}_+^*$, let $d_G(\epsilon) = \epsilon$-G-dim$(\mathcal{F})$. Then for every $\gamma \in (0,1]$, $\epsilon \in (0,\gamma]$ and $n \in \mathbb{N}^*$ such that $n \geqslant d_G\left(\frac{\epsilon}{4}\right)$,*

$$\mathcal{M}_\infty(\epsilon, \mathcal{F}_\gamma, n) \leqslant \left(\frac{6\gamma n}{\epsilon}\right)^{d_G\left(\frac{\epsilon}{4}\right)\log_2\left(\frac{2\gamma e n}{d_G\left(\frac{\epsilon}{4}\right)\epsilon}\right)}. \tag{11}$$

Inequality (11) compares with the application of Lemma 3.5 in Alon et al. (1997). This application produces

$$\mathcal{M}_\infty(\epsilon, \mathcal{F}_\gamma, n) < 2\left(\frac{4\gamma^2 n}{\epsilon^2}\right)^{\left\lceil d_G\left(\frac{\epsilon}{4}\right)\log_2\left(\frac{2\gamma e n}{d_G\left(\frac{\epsilon}{4}\right)\epsilon}\right)\right\rceil}.$$

A gain can be noticed, which appears especially clearly for $\epsilon = \frac{\gamma}{2}$, the combination of practical interest as will be seen in Section 7.2. We now turn to the case $p = 2$.

Lemma 7. *Let $\mathcal{F}$ be a function class defined as in Definition 8. For $\epsilon \in \mathbb{R}_+^*$, let $d_G(\epsilon) = \epsilon$-G-dim$(\mathcal{F})$. Then for every $\gamma \in (0,1]$, $\epsilon \in (0,\gamma]$ and $n \in \mathbb{N}^*$,*

$$\mathcal{M}_2(\epsilon, \mathcal{F}_\gamma, n) \leqslant \left(\frac{5\gamma}{\epsilon}\right)^{12 d_G\left(\frac{\epsilon}{24}\right)}. \tag{12}$$

Inequality (12) compares with the formula obtained with Theorem 1 in Mendelson and Vershynin (2003). This time, the improvement is limited to an optimization of the constants which is not induced by the direct connection between the packing numbers of $\mathcal{F}_\gamma$ and the margin Graph dimension of $\mathcal{F}$.

5.2 Margin Natarajan Dimension

As for the margin Natarajan dimension, with Lemma 5 at hand, Lemmas 6 and 7 also provide us with combinatorial results involving this capacity measure. However, sharper bounds should spring from following the direct path, i.e., working directly with this latter dimension (without involving the margin Graph dimension). We now state the corresponding combinatorial results (for $p = \infty$ then $p = 2$) and perform the comparison.

Lemma 8. *Let $\mathcal{F}$ be a function class defined as in Definition 8. For $\epsilon \in \mathbb{R}_+^*$, let $d_N(\epsilon) = \epsilon$-N-dim$(\mathcal{F})$. Then for every $\gamma \in (0,1]$, $\epsilon \in (0,\gamma]$ and $n \in \mathbb{N}^*$ such that $n \geqslant d_N\left(\frac{\epsilon}{4}\right)$,*

$$\mathcal{M}_\infty(\epsilon, \mathcal{F}_\gamma, n) \leqslant \left(\frac{6\gamma\sqrt{C-1}n}{\epsilon}\right)^{d_N\left(\frac{\epsilon}{4}\right)\log_2\left(\frac{2\gamma(C-1)en}{d_N\left(\frac{\epsilon}{4}\right)\epsilon}\right)}. \tag{13}$$

Lemma 9. *Let $\mathcal{F}$ be a function class defined as in Definition 8. For $\epsilon \in \mathbb{R}_+^*$, let $d_N(\epsilon) = \epsilon$-N-dim$(\mathcal{F})$. Then for every $\gamma \in (0,1]$, $\epsilon \in (0,\gamma]$ and $n \in \mathbb{N}^*$,*

$$\mathcal{M}_2(\epsilon, \mathcal{F}_\gamma, n) \leqslant \left((C-1)\left(\frac{4\gamma}{\epsilon}\right)^5\right)^{\frac{3}{2}\log_2\left(2\left(\frac{14\gamma}{\epsilon}\right)^2(C-1)\right)d_N\left(\frac{\epsilon}{28}\right)}. \tag{14}$$

As expected, as close to 1 as $\alpha(C)$ may be, Inequalities (13) and (14) are better than the bounds obtained by substitution of the right-hand side term of (10) in the formulas involving the margin Graph dimension: (11) and (12), respectively. Precisely, in both cases, the dependence on n is unchanged, while the dependences on C and ϵ are slightly improved.

5.3 Discussion

Overall, we have seen that the combinatorial results involving directly the margin Graph dimension, Lemmas 6 and 7, provide sharper bounds on the packing numbers of $\mathcal{F}_\gamma$ (and thus also $\rho_{\mathcal{G},\gamma}$) than their counterparts from the literature: Lemma 3.5 in Alon et al. (1997) and Theorem 1 in Mendelson and Vershynin (2003), respectively. The gain is more important in the first case. Furthermore, the use of the margin Natarajan dimension appears as a promising alternative. Indeed, Lemmas 8 and 9 provide sharper bounds than those resulting from the combination of Lemmas 6 and 7 with Lemma 5.

6 Structural Results for the Margin Natarajan Dimension

Since the margin Graph dimension is upper bounded by the fat-shattering dimension (Lemma 4), Lemma 3 is also a structural result for the margin Graph dimension. Loosely speaking, this result can be improved by substituting in its proof the L_2-norm with the norm $L_{\lceil \log_2(C) \rceil}$, i.e., by applying the idea developed in Musayeva et al. (2019). The gain then regards the dependence of the bound on C. However, the prohibitive drawback identified in Section 3.2 remains. It is possible to upper bound directly (without resorting to Lemma 3) the margin Graph dimension of specific classifiers from the literature, but this comes at the expense of difficult computations, that go beyond the scope of this article. In that respect, it is far easier to exploit the appealing properties of the margin Natarajan dimension, as will be seen in the sequel.

Lemma 10. *Let $\mathcal{G}$ be a function class defined as in Definition 1 and let $\mathcal{D}_\mathcal{G}$ be the function class $\left\{\frac{1}{2}(g_k - g_l) : g \in \mathcal{G},\ 1 \leqslant k < l \leqslant C\right\}$. Then for every value of γ in $(0, M_\mathcal{G}]$,*

$$\gamma\text{-}N\text{-}dim(\rho_\mathcal{G}) \leqslant \binom{C}{2} \cdot \gamma\text{-}dim(\mathcal{D}_\mathcal{G}) \tag{15}$$

and

$$\gamma\text{-}N\text{-}dim(\rho_\mathcal{G}) \leqslant 384\binom{C}{2} \log_2\left(\frac{20M_\mathcal{G}}{\gamma}\right)\left(\frac{\gamma}{48}\right)\text{-}dim(\mathcal{G}_0). \tag{16}$$

Formula (16) is almost as unsatisfactory as Formula (4). However, Formula (15) can be used to derive a sharper structural result for popular classifiers associated with function classes $\mathcal{G}$ such that $\mathcal{G}_0$ has the closure property $\mathcal{D}_{\mathcal{G}} \subset \mathcal{G}_0$. Corollary 1 illustrates this behaviour.

Corollary 1. *Let $\mathcal{H}^{(1)}$ be a class of functions from $\mathcal{X}$ into a Hilbert space $(\mathbf{H}, \langle \cdot, \cdot \rangle_{\mathbf{H}})$ and $(\Lambda_1, \Lambda_2) \in \left(\mathbb{R}_+^*\right)^2$. Let $\mathcal{H}^{(2)}$ be the class of functions $h^{(2)}$ from $\mathcal{X}$ into $[-\Lambda_1\Lambda_2, \Lambda_1\Lambda_2]^C$ of the form:*

$$\forall x \in \mathcal{X}, \;\; h^{(2)}(x) = \left(\left\langle \mathbf{w}_k, h^{(1)}(x) \right\rangle_{\mathbf{H}} \right)_{1 \leqslant k \leqslant C},$$

where $h^{(1)} \in \mathcal{H}^{(1)}$ satisfies $\sup_{x \in \mathcal{X}} \left\| h^{(1)}(x) \right\|_{\mathbf{H}} \leqslant \Lambda_1$ and the vector $(\mathbf{w}_k)_{1 \leqslant k \leqslant C} \in \mathbf{H}^C$ satisfies $\max_{1 \leqslant k \leqslant C} \|\mathbf{w}_k\|_{\mathbf{H}} \leqslant \Lambda_2$. Let $\mathcal{H}_0^{(2)}$ be the class of all the component functions of the functions in $\mathcal{H}^{(2)}$. Then,

$$\forall \gamma \in (0, \Lambda_1\Lambda_2], \;\; \gamma\text{-}N\text{-}dim\left(\rho_{\mathcal{H}^{(2)}}\right) \leqslant \binom{C}{2} \cdot \gamma\text{-}dim\left(\mathcal{H}_0^{(2)}\right). \tag{17}$$

Corollary 1 is actually a consequence of Formula (15) since the class $\mathcal{H}^{(2)}$ has been specified so as to ensure the satisfaction of the closure property. It applies to classifiers of reference such as the multi-layer perceptrons (MLPs) (Anthony and Bartlett, 1999) with linear output units and the C-category support vector machines (SVMs) (Doğan et al., 2016). In the second case, $\mathcal{H}^{(1)}$ is restricted to one single function, the feature map, which can be defined from the kernel κ (Berlinet and Thomas-Agnan, 2004) as: $\forall x \in \mathcal{X}$, $h^{(1)}(x) = \kappa_x = \kappa(\cdot, x)$. Then, $\mathbf{H}$ is the reproducing kernel Hilbert space (RKHS) of κ. An application of the standard upper bound on the fat-shattering dimension of (binary) SVMs, Theorem 4.6 in Bartlett and Shawe-Taylor (1999), produces the instantiation of Inequality (17) for C-category SVMs:

$$\forall \gamma \in (0, \Lambda_1\Lambda_2], \;\; \gamma\text{-N-dim}\left(\rho_{\mathcal{H}^{(2)}}\right) \leqslant \binom{C}{2} \left(\frac{\Lambda_1\Lambda_2}{\gamma}\right)^2. \tag{18}$$

It is noticeable that such a simple algebraic property as $\mathcal{D}_{\mathcal{G}} \subset \mathcal{G}_0$ proves enough to replace Inequality (16) with a far sharper bound. Corollary 1 provides a first illustration of the capacity of the margin Natarajan dimension to exploit a coupling between the component functions of the classifier. This exploitation can be carried out further. We illustrate the phenomenon by refining the study of the case of the C-category SVMs. To than end, a specific definition of these machines is used, which is based on the concept of RKHS of $\mathbb{R}^C$-valued functions (Wahba, 1992).

Definition 10 (RKHS $\mathbf{H}_{\kappa,C}$)**.** *Let κ be a real-valued positive type function on $\mathcal{X}^2$ and let $\left(\mathbf{H}_\kappa, \langle\cdot,\cdot\rangle_{\mathbf{H}_\kappa}\right)$ be its RKHS. Let $\tilde{\kappa}$ be the real-valued positive type function on $\mathcal{Z}^2$ deduced from κ as follows: $\forall\,(z,z') \in \mathcal{Z}^2$, $\tilde{\kappa}\,(z,z') = \delta_{y,y'}\kappa\,(x,x')$, where δ is the Kronecker delta. For every $z \in \mathcal{Z}$, let us define the $\mathbb{R}^C$-valued function $\tilde{\kappa}_z^{(C)}$ on $\mathcal{X}$ by the formula*

$$\tilde{\kappa}_z^{(C)}\,(\cdot) = \left(\tilde{\kappa}\,(z,(\cdot,k))\right)_{1\leqslant k\leqslant C}\,. \tag{19}$$

The RKHS of $\mathbb{R}^C$-valued functions at the basis of a C-category SVM with kernel κ, $\left(\mathbf{H}_{\kappa,C}, \langle\cdot,\cdot\rangle_{\mathbf{H}_{\kappa,C}}\right)$, consists of the linear manifold of all finite linear combinations of functions of the form (19) and its closure with respect to the inner product: $\forall\,(z,z') \in \mathcal{Z}^2$, $\left\langle \tilde{\kappa}_z^{(C)}, \tilde{\kappa}_{z'}^{(C)} \right\rangle_{\mathbf{H}_{\kappa,C}} = \tilde{\kappa}\,(z,z')$.

With Definition 10 at hand, the specification of the function class at the basis of a C-category SVM rests on the condition controlling the capacity through a coupling between the component functions. We consider the standard one, used for instance by Lei et al. (2015).

Definition 11 (Function class $\mathcal{H}_\Lambda$)**.** *Let κ be a kernel on $\mathcal{X}^2$ and let $\Lambda \in \mathbb{R}_+^*$. Let $\left(\mathbf{H}_{\kappa,C}, \langle\cdot,\cdot\rangle_{\mathbf{H}_{\kappa,C}}\right)$ be the RKHS of $\mathbb{R}^C$-valued functions spanned by κ according to Definition 10. Then the function class $\mathcal{H}_\Lambda$ associated with the C-category SVM parameterized by (κ,Λ) is: $\mathcal{H}_\Lambda = \left\{h = (h_k)_{1\leqslant k\leqslant C} \in \mathbf{H}_{\kappa,C} :\ \sum_{k=1}^C h_k = 0_{\mathbf{H}_\kappa} \text{ and } \|h\|_{\mathbf{H}_{\kappa,C}} \leqslant \Lambda\right\}$.*

The class $\mathcal{H}_\Lambda$ can be seen as an instance of the class $\mathcal{H}^{(2)}$ of Corollary 1 for which $\Lambda_1 = \sup_{x\in\mathcal{X}} \|\kappa_x\|_{\mathbf{H}_\kappa}$ and $\Lambda_2 = 2^{-\frac{1}{2}}\Lambda$. This instance exhibits a stronger coupling between the component functions. Lemma 11 takes this property into account to produce a sharper bound on the margin Natarajan dimension.

Lemma 11. *For $\Lambda \in \mathbb{R}_+^*$, let $\mathcal{H}_\Lambda$ be the function class of Definition 11. Suppose that for every $x \in \mathcal{X}$, κ_x belongs to the closed ball of radius $\Lambda_\mathcal{X}$ about the origin in $\mathbf{H}_\kappa$. Then,*

$$\forall\gamma \in (0, \Lambda\Lambda_\mathcal{X}]\,,\ \gamma\text{-}N\text{-}dim\,(\rho_{\mathcal{H}_\Lambda}) \leqslant C\left(\frac{\Lambda\Lambda_\mathcal{X}}{2\gamma}\right)^2. \tag{20}$$

In words, the stronger coupling between the component functions could be exploited so as to turn the quadratic dependence on C of Formula (18) into a linear one.

7 Guaranteed Risks

The combinatorial and structural results of the two previous sections provide us with new upper bounds on the metric entropies which are based on the margin Natarajan

dimension. Their comparison with the bounds of reference, for instance Inequalities (5) and (6), requires two generic formulas. The first one is an upper bound on the γ-dimension of the class $\mathcal{G}_0$. The second one is the corresponding structural result for the margin Natarajan dimension of $\rho_{\mathcal{G}}$.

7.1 Bounds on the Metric Entropies

For the first formula, we use the standard hypothesis: that of polynomial γ-dimensions (van der Vaart and Wellner, 1996; Mendelson, 2003). We have already seen that it is satisfied by SVMs (Formula (18)). This is also the case for MLPs with linear output units (see for instance Theorem 14.19 in Anthony and Bartlett, 1999). The second formula is designed to incorporate the hypothesis of polynomial γ-dimensions in a decomposition result taking benefit from the coupling between the component functions of the functions g. It is thus primarily inspired by Inequalities (17) and (20).

Hypothesis 1. *We consider function classes $\mathcal{G}$ defined as in Definition 1 for which there exists a quadruplet $\left(d_{\mathcal{G},C}, d_{\mathcal{G},\gamma}, K_{\mathcal{G}_0}, K_{\rho_{\mathcal{G}}}\right) \in (0,2] \times \left(\mathbb{R}_+^*\right)^3$ such that*

$$\forall \epsilon \in (0, M_{\mathcal{G}}], \begin{cases} \epsilon\text{-}dim\left(\mathcal{G}_0\right) \leqslant K_{\mathcal{G}_0} \epsilon^{-d_{\mathcal{G},\gamma}} & \text{(21a)} \\ \epsilon\text{-}N\text{-}dim\left(\rho_{\mathcal{G}}\right) \leqslant K_{\rho_{\mathcal{G}}} C^{d_{\mathcal{G},C}} \epsilon^{-d_{\mathcal{G},\gamma}}. & \text{(21b)} \end{cases}$$

Under Hypothesis 1, the combinatorial results dedicated to the margin Natarajan dimension (Lemmas 8 and 9) give birth to the following bounds on the metric entropies.

Theorem 1. *Let $\mathcal{G}$ be a function class satisfying Hypothesis 1. For every $\gamma \in (0,1]$, $\epsilon \in (0,\gamma]$ and $n \in \mathbb{N}^*$ such that $n \geqslant K_{\rho_{\mathcal{G}}} C^{d_{\mathcal{G},C}} \left(\frac{4}{\epsilon}\right)^{d_{\mathcal{G},\gamma}}$,*

$$\log_2\left(\mathcal{N}_\infty^{int}\left(\epsilon, \rho_{\mathcal{G},\gamma}, n\right)\right) \leqslant K_{\rho_{\mathcal{G}}} C^{d_{\mathcal{G},C}} \log_2^2\left(\frac{6\gamma\left(C-1\right)n}{\epsilon}\right)\left(\frac{4}{\epsilon}\right)^{d_{\mathcal{G},\gamma}}. \tag{22}$$

For every $\gamma \in (0,1]$, $\epsilon \in (0,\gamma]$ and $n \in \mathbb{N}^*$,

$$\log_2\left(\mathcal{N}_2^{int}\left(\epsilon, \rho_{\mathcal{G},\gamma}, n\right)\right) \leqslant \frac{3}{2} K_{\rho_{\mathcal{G}}} C^{d_{\mathcal{G},C}} \log_2^2\left(\left(C-1\right)\left(\frac{4\gamma}{\epsilon}\right)^5\right)\left(\frac{28}{\epsilon}\right)^{d_{\mathcal{G},\gamma}}. \tag{23}$$

As expected, those two bounds are significantly better than the bounds obtained with the structural result dedicated to the fat-shattering dimension of $\rho_{\mathcal{G}}$: Lemma 3. Thus, a partial answer to Question 1 emerges: the margin Natarajan dimension is the first scale-sensitive combinatorial dimension which can be considered to handle the class of margin functions $\rho_{\mathcal{G}}$ for a vast family of function classes $\mathcal{G}$. The remaining comparison to be done is with the bounds using the structural result involving covering numbers (Lemma 2), i.e.,

Inequalities (5) and (6). The rest of the section is devoted to this comparison. To make it more concrete, we use as touchstones the functions F_i corresponding to the state-of-the-art basic supremum inequalities associated with the two most popular margin loss functions.

7.2 Guaranteed Risk for the Indicator Margin Loss Function

The class of margin loss functions contains two main families: the family of indicator functions and the one of Lipschitz continuous functions. Our first guaranteed risk involves the classical indicator function: $\phi_{\infty,\gamma}$. It is given by:

$$\forall t \in \mathbb{R}, \quad \phi_{\infty,\gamma}(t) = \mathbb{1}_{\{t<\gamma\}}.$$

To the best of our knowledge, the sharpest instance of Inequality (1) involving this loss function is provided by Theorem 2 in Guermeur (2017). This bound corresponds to the case $L_* = L$ and produces:

$$F_i(m, \gamma, \delta, \mathrm{cap}(\rho_{\mathcal{G},\gamma})) = \sqrt{\frac{2}{m}\left(\ln\left(\mathcal{N}_\infty^{\mathrm{int}}\left(\frac{\gamma}{2}, \rho_{\mathcal{G},\gamma}, 2m\right)\right) + \ln\left(\frac{2}{\delta}\right)\right)} + \frac{1}{m}. \tag{24}$$

By application of (22) (for $m \geqslant \frac{1}{2} K_{\rho_{\mathcal{G}}} C^{d_{\mathcal{G},C}} \left(\frac{8}{\gamma}\right)^{d_{\mathcal{G},\gamma}}$):

$$\log_2\left(\mathcal{N}_\infty^{\mathrm{int}}\left(\frac{\gamma}{2}, \rho_{\mathcal{G},\gamma}, 2m\right)\right) \leqslant K_{\rho_{\mathcal{G}}} C^{d_{\mathcal{G},C}} \log_2^2\left(24\,(C-1)\,m\right)\left(\frac{8}{\gamma}\right)^{d_{\mathcal{G},\gamma}}. \tag{25}$$

The function F_f obtained by substitution of (25) into (24) decreases with the sample size m as a $O\left(\frac{\ln(m)}{\sqrt{m}}\right)$. This convergence rate is that of the literature (for the margin loss function considered). The dependence on the number C of categories is a $O\left(C^{\frac{d_{\mathcal{G},C}}{2}} \ln(C)\right)$, implying that it is always sublinear except in the worst case $d_{\mathcal{G},C} = 2$. At last, the dependence on $\frac{1}{\gamma}$ is a $O\left(\left(\frac{1}{\gamma}\right)^{\frac{d_{\mathcal{G},\gamma}}{2}}\right)$. Even though the decrease of F_f with γ was expected, it calls for an explanation, since for a fixed value of the scale parameter ϵ, the metric entropy and its upper bound (Formula (22)) increase with γ. The obvious reason is that F_i introduces a linear dependence of ϵ on γ ($\epsilon = \frac{\gamma}{2}$).

7.3 Guaranteed Risk for the Parameterized Truncated Hinge Loss

In the family of Lipschitz continuous margin loss functions, the option of choice is the parameterized truncated hinge loss $\phi_{2,\gamma}$ given by:

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

For this loss function, the best instance of Inequality (1) is provided by Theorem 5 in Guermeur (2017). It corresponds to $L_* = L_\gamma$ and the analytical expression of function F_i is:

$$F_i\left(m, \gamma, \delta, \mathrm{cap}\left(\rho_{\mathcal{G},\gamma}\right)\right) = \frac{2}{\gamma} R_m\left(\rho_{\mathcal{G},\gamma}\right) + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2m}}. \tag{26}$$

Formula (26) calls for a discussion. On the one hand, a non trivial expression for the function F_f should decrease with γ. On the other hand, the Rademacher complexity is increasing with this parameter. This implies that we are looking for an upper bound on $R_m\left(\rho_{\mathcal{G},\gamma}\right)$ increasing at most linearly with γ. In accordance with both graphs of transitions (Figures 1 and 2), this capacity measure is upper bounded as a function of the L_2-norm metric entropy by means of Dudley's chaining method. We use the following formula, whose degrees of freedom can be exploited to optimize the dependence on the basic parameters.

Theorem 2 (Theorem 9 in Guermeur, 2017)**.** *Let $\mathcal{F}$ be a class of bounded real-valued functions on $\mathcal{T}$. For $n \in \mathbb{N}^*$, let $\mathbf{t}_n \in \mathcal{T}^n$ and let $diam\left(\mathcal{F}\right)$ be the diameter of $\mathcal{F}$ with respect to the pseudo-metric $d_{2,\mathbf{t}_n}$. Let h be a positive and decreasing function on $\mathbb{N}$ such that $h\left(0\right) \geqslant diam\left(\mathcal{F}\right)$. Then for $N \in \mathbb{N}^*$,*

$$\hat{R}_n\left(\mathcal{F}\right) \leqslant h\left(N\right) + 2\sum_{j=1}^{N}\left(h\left(j\right) + h\left(j-1\right)\right)\sqrt{\frac{\ln\left(\mathcal{N}^{int}\left(h\left(j\right), \mathcal{F}, d_{2,\mathbf{t}_n}\right)\right)}{n}}. \tag{27}$$

A substitution of Inequality (23) into (27) gives:

$$R_m\left(\rho_{\mathcal{G},\gamma}\right) \leqslant h\left(N\right) + \frac{8}{3}\sqrt{\frac{F_1\left(C\right)}{m}}\sum_{j\in\mathcal{J}}\frac{h\left(j\right) + h\left(j-1\right)}{h\left(j\right)^{\frac{d_{\mathcal{G},\gamma}}{2}}}\log_2\left(\left(C-1\right)\left(\frac{4\gamma}{h\left(j\right)}\right)^5\right) \tag{28}$$

where

$$F_1\left(C\right) = 28^{d_{\mathcal{G},\gamma}} K_{\rho_{\mathcal{G}}} C^{d_{\mathcal{G},C}}, \tag{29}$$

with $\mathcal{J} = \left\{j \in [\![1; N]\!] : \; h\left(j\right) \leqslant \gamma\right\}$. With the last formula at hand, the derivation of the confidence interval amounts to studying the phase transitions highlighted by Theorem 18 in Mendelson (2003).

Theorem 3. *Let $\mathcal{G}$ be a function class satisfying Hypothesis 1. The following statements hold true for every value of γ in $(0, 1]$.*
If $d_{\mathcal{G},\gamma} \in (0, 2)$, then

$$R_m\left(\rho_{\mathcal{G},\gamma}\right) \leqslant 12\left(1 + 2^{\frac{2}{2-d_{\mathcal{G},\gamma}}}\right)\sqrt{\frac{F_1\left(C\right)}{m}} F_2\left(C\right)\gamma^{1-\frac{d_{\mathcal{G},\gamma}}{2}},$$

where $F_1\left(C\right)$ is given by Equation (29) and $F_2\left(C\right) = \ln\left(\left(C-1\right)4^5\right) + 10\frac{1+\ln(2)}{2-d_{\mathcal{G},\gamma}}$.

If $d_{\mathcal{G},\gamma} = 2$ *and* $m \geqslant 2$*, then*

$$R_m(\rho_{\mathcal{G},\gamma}) \leqslant \gamma \frac{\log_2(m)}{\sqrt{m}} + 8\sqrt{\frac{F_1(C)}{m}} \left\lceil \log_2\left(\frac{\sqrt{m}}{\log_2(m)}\right)\right\rceil \log_2\left((C-1)\left(4\frac{\sqrt{m}}{\log_2(m)}\right)^5\right).$$

At last, if $d_{\mathcal{G},\gamma} > 2$ *and* $m \geqslant 2$*, then*

$$R_m(\rho_{\mathcal{G},\gamma}) \leqslant \gamma \left(\frac{\log_2(m)}{m}\right)^{\frac{1}{d_{\mathcal{G},\gamma}}}$$

$$\times \left[1 + \frac{16}{3}\left(1 + 2^{\frac{2}{d_{\mathcal{G},\gamma}-2}}\right)\left(\frac{1}{\gamma}\right)^{\frac{d_{\mathcal{G},\gamma}}{2}} \sqrt{\frac{F_1(C)}{\log_2(m)}} \log_2\left((C-1)\left(4\left(\frac{m}{\log_2(m)}\right)^{\frac{1}{d_{\mathcal{G},\gamma}}}\right)^5\right)\right].$$

The dependence of F_f on m is that of Mendelson's formulas, except for a multiplicative factor $\sqrt{\ln(m)}$ in the case when $d_{\mathcal{G},\gamma} \geqslant 2$ (complex classifiers). The dependences on C and $\frac{1}{\gamma}$ are the same as with the indicator margin loss function.

7.4 Comparison with the Use of the State-of-the-Art Structural Results

Compared to the use of the structural result involving Rademacher complexities (Lemma 1), the obvious advantage of our approach is to allow to exploit a possible coupling between the component functions of the classifier to produce a sublinear dependence of F_f on the number C of categories. The comparison with the decomposition involving covering numbers (Lemma 2) produces a mixed result. On the one hand, whatever the choice of the margin loss function, the dependence of F_f on the inverse of the margin parameter γ is improved. The gain is a factor $\ln\left(\frac{1}{\gamma}\right)$ with the indicator margin loss function and a factor $\sqrt{\ln\left(\frac{1}{\gamma}\right)}$ with the parameterized truncated hinge loss. The only prize to pay occurs for this latter margin loss function, when $d_{\mathcal{G},\gamma} \geqslant 2$ (complex classifiers). Then, the dependence on the sample size m increases by a factor $\sqrt{\ln(m)}$.

8 Conclusions

We have established that the combinatorial results involving the fat-shattering dimension of the class of margin functions $\rho_{\mathcal{G}}$ can always be improved by replacing this dimension with the margin Graph dimension of the same class (Lemmas 4, 6 and 7). Currently, the gain is limited by the lack of a structural result specific to the margin Graph dimension (a structural result that would be significantly better than Lemma 3). Fortunately, the use of another γ-Ψ-dimension, the margin Natarajan dimension, makes it possible to exploit basic features of the classifier of interest (Corollary 1 and Lemma 11) to derive useful structural

results. The major consequence is an improved dependence of the confidence interval of the guaranteed risk on the margin parameter γ. This holds true both with the L_∞-norm (Inequality (25)) and the L_2-norm (Theorem 3). Except in the worst case $d_{\mathcal{G},C} = 2$, the dependence on the number C of categories is sublinear. The only drawback is that the convergence rate of the guaranteed risk associated with the parameterized truncated hinge loss can be worsened by a factor $\sqrt{\ln(m)}$ when the class $\mathcal{G}_0$ is complex (large value of $d_{\mathcal{G},\gamma}$). The phenomenon is a direct consequence of the appearance of the logarithmic function of $\frac{\gamma}{\epsilon}$ in the exponent of the right-hand side of Inequality (14) (compared to Formula 2 in Mendelson and Vershynin, 2003). Whether this term can be replaced with a logarithmic function of C only is an open question which is the subject of an ongoing research.

Acknowledgements The author would like to thank R. Vershynin for his explanations on the proof of Theorem 1 in Mendelson and Vershynin (2003). Thanks are also due to F. Lauer and T. Masini for carefully reading this manuscript. This work was partly funded by a CNRS research grant (PEPS).

References

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

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

P.L. 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):525–536, 1998.

P.L. Bartlett and P.M. Long. Prediction, learning, uniform convergence, and scale-sensitive dimensions. *Journal of Computer and System Sciences*, 56(2):174–190, 1998.

P.L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

P.L. Bartlett and J. Shawe-Taylor. Generalization performance of support vector machines and other pattern classifiers. In B. Schölkopf, C.J.C. Burges, and A. Smola, editors,

Advances in Kernel Methods - Support Vector Learning, chapter 4, pages 43–54. The MIT Press, Cambridge, MA, 1999.

S. Ben-David, N. Cesa-Bianchi, D. Haussler, and P.M. Long. Characterizations of learnability for classes of $\{0, \ldots, n\}$-valued functions. *Journal of Computer and System Sciences*, 50(1):74–86, 1995.

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

U. Doğan, T. Glasmachers, and C. Igel. A unified view on multi-class support vector classification. *Journal of Machine Learning Research*, 17(45):1–32, 2016.

H.H. Duan. Bounding the fat shattering dimension of a composition function class built using a continuous logic connective. *The Waterloo Mathematics Review*, 2(1):1–21, 2012.

R.M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

R.M. Dudley, E. Giné, and J. Zinn. Uniform and universal Glivenko-Cantelli classes. *Journal of Theoretical Probability*, 4(3):485–510, 1991.

Y. Guermeur. VC theory of large margin multi-category classifiers. *Journal of Machine Learning Research*, 8:2551–2594, 2007.

Y. Guermeur. L_p-norm Sauer-Shelah lemma for margin multi-category classifiers. *Journal of Computer and System Sciences*, 89:450–473, 2017.

M.J. Kearns and R.E. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3):464–497, 1994.

A.N. Kolmogorov and V.M. Tihomirov. ϵ-entropy and ϵ-capacity of sets in functional spaces. *American Mathematical Society Translations, series 2*, 17:277–364, 1961.

A. Kontorovich and R. Weiss. Maximum margin multiclass nearest neighbors. In *ICML'14*, 2014.

Y. Lei, U. Doğan, A. Binder, and M. Kloft. Multi-class SVMs: From tighter data-dependent generalization bounds to novel algorithms. In *NIPS 28*, pages 2026–2034, 2015.

A. Maurer. A vector-contraction inequality for Rademacher complexities. In *ALT'16*, pages 3–17, 2016.

S. Mendelson. A few notes on statistical learning theory. In S. Mendelson and A.J. Smola, editors, *Advanced Lectures on Machine Learning*, chapter 1, pages 1–40. Springer-Verlag, Berlin, Heidelberg, New York, 2003.

S. Mendelson and R. Vershynin. Entropy and the combinatorial dimension. *Inventiones mathematicae*, 152:37–55, 2003.

M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. The MIT Press, Cambridge, MA, second edition, 2018.

K. Musayeva, F. Lauer, and Y. Guermeur. Rademacher complexity and generalization performance of multi-category margin classifiers. *Neurocomputing*, 342:6–15, 2019.

D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, New York, 1984.

M. Rudelson and R. Vershynin. Combinatorics of random processes and sections of convex bodies. *Annals of Mathematics*, 164(2):603–648, 2006.

M. Talagrand. Vapnik-Chervonenkis type conditions and uniform Donsker classes of functions. *The Annals of Probability*, 31(3):1565–1582, 2003.

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

A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes, With Applications to Statistics*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

V.N. Vapnik and A.Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, XVI(2):264–280, 1971.

G. Wahba. Multivariate function and operator estimation, based on smoothing splines and reproducing kernels. In M. Casdagli and S. Eubank, editors, *Nonlinear Modeling and Forecasting, SFI Studies in the Sciences of Complexity*, volume XII, pages 95–112. Addison-Wesley, 1992.

A Proofs of the Connections Between the Scale-Sensitive Combinatorial Dimensions

The proofs in this appendix and the following ones make use of a standard convention: a function class $\mathcal{F}$ is said to γ-N-shatter a triplet $(s_{\mathcal{Z}^n}, \mathbf{b}_n, \mathbf{c}_n)$ if $\mathcal{F}$ γ-N-shatters $s_{\mathcal{Z}^n}$ and $(\mathbf{b}_n, \mathbf{c}_n)$ is a witness to this shattering. The corresponding convention for the margin Graph dimension is also used. The proof of Lemma 4 is the following one.

Proof. Let f be any function in $\mathcal{F}$ and $z = (x, y) \in \mathcal{Z}$. Let $\gamma \in (0, 1]$, $\epsilon \in \left(0, \frac{\gamma}{2}\right]$ and $b \in [\epsilon, \gamma - \epsilon]$. Then,

$$f_\gamma(z) - b \geqslant \epsilon \Longrightarrow f_\gamma(z) \geqslant 2\epsilon > 0 \Longrightarrow f(z) \geqslant f_\gamma(z).$$

Consequently,

$$f_\gamma(z) - b \geqslant \epsilon \Longrightarrow f(z) - b \geqslant \epsilon. \tag{30}$$

Suppose that $f_\gamma(z) = 0$ (implying that $f(z) \leqslant 0$). Then it results from Equation (2) that

$$-f_\gamma(z) \leqslant \max_{k \neq y} f(x, k) \leqslant -f(z).$$

Suppose now that $f_\gamma(z) > 0$. Then,

$$-f_\gamma(z) + b \geqslant \epsilon \Longrightarrow f_\gamma(z) \leqslant \gamma - 2\epsilon < \gamma.$$

Thus, $f_\gamma(z)$ belongs to the open interval $(0, \gamma)$, so that $f(z) = f_\gamma(z)$. Consequently,

$$-f_\gamma(z) = \max_{k \neq y} f(x, k) = -f(z).$$

To sum up, for all possible values of $f_\gamma(z)$ (positive or null),

$$-f_\gamma(z) + b \geqslant \epsilon \Longrightarrow \max_{k \neq y} f(x, k) + b \geqslant \epsilon \Longrightarrow -f(z) + b \geqslant \epsilon. \tag{31}$$

Formula (9) directly springs from Formulas (30) and (31). □

The proof of Lemma 5 is the following one.

Proof. The left-hand side inequality of Formula (10) is obvious. Let us set $d_G = \gamma\text{-G-dim}(\mathcal{F})$ and $d_N = \gamma\text{-N-dim}(\mathcal{F})$. The right-hand side inequality of Formula (10) is trivially true for $d_G = 0$. Thus, we prove it under the assumption that $d_G \geqslant 1$. Let $\tilde{\mathcal{F}}$ be any subset of $\mathcal{F}$ of cardinality 2^{d_G} that γ-G-shatters a subset $s_{\mathcal{Z}^{d_G}}$ of $\mathcal{Z}$ of cardinality d_G. The proof rests on the derivation of a separating tree of $\tilde{\mathcal{F}}$. For notational simplicity, we set

$s_{\mathcal{Z}^{d_G}} = \{z_i : 1 \leqslant i \leqslant d_G\}$. A subset of $s_{\mathcal{Z}^{d_G}}$ of cardinality $n \in [\![1; d_N]\!]$ is denoted by $s'_{\mathcal{Z}^n} = \{z'_i : \ 1 \leqslant i \leqslant n\}$, with the convention

$$\forall\, (i,j) : 1 \leqslant i < j \leqslant n,\ \left(z'_i, z'_j\right) = (z_v, z_w) \Longrightarrow 1 \leqslant v < w \leqslant d_G.$$

Let the vector $\mathbf{b}_{d_G} = (b_i)_{1 \leqslant i \leqslant d_G} \in \mathbb{R}_+^{d_G}$ be a witness to the γ-G-shattering of $s_{\mathcal{Z}^{d_G}}$ by $\tilde{\mathcal{F}}$ and let $(s'_{\mathcal{Z}^n}, \mathbf{b}'_n, \mathbf{c}'_n)$ be any triplet such that

$$\begin{cases} s'_{\mathcal{Z}^n} \subset s_{\mathcal{Z}^{d_G}} \\ \mathbf{b}'_n \in \mathbb{R}_+^n : \ \forall i \in [\![1; n]\!]\,,\ z'_i = z_j \Longrightarrow b'_i = b_j \\ \mathbf{c}'_n \in \mathcal{Y}^n : \ \forall i \in [\![1; n]\!]\,,\ c'_i \in \mathcal{Y} \setminus \{y'_i\} \end{cases}.$$

For every subset $\bar{\mathcal{F}}$ of $\tilde{\mathcal{F}}$, let $s\left(\bar{\mathcal{F}}\right)$ denote the number of triplets $(s'_{\mathcal{Z}^n}, \mathbf{b}'_n, \mathbf{c}'_n)$ γ-N-shattered by $\bar{\mathcal{F}}$. Combinatorics produces

$$s\left(\tilde{\mathcal{F}}\right) \leqslant \sum_{n=1}^{d_N} \binom{d_G}{n} (C-1)^n\,,$$

which gives birth to a handy formula thanks to a well-known computation (see for instance the proof of Corollary 3.18 in Mohri et al., 2018):

$$s\left(\tilde{\mathcal{F}}\right) \leqslant \left(\frac{(C-1)\, e d_G}{d_N}\right)^{d_N} - 1. \tag{32}$$

The derivation of the separating tree of $\tilde{\mathcal{F}}$ provides us with a lower bound on $s\left(\tilde{\mathcal{F}}\right)$. Let $\bar{\mathcal{F}}$ be any of its nodes such that $\left|\bar{\mathcal{F}}\right| \geqslant 2$ (inner node). Its two sons, $\bar{\mathcal{F}}_+$ and $\bar{\mathcal{F}}_-$, are built as follows. Split $\bar{\mathcal{F}}$ arbitrarily into $\left\lfloor \frac{|\bar{\mathcal{F}}|}{2} \right\rfloor$ pairs (with possibly a function remaining alone). For each pair (f, f'), find $z_i \in s_{\mathcal{Z}^{d_G}}$ such that

$$\begin{cases} f\left(z_i\right) - b_i \geqslant \gamma \\ \max_{k \neq y_i} f'\left(x_i, k\right) + b_i \geqslant \gamma \end{cases}$$

or vice versa. By the pigeonhole principle, the same example is picked for at least $\left\lceil \left\lfloor \frac{|\bar{\mathcal{F}}|}{2} \right\rfloor \frac{1}{d_G} \right\rceil$ pairs. Let z_{i_0} be such an example, and let (f_+, f_-) denote the corresponding pairs, whose components are reordered (when needed) so that

$$\begin{cases} f_+\left(z_{i_0}\right) - b_{i_0} \geqslant \gamma \\ \max_{k \neq y_{i_0}} f_-\left(x_{i_0}, k\right) + b_{i_0} \geqslant \gamma \end{cases}.$$

$\bar{\mathcal{F}}_+$ is the set of the functions f_+. Once more by the pigeonhole principle, there exists a value $c_{i_0} \in \mathcal{Y} \setminus \{y_{i_0}\}$ such that at least $\left\lceil \left\lceil \left\lfloor \frac{|\bar{\mathcal{F}}|}{2} \right\rfloor \frac{1}{d_G} \right\rceil \frac{1}{C-1} \right\rceil$ functions f_- satisfy

$$f_-\left(x_{i_0}, c_{i_0}\right) + b_{i_0} \geqslant \gamma.$$

Let $\bar{\mathcal{F}}_-$ be the set of these functions f_-. The two sons $\bar{\mathcal{F}}_+$ and $\bar{\mathcal{F}}_-$ of $\bar{\mathcal{F}}$ have been built in such a way that $\left|\bar{\mathcal{F}}_+\right| \geqslant \frac{|\bar{\mathcal{F}}|}{3d_G}$, $\left|\bar{\mathcal{F}}_-\right| \geqslant \frac{|\bar{\mathcal{F}}|}{3d_G(C-1)}$ and

$$\begin{cases} \forall f_+ \in \bar{\mathcal{F}}_+, \;\; f_+\left(z_{i_0}\right) - b_{i_0} \geqslant \gamma \\ \forall f_- \in \bar{\mathcal{F}}_-, \;\; f_-\left(x_{i_0}, c_{i_0}\right) + b_{i_0} \geqslant \gamma \end{cases}. \tag{33}$$

Since $\bar{\mathcal{F}}_+ \bigcup \bar{\mathcal{F}}_- \subset \bar{\mathcal{F}}$, any triplet γ-N-shattered by either $\bar{\mathcal{F}}_+$ or $\bar{\mathcal{F}}_-$ is also γ-N-shattered by $\bar{\mathcal{F}}$. Furthermore, according to (33), if the triplet $(s'_{\mathcal{Z}^n}, \mathbf{b}'_n, \mathbf{c}'_n)$ is γ-N-shattered by both $\bar{\mathcal{F}}_+$ and $\bar{\mathcal{F}}_-$, then $\bar{\mathcal{F}}$ also γ-N-shatters the triplet $\left(s''_{\mathcal{Z}^{n+1}}, \mathbf{b}''_{n+1}, \mathbf{c}''_{n+1}\right)$ such that $s''_{\mathcal{Z}^{n+1}} = s'_{\mathcal{Z}^n} \bigcup \{z_{i_0}\}$, the vector $\mathbf{b}''_{n+1}$ is deduced from $\mathbf{b}'_n$ by inserting the component b_{i_0}, and the vector $\mathbf{c}''_{n+1}$ is deduced from $\mathbf{c}'_n$ by inserting the component c_{i_0}. Clearly, neither $\bar{\mathcal{F}}_+$ nor $\bar{\mathcal{F}}_-$ γ-N-shatters the triplet $\left(s''_{\mathcal{Z}^{n+1}}, \mathbf{b}''_{n+1}, \mathbf{c}''_{n+1}\right)$, simply because (contrary to $\bar{\mathcal{F}}$) they do not γ-N-shatter the triplet $\left(\{z_{i_0}\}, b_{i_0}, c_{i_0}\right)$. A synthesis of the different cases produces:

$$s\left(\bar{\mathcal{F}}\right) \geqslant s\left(\bar{\mathcal{F}}_+\right) + s\left(\bar{\mathcal{F}}_-\right) + 1. \tag{34}$$

To further bound frow below $s\left(\bar{\mathcal{F}}\right)$, we introduce the function ℓ which returns the number of leaves of the (sub)tree whose root is its argument. Using (34), the following simple connection between the two functions can be proved by induction:

$$s\left(\bar{\mathcal{F}}\right) \geqslant \ell\left(\bar{\mathcal{F}}\right) - 1. \tag{35}$$

Once more by induction, we now establish that any node $\bar{\mathcal{F}}$ (even a leaf) satisfies:

$$\ell\left(\bar{\mathcal{F}}\right) \geqslant \left|\bar{\mathcal{F}}\right|^{\frac{1}{\log_2\left(3\sqrt{C-1}d_G\right)}}. \tag{36}$$

Inequality (36) is obviously true for the leafs (for which $\ell\left(\bar{\mathcal{F}}\right) = \left|\bar{\mathcal{F}}\right| = 1$). Suppose now that it holds true for the two sons $\bar{\mathcal{F}}_+$ and $\bar{\mathcal{F}}_-$ of the inner node $\bar{\mathcal{F}}$. Then,

$$
\begin{aligned}
\ell\left(\bar{\mathcal{F}}\right) = & \ \ell\left(\bar{\mathcal{F}}_{+}\right) + \ell\left(\bar{\mathcal{F}}_{-}\right) \\
\geqslant & \ \left(\frac{\left|\bar{\mathcal{F}}\right|}{3d_G}\right)^{\frac{1}{\log_2\left(3\sqrt{C-1}d_G\right)}} + \left(\frac{\left|\bar{\mathcal{F}}\right|}{3d_G\left(C-1\right)}\right)^{\frac{1}{\log_2\left(3\sqrt{C-1}d_G\right)}} \\
= & \ \frac{1}{2}\left(\left(\sqrt{C-1}\right)^{\frac{1}{\log_2\left(3\sqrt{C-1}d_G\right)}} + \left(\sqrt{C-1}\right)^{-\frac{1}{\log_2\left(3\sqrt{C-1}d_G\right)}}\right)\left|\bar{\mathcal{F}}\right|^{\frac{1}{\log_2\left(3\sqrt{C-1}d_G\right)}} \\
\geqslant & \ \frac{1}{2}\min_{t\in\mathbb{R}_{+}^{*}}\left(t+\frac{1}{t}\right)\left|\bar{\mathcal{F}}\right|^{\frac{1}{\log_2\left(3\sqrt{C-1}d_G\right)}} \\
= & \ \left|\bar{\mathcal{F}}\right|^{\frac{1}{\log_2\left(3\sqrt{C-1}d_G\right)}}.
\end{aligned}
$$

Since $\left|\tilde{\mathcal{F}}\right| = 2^{d_G}$, we thus get for the whole tree:

$$
\ell\left(\tilde{\mathcal{F}}\right) \geqslant 2^{\frac{d_G}{\log_2\left(3\sqrt{C-1}d_G\right)}}. \tag{37}
$$

A substitution of (37) into (35) provides the lower bound on $s\left(\tilde{\mathcal{F}}\right)$ announced:

$$
s\left(\tilde{\mathcal{F}}\right) \geqslant 2^{\frac{d_G}{\log_2\left(3\sqrt{C-1}d_G\right)}} - 1. \tag{38}
$$

Combining (38) and the upper bound, (32), gives by transitivity:

$$
\begin{aligned}
d_G \leqslant & \ d_N \log_2\left(\frac{\left(C-1\right)ed_G}{d_N}\right)\log_2\left(3\sqrt{C-1}d_G\right) \\
\leqslant & \ \frac{1}{\ln^2\left(2\right)}d_N \ln\left(F\left(C\right)\frac{d_G}{d_N}\right)\ln\left(F\left(C\right)d_G\right),
\end{aligned} \tag{39}
$$

where $F\left(C\right) = e\left(C-1\right)$. To bound from above the right-hand side of Inequality (39), we resort to the following statement:

$$
\forall\left(u,u_0\right)\in\left[1,+\infty\right)^2,\ \ln\left(u\right) \leqslant 2u_0 u^{\frac{1}{4u_0}}, \tag{40}
$$

with $u_0 = \ln\left(F\left(C\right)\right)$. We then obtain

$$
\begin{cases}
\ln\left(F\left(C\right)\frac{d_G}{d_N}\right) \leqslant 2e^{\frac{1}{4}}\ln\left(F\left(C\right)\right)\left(\frac{d_G}{d_N}\right)^{\frac{1}{4\ln\left(F\left(C\right)\right)}} \\
\ln\left(F\left(C\right)d_G\right) \leqslant 2e^{\frac{1}{4}}\ln\left(F\left(C\right)\right)d_G^{\frac{1}{4\ln\left(F\left(C\right)\right)}}
\end{cases}.
$$

By substitution into (39),

$$
\begin{aligned}
d_G &\leqslant \frac{4\sqrt{e}}{\ln^2(2)} \ln^2(F(C))\, d_G^{\frac{1}{2\ln(F(C))}} d_N^{\frac{4\ln(F(C))-1}{4\ln(F(C))}} \\
&\leqslant \left(\frac{4\sqrt{e}}{\ln^2(2)}\right)^{\frac{2\ln(F(C))}{2\ln(F(C))-1}} (\ln(F(C)))^{\frac{4\ln(F(C))}{2\ln(F(C))-1}} d_N^{\frac{4\ln(F(C))-1}{4\ln(F(C))-2}} \\
&= \left(\frac{4\sqrt{e}}{\ln^2(2)}\right)^{\frac{2\ln(F(C))}{2\ln(F(C))-1}} (\ln(F(C)))^{\frac{2}{2\ln(F(C))-1}} \ln^2(2) \log_2^2(F(C))\, d_N^{\frac{4\ln(F(C))-1}{4\ln(F(C))-2}} \\
&< 32 \log_2^2(F(C))\, d_N^{1+\frac{1}{4\ln(F(C))-2}}.
\end{aligned}
$$

□

B Proofs of the Combinatorial Results

This appendix gathers the proofs of the four new combinatorial results. It starts with three lemmas which are common to all proofs.

B.1 Shared Technical Lemmas

Each of the combinatorial results in the literature is built upon a basic lemma that involves two (possibly identical) function classes whose codomains are discrete. The domain and codomain of the first one are finite sets, so that its cardinality is also finite. This cardinality is upper bounded in terms of a combinatorial dimension of the second function class. In the case of margin classifiers, the combinatorial dimension of the basic lemma is a variant of the scale-sensitive dimension of the combinatorial result, variant designed to take benefit from the aforementioned restrictions. The first capacity measure of this kind is a variant of the γ-dimension: the strong dimension (Definition 3.1 in Alon et al., 1997). The strong Ψ-dimensions extend the γ-Ψ-dimensions according to the same principle.

Definition 12 (Strong Ψ-dimensions)**.** *Let $\mathcal{F}$ be a function class defined as in Definition 8. Suppose further that the functions in $\mathcal{F}$ take their values in $\mathbb{Z}$. Let Ψ be a family of mappings from $\mathcal{Y}$ into $\{-1,0,1\}$. A subset $s_{\mathcal{Z}^n} = \{z_i = (x_i, y_i) : 1 \leqslant i \leqslant n\}$ of $\mathcal{Z}$ is said to be* strongly Ψ-shattered *by $\mathcal{F}$ if there is a vector $\boldsymbol{\psi}_n = \left(\psi^{(i)}\right)_{1\leqslant i\leqslant n} \in \Psi^n$ satisfying $\left(\psi^{(i)}(y_i)\right)_{1\leqslant i\leqslant n} = \mathbf{1}_n$, and a vector $\mathbf{b}_n = (b_i)_{1\leqslant i\leqslant n} \in \mathbb{N}^n$ 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*

$$
\forall i \in [\![1;n]\!],\ \ s_i \left(s_i \max_{\left\{k:\, \psi^{(i)}(k) = s_i\right\}} f_{\mathbf{s}_n}(x_i, k) - b_i \right) \geqslant 1.
$$

The strong Ψ-dimension *of $\mathcal{F}$, denoted by S-Ψ-$dim(\mathcal{F})$, is the maximal cardinality of a subset of $\mathcal{Z}$ strongly Ψ-shattered by $\mathcal{F}$, if such maximum exists. Otherwise, $\mathcal{F}$ is said to have infinite strong Ψ-dimension.*

In what follows, the finiteness of the domain is simply obtained by application of a restriction to (projection on) an appropriately chosen set of data points. The discretization of the codomain results from the application of the following operator.

Definition 13 (η-discretization operator, Definition 33 in Guermeur, 2007)**.** *Let $\mathcal{F} \subset \mathbb{R}^{\mathcal{T}}$. For $\eta \in \mathbb{R}_+^*$, define the* η-discretization *as an operator on $\mathcal{F}$ such that:*

$$\begin{array}{rccc} (\cdot)^{(\eta)}: & \mathcal{F} & \longrightarrow & \mathcal{F}^{(\eta)} \\ & f & \mapsto & f^{(\eta)} \end{array}$$

$$\forall t \in \mathcal{T}, \quad f^{(\eta)}(t) = \operatorname{sign}(f(t)) \cdot \left\lfloor \frac{|f(t)|}{\eta} \right\rfloor.$$

The transitions from continuous functions to discrete ones and back are obtained by application of the two following lemmas.

Lemma 12. *Let $\mathcal{F}$ be a class of functions on $\mathcal{T}$ taking nonnegative values ($\mathcal{F} \subset \mathbb{R}_+^{\mathcal{T}}$). For $n \in \mathbb{N}^*$, let $\mathbf{t}_n = (t_i)_{1 \leqslant i \leqslant n} \in \mathcal{T}^n$. Let N be a positive integer. For every $\epsilon \in \mathbb{R}_+^*$ and every $\eta \in \left(0, \frac{\epsilon}{N+1}\right]$,*

$$\forall (f, f') \in \mathcal{F}^2, \quad d_{2,\mathbf{t}_n}(f, f') \geqslant \epsilon \Longrightarrow d_{2,\mathbf{t}_n}\left(f^{(\eta)}, f'^{(\eta)}\right) \geqslant N, \tag{41}$$

with the consequence that if the subset $\bar{\mathcal{F}}$ of $\mathcal{F}$ is ϵ-separated with respect to the pseudo-metric $d_{2,\mathbf{t}_n}$, then it is in bijection with the subset $\bar{\mathcal{F}}^{(\eta)}$ of $\mathcal{F}^{(\eta)}$, which is N-separated with respect to the same pseudo-metric. Similarly, for every $\epsilon \in \mathbb{R}_+^*$ and every $\eta \in \left(0, \frac{\epsilon}{2}\right]$,

$$\mathcal{M}(\epsilon, \mathcal{F}, d_{\infty,\mathbf{t}_n}) \leqslant \mathcal{M}\left(2, \mathcal{F}^{(\eta)}, d_{\infty,\mathbf{t}_n}\right). \tag{42}$$

Proof. For $f \in \mathcal{F}$ and $i \in [\![1; n]\!]$, let us denote the Euclidean division of $f(t_i)$ by η as follows:

$$\forall i \in [\![1; n]\!], \quad f(t_i) = \eta f^{(\eta)}(t_i) + r_i.$$

With the notation introduced above,

$$d_{2,\mathbf{t}_n}(f, f')^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\eta\left(f^{(\eta)}(t_i) - f'^{(\eta)}(t_i)\right) + r_i - r_i'\right)^2.$$

For $i \in [\![1;n]\!]$, let $\delta_i = \left| f^{(\eta)}(t_i) - f'^{(\eta)}(t_i) \right|$.

$$\begin{aligned}
\left(d_{2,\mathbf{t}_n}(f,f') \geqslant \epsilon\right) \text{ and } \left(\eta \in \left(0, \frac{\epsilon}{N+1}\right]\right) &\implies \left(\frac{1}{n}\sum_{i=1}^{n}\left(\eta\delta_i + \left|r_i - r'_i\right|\right)^2\right)^{\frac{1}{2}} \geqslant \epsilon \\
&\implies \left(\frac{1}{n}\sum_{i=1}^{n}\left(\delta_i+1\right)^2\right)^{\frac{1}{2}} \geqslant \frac{\epsilon}{\eta} \\
&\implies \left(\frac{1}{n}\sum_{i=1}^{n}\left(\delta_i+1\right)^2\right)^{\frac{1}{2}} \geqslant N+1 \qquad (43) \\
&\implies \left(\frac{1}{n}\sum_{i=1}^{n}\delta_i^2\right)^{\frac{1}{2}} + 1 \geqslant N+1 \qquad (44) \\
&\implies d_{2,\mathbf{t}_n}\left(f^{(\eta)}, f'^{(\eta)}\right) \geqslant N,
\end{aligned}$$

where the transition from (43) to (44) is provided by the triangle inequality. To sum up, we have established (41), i.e., the part of the lemma dealing with the L_2-norm. To prove (42), it is enough to observe that

$$\begin{aligned}
f(t) - f'(t) \geqslant \epsilon &\iff \frac{\epsilon}{2}\left(f^{\left(\frac{\epsilon}{2}\right)}(t) - f'^{\left(\frac{\epsilon}{2}\right)}(t)\right) + r - r' \geqslant \epsilon \\
&\implies \frac{\epsilon}{2}\left(f^{\left(\frac{\epsilon}{2}\right)}(t) - f'^{\left(\frac{\epsilon}{2}\right)}(t) + 1\right) > \epsilon \\
&\implies f^{\left(\frac{\epsilon}{2}\right)}(t) - f'^{\left(\frac{\epsilon}{2}\right)}(t) \geqslant 2.
\end{aligned}$$

□

Lemma 13. *Let $\mathcal{F}$ be a function class defined as in Definition 8. For every $\eta \in \mathbb{R}_+^*$ and every $\epsilon \in \left(0, \frac{\eta}{2}\right]$,*

$$S\text{-}\Psi\text{-}dim\left(\mathcal{F}^{(\eta)}\right) \leqslant \epsilon\text{-}\Psi\text{-}dim\left(\mathcal{F}\right). \qquad (45)$$

Proof. To prove (45), it is enough to notice that any set $s_{\mathcal{Z}^n}$ strongly Ψ-shattered by $\mathcal{F}^{(\eta)}$ according to the vector $\mathbf{b}_n = (b_i)_{1\leqslant i\leqslant n} \in \mathbb{N}^n$ is also $\frac{\eta}{2}$-Ψ-shattered by $\mathcal{F}$ according to $\mathbf{b}'_n = \left(\eta\left(b_i + \frac{1}{2}\right)\right)_{1\leqslant i\leqslant n} \in \mathbb{R}_+^n$. □

Lemma 14. *Let $\mathcal{F}$ be a function class defined as in Definition 8. Suppose that there exist $(f,f') \in \mathcal{F}^2$, $\gamma \in (0,1]$, $\eta \in \left(0, \frac{\gamma}{2}\right]$, and $z = (x,y) \in \mathcal{Z}$ such that*

$$f_\gamma^{(\eta)}(z) - f'^{(\eta)}_\gamma(z) \geqslant 2,$$

where $f_\gamma^{(\eta)} = (\pi_\gamma \circ f)^{(\eta)}$ and $f'^{(\eta)}_\gamma$ is defined accordingly. For every $b \in \left[\!\left[f'^{(\eta)}_\gamma(z) + 1; f_\gamma^{(\eta)}(z) - 1 \right]\!\right]$ and $c \in \operatorname{argmax}_{k\neq y} f'^{(\eta)}(x,k)$,

1. *the set* $\left\{f_\gamma^{(\eta)}, f_\gamma'^{(\eta)}\right\}$ *strongly shatters the pair* $(\{z\}, b)$*;*

2. *the set* $\left\{f^{(\eta)}, f'^{(\eta)}\right\}$ *strongly G-shatters the same pair;*

3. *the set* $\left\{f^{(\eta)}, f'^{(\eta)}\right\}$ *strongly N-shatters the triplets* $(\{z\}, b, c)$*.*

Proof. 1. The first assertion directly springs from the definition of the fat-shattering dimension.

2. The second assertion can be derived from the first one following the line of reasoning of the proof of Lemma 4.

3. The third assertion directly springs from the second one and the definition of the margin Natarajan dimension.

□

B.2 Margin Graph Dimension - Uniform Convergence Norm

The proof of Lemma 6 borrows from the proofs of classical results, including the two state-of-the-art combinatorial results: Lemma 3.5 in Alon et al. (1997) and Theorem 1 in Mendelson and Vershynin (2003). Central in this proof is the following basic combinatorial result.

Lemma 15. *Let* $\mathcal{F}$ *be a function class defined as in Definition 8. For every* $s_{\mathcal{Z}^n} = \{z_i = (x_i, y_i) : 1 \leqslant i \leqslant n\} \subset \mathcal{Z}$*,* $\gamma \in (0, 1]$ *and* $\eta \in \left(0, \frac{\gamma}{2}\right]$*, if* $\mathcal{F}_\gamma^{(\eta)}\Big|_{s_{\mathcal{Z}^n}}$ *is 2-separated in the metric* $d_{\infty,\mathbf{z}_n}$*, then*

$$\left|\mathcal{F}_\gamma^{(\eta)}\Big|_{s_{\mathcal{Z}^n}}\right| \leqslant (3M_\gamma n)^{\log_2(\Sigma)}, \tag{46}$$

where $\Sigma = \sum_{u=0}^{d_G} \binom{n}{u} M_\gamma^u$ *with* $M_\gamma = \left\lfloor \frac{\gamma}{\eta} \right\rfloor$ *and* $d_G = S\text{-}G\text{-}dim\left(\mathcal{F}^{(\eta)}\right)$*.*

Proof. Notice first that Inequality (46) is trivially true for $\left|\mathcal{F}_\gamma^{(\eta)}\Big|_{s_{\mathcal{Z}^n}}\right| = 1$. Indeed, the minimal value of its right-hand side, corresponding to $d_G = 0$, is 1. Thus, the rest of the proof makes use of the restriction $\left|\mathcal{F}_\gamma^{(\eta)}\Big|_{s_{\mathcal{Z}^n}}\right| \geqslant 2$. A direct consequence is that according to Lemma 14, $d_G \geqslant 1$. A subset of $s_{\mathcal{Z}^n}$ of cardinality $u \in [\![1; n]\!]$ is denoted by $s'_{\mathcal{Z}^u} = \{z'_i : 1 \leqslant i \leqslant u\}$, with the convention

$$\forall\, (i, j) : 1 \leqslant i < j \leqslant u,\ \left(z'_i, z'_j\right) = (z_v, z_w) \Longrightarrow 1 \leqslant v < w \leqslant n.$$

For every subset $\bar{\mathcal{F}}$ of $\mathcal{F}$, let $s\left(\bar{\mathcal{F}}^{(\eta)}\right)$ denote the number of pairs $(s'_{\mathcal{Z}^u}, \mathbf{b}'_u)$ with $s'_{\mathcal{Z}^u} \subset s_{\mathcal{Z}^n}$ and $\mathbf{b}'_u \in [\![1; M_\gamma - 1]\!]^u$ strongly G-shattered by $\bar{\mathcal{F}}^{(\eta)}$ (the convention above has been

introduced to avoid handling duplicates). Let $\tilde{\mathcal{F}}$ be any subset of $\mathcal{F}$ such that $\tilde{\mathcal{F}}_\gamma^{(\eta)}\Big|_{s_{\mathcal{Z}^n}} = \mathcal{F}_\gamma^{(\eta)}\Big|_{s_{\mathcal{Z}^n}}$ and $\left|\tilde{\mathcal{F}}\right| = \left|\mathcal{F}_\gamma^{(\eta)}\Big|_{s_{\mathcal{Z}^n}}\right|$ ($\tilde{\mathcal{F}}$ and $\mathcal{F}_\gamma^{(\eta)}\Big|_{s_{\mathcal{Z}^n}}$ are in bijection). Since $\tilde{\mathcal{F}}^{(\eta)} \subset \mathcal{F}^{(\eta)}$ and $d_G \geqslant 1$, combinatorics gives:

$$\begin{aligned} s\left(\tilde{\mathcal{F}}^{(\eta)}\right) &\leqslant s\left(\mathcal{F}^{(\eta)}\right) \\ &\leqslant \sum_{u=1}^{d_G} \binom{n}{u} M_\gamma^u = \Sigma - 1. \end{aligned} \tag{47}$$

In order to derive a lower bound on $s\left(\tilde{\mathcal{F}}^{(\eta)}\right)$, we build a 2-separating tree of $\tilde{\mathcal{F}}$, i.e., a 2-separating tree of $\mathcal{F}_\gamma^{(\eta)}\Big|_{s_{\mathcal{Z}^n}}$ (see Definition 3.4 in Rudelson and Vershynin, 2006). Let $\bar{\mathcal{F}}$ be one of its nodes such that $\left|\bar{\mathcal{F}}\right| \geqslant 2$ (inner node). Its two sons, $\bar{\mathcal{F}}_+$ and $\bar{\mathcal{F}}_-$, are built as follows. Split $\bar{\mathcal{F}}$ arbitrarily into $\left\lfloor \frac{\left|\bar{\mathcal{F}}\right|}{2} \right\rfloor$ pairs (with possibly a function remaining alone). For each pair (f, f'), find $z_i \in s_{\mathcal{Z}^n}$ such that $\left|f_\gamma^{(\eta)}(z_i) - f'^{(\eta)}_\gamma(z_i)\right| \geqslant 2$. By the pigeonhole principle, the same example is picked for at least $\left\lceil \left\lfloor \frac{\left|\bar{\mathcal{F}}\right|}{2} \right\rfloor \frac{1}{n} \right\rceil$ pairs. Let z_{i_0} be such an example, and let (f_+, f_-) denote the corresponding pairs, whose components are reordered (when needed) so that

$$f_{+,\gamma}^{(\eta)}(z_{i_0}) > f_{-,\gamma}^{(\eta)}(z_{i_0}).$$

Among the functions $f_{+,\gamma}^{(\eta)}$, at least $\left\lceil \left\lceil \left\lfloor \frac{\left|\bar{\mathcal{F}}\right|}{2} \right\rfloor \frac{1}{n} \right\rceil \frac{1}{M_\gamma - 1} \right\rceil$ take the same value at z_{i_0}. Let $v(z_{i_0})$ be such a value. We define $\bar{\mathcal{F}}_+$ (resp. $\bar{\mathcal{F}}_-$) to be the set of functions f_+ (resp. f_-) belonging to a pair associated with $(z_{i_0}, v(z_{i_0}))$. We obtain by construction:

$$\left|\bar{\mathcal{F}}_+\right| = \left|\bar{\mathcal{F}}_-\right| \geqslant \frac{\left|\bar{\mathcal{F}}\right|}{3M_\gamma n}. \tag{48}$$

Furthermore, according to Lemma 14, the sets $\bar{\mathcal{F}}_+^{(\eta)}$ and $\bar{\mathcal{F}}_-^{(\eta)}$ satisfy:

$$\begin{cases} \forall f_+ \in \bar{\mathcal{F}}_+, \;\; f_+^{(\eta)}(z_{i_0}) - b_{i_0} \geqslant 1 \\ \forall f_- \in \bar{\mathcal{F}}_-, \;\; \max_{k \neq y_{i_0}} f_-^{(\eta)}(z_{i_0}) + b_{i_0} \geqslant 1 \end{cases} \tag{49}$$

with $b_{i_0} = v(z_{i_0}) - 1$. Since $\bar{\mathcal{F}}_+^{(\eta)} \bigcup \bar{\mathcal{F}}_-^{(\eta)} \subset \bar{\mathcal{F}}^{(\eta)}$, obviously, any pair strongly G-shattered by either $\bar{\mathcal{F}}_+^{(\eta)}$ or $\bar{\mathcal{F}}_-^{(\eta)}$ is also strongly G-shattered by $\bar{\mathcal{F}}^{(\eta)}$. Furthermore, according to (49), $\bar{\mathcal{F}}^{(\eta)}$ strongly G-shatters the pair $(\{z_{i_0}\}, b_{i_0})$ which is strongly G-shattered by neither $\bar{\mathcal{F}}_+^{(\eta)}$ nor $\bar{\mathcal{F}}_-^{(\eta)}$. At last, let us consider any pair $(s'_{\mathcal{Z}^u}, \mathbf{b}'_u)$ strongly G-shattered by both $\bar{\mathcal{F}}_+^{(\eta)}$ and $\bar{\mathcal{F}}_-^{(\eta)}$. Let the pair $\left(s''_{\mathcal{Z}^{u+1}}, \mathbf{b}''_{u+1}\right)$ be such that $s''_{\mathcal{Z}^{u+1}} = s'_{\mathcal{Z}^u} \bigcup \{z_{i_0}\}$ and the vector $\mathbf{b}''_{u+1}$ is deduced from $\mathbf{b}'_u$ by inserting the component b_{i_0} at the right place. Clearly, neither $\bar{\mathcal{F}}_+^{(\eta)}$ nor $\bar{\mathcal{F}}_-^{(\eta)}$ strongly G-shatters $\left(s''_{\mathcal{Z}^{u+1}}, \mathbf{b}''_{u+1}\right)$, simply because they do not strongly G-shatter

the pair $(\{z_{i_0}\}, b_{i_0})$. On the contrary, it springs once more from (49) that $\left(s''_{\mathcal{Z}^{u+1}}, \mathbf{b}''_{u+1}\right)$ is strongly G-shattered by $\bar{\mathcal{F}}^{(\eta)}$. Summarizing, for each pair $(s'_{\mathcal{Z}^u}, \mathbf{b}'_u)$ strongly G-shattered by both $\bar{\mathcal{F}}^{(\eta)}_+$ and $\bar{\mathcal{F}}^{(\eta)}_-$, we can exhibit by means of an injective mapping a pair $\left(s''_{\mathcal{Z}^{u+1}}, \mathbf{b}''_{u+1}\right)$ strongly G-shattered by $\bar{\mathcal{F}}^{(\eta)}$ but not by $\bar{\mathcal{F}}^{(\eta)}_+$ or $\bar{\mathcal{F}}^{(\eta)}_-$. Collecting all terms, we obtain

$$\begin{aligned} s\left(\bar{\mathcal{F}}^{(\eta)}\right) &\geqslant s\left(\bar{\mathcal{F}}^{(\eta)}_+\right) + s\left(\bar{\mathcal{F}}^{(\eta)}_-\right) + 1 \\ &\geqslant \ell\left(\bar{\mathcal{F}}\right) - 1, \end{aligned} \tag{50}$$

where the function ℓ returns the number of leaves of the (sub)tree whose root is its argument. Thus, finishing the proof boils down to exhibiting the appropriate lower bound on $\ell\left(\bar{\mathcal{F}}\right)$. To that end, we proceed by induction on the depth of the node. The hypothesis is that

$$\ell\left(\bar{\mathcal{F}}\right) \geqslant \left|\bar{\mathcal{F}}\right|^{\frac{1}{\log_2(3M_\gamma n)}}. \tag{51}$$

It is obviously true for the leaves (which are of cardinality 1). Suppose now that it is true for the two sons of an inner node. Then, Inequality (48) gives:

$$\begin{aligned} \ell\left(\bar{\mathcal{F}}\right) &= \ell\left(\bar{\mathcal{F}}_+\right) + \ell\left(\bar{\mathcal{F}}_-\right) \\ &\geqslant \left|\bar{\mathcal{F}}_+\right|^{\frac{1}{\log_2(3M_\gamma n)}} + \left|\bar{\mathcal{F}}_-\right|^{\frac{1}{\log_2(3M_\gamma n)}} \\ &\geqslant 2\left(\frac{\left|\bar{\mathcal{F}}\right|}{3M_\gamma n}\right)^{\frac{1}{\log_2(3M_\gamma n)}} \\ &= \left|\bar{\mathcal{F}}\right|^{\frac{1}{\log_2(3M_\gamma n)}}. \end{aligned}$$

The induction hypothesis has been proved. Combining Inequalities (47), (50) and (51) produces by transitivity:

$$\left|\tilde{\mathcal{F}}\right|^{\frac{1}{\log_2(3M_\gamma n)}} \leqslant \Sigma,$$

or equivalently

$$\begin{aligned} \left|\tilde{\mathcal{F}}\right| &\leqslant \Sigma^{\log_2(3M_\gamma n)} \\ &= (3M_\gamma n)^{\log_2(\Sigma)}, \end{aligned}$$

i.e., Inequality (46), the result announced. □

With Lemma 15 at hand, the proof of Lemma 6 is straightforward.

Proof. Let us consider any vector $\mathbf{z}_n \in \mathcal{Z}^n$ and let $s_{\mathcal{Z}^n} = \{z_i : 1 \leqslant i \leqslant n\}$ be the smallest subset of $\mathcal{Z}$ containing all the components of $\mathbf{z}_n$. Note that its cardinality can be strictly

inferior to n, in case that $\mathbf{z}_n$ has two identical components. Setting $\eta = \frac{\epsilon}{2}$ in (42), one obtains:

$$\mathcal{M}\left(\epsilon, \mathcal{F}_\gamma, d_{\infty,\mathbf{z}_n}\right) \leqslant \mathcal{M}\left(2, \mathcal{F}_\gamma^{\left(\frac{\epsilon}{2}\right)}, d_{\infty,\mathbf{z}_n}\right).$$

Furthermore, by definition,

$$\mathcal{M}\left(2, \mathcal{F}_\gamma^{\left(\frac{\epsilon}{2}\right)}, d_{\infty,\mathbf{z}_n}\right) = \mathcal{M}\left(2, \mathcal{F}_\gamma^{\left(\frac{\epsilon}{2}\right)}\Big|_{s_{\mathcal{Z}^n}}, d_{\infty,\mathbf{z}_n}\right).$$

Let $\tilde{\mathcal{F}}$ be a subset of $\mathcal{F}$ such that $\tilde{\mathcal{F}}_\gamma^{\left(\frac{\epsilon}{2}\right)}\Big|_{s_{\mathcal{Z}^n}}$ is 2-separated in the metric $d_{\infty,\mathbf{z}_n}$ and of cardinality $\mathcal{M}\left(2, \mathcal{F}_\gamma^{\left(\frac{\epsilon}{2}\right)}\Big|_{s_{\mathcal{Z}^n}}, d_{\infty,\mathbf{z}_n}\right)$ (maximal cardinality). $\tilde{\mathcal{F}}$ has been built so as to satisfy the hypotheses of Lemma 15. Since $\left|\tilde{\mathcal{F}}_\gamma^{\left(\frac{\epsilon}{2}\right)}\Big|_{s_{\mathcal{Z}^n}}\right| = \mathcal{M}\left(2, \mathcal{F}_\gamma^{\left(\frac{\epsilon}{2}\right)}\Big|_{s_{\mathcal{Z}^n}}, d_{\infty,\mathbf{z}_n}\right)$ and S-G-dim $\left(\tilde{\mathcal{F}}^{\left(\frac{\epsilon}{2}\right)}\right) \leqslant$ S-G-dim $\left(\mathcal{F}^{\left(\frac{\epsilon}{2}\right)}\right)$, the application of the lemma provides us with:

$$\mathcal{M}\left(2, \mathcal{F}_\gamma^{\left(\frac{\epsilon}{2}\right)}\Big|_{s_{\mathcal{Z}^n}}, d_{\infty,\mathbf{z}_n}\right) \leqslant \left(\frac{6\gamma n}{\epsilon}\right)^{\log_2(\Sigma)} \tag{52}$$

where $\Sigma = \sum_{u=0}^{d_G} \binom{n}{u} \left(\frac{2\gamma}{\epsilon}\right)^u$ with d_G standing for S-G-dim $\left(\mathcal{F}^{\left(\frac{\epsilon}{2}\right)}\right)$. According to (45),

$$\text{S-G-dim}\left(\mathcal{F}^{\left(\frac{\epsilon}{2}\right)}\right) \leqslant d_G\left(\frac{\epsilon}{4}\right).$$

Since by hypothesis, $n \geqslant d_G\left(\frac{\epsilon}{4}\right)$, Σ can be bounded from above by replacing in its formula d_G with $d_G\left(\frac{\epsilon}{4}\right)$ and resorting to Corollary 3.18 in Mohri et al. (2018), leading to:

$$\begin{aligned}\Sigma &\leqslant \sum_{u=0}^{d_G\left(\frac{\epsilon}{4}\right)} \binom{n}{u} \left(\frac{2\gamma}{\epsilon}\right)^u \\ &\leqslant \left(\frac{2\gamma e n}{d_G\left(\frac{\epsilon}{4}\right)\epsilon}\right)^{d_G\left(\frac{\epsilon}{4}\right)},\end{aligned} \tag{53}$$

where the standard convention that the last term takes the value 1 for $d_G\left(\frac{\epsilon}{4}\right) = 0$ is made. Substituting (53) into (52) and taking the supremum over $\mathcal{Z}^n$ concludes the proof of (11). □

B.3 Margin Graph Dimension - L_2-norm

The sketches of the proofs of the two L_2-norm combinatorial results, Lemma 7 and Lemma 9, are basically the same. Compared to the sketch of the proof of Lemma 6, they exhibit two major differences. First, the construction of the 2-separating tree is more sophisticated, since it rests on a small deviation principle (in place of the sole pigeonhole principle). Second, one additional step is involved, which implements a probabilistic extraction principle.

This additional step makes the result dimension free. We begin the proof with the formulation of the small deviation principle. This extension of Lemma 5 in Mendelson and Vershynin (2003) is tailored to our needs.

Lemma 16. *Let T be a random variable taking values in $[\![0; M]\!]$ with $M \geqslant 2$. Suppose that $Var[T] \geqslant 9$. Then there exists either $(\alpha, \beta) \in [\![1; M-1]\!] \times \left[\frac{1}{M^2}, \frac{1}{2}\right]$ such that*

$$\begin{cases} \mathbb{P}\{T \geqslant \alpha + 1\} \geqslant \max\left\{\frac{1}{2}\beta, \frac{1}{M^2}\right\} \\ \mathbb{P}\{T \leqslant \alpha - 1\} \geqslant 1 - \beta \end{cases}$$

or $(\alpha', \beta') \in [\![1; M-1]\!] \times \left[\frac{1}{M^2}, \frac{1}{2}\right]$ such that

$$\begin{cases} \mathbb{P}\{T \geqslant \alpha' + 1\} \geqslant 1 - \beta' \\ \mathbb{P}\{T \leqslant \alpha' - 1\} \geqslant \max\left\{\frac{1}{2}\beta', \frac{1}{M^2}\right\} \end{cases}.$$

Proof. We first note that the hypothesis $\mathrm{Var}[T] \geqslant 9$ implies that $M \geqslant 6$. Let M_T be the smallest median of T belonging to $[\![0; M]\!]$. Then, several cases must be distinguished, according to the values of M_T and $M - M_T$. Since they can all be treated the same and the one implying the largest upper bound on the variance, i.e., the one from which springs the hypothesis on the variance, is $M - M_T \geqslant 2$ and $M_T \geqslant 2$, we focus on it in the sequel. Let us define the sequences $(\beta_k)_{k\in\mathbb{N}^*}$ and $(\beta'_k)_{k\in\mathbb{N}^*}$ as follows:

$$\forall k \in \mathbb{N}^*, \quad \begin{cases} \beta_k = \mathbb{P}\{T \geqslant M_T + k\} \\ \beta'_k = \mathbb{P}\{T \leqslant M_T - k\} \end{cases}.$$

Note that by definition of M_T, both β_1 and β'_1 are inferior or equal to $\frac{1}{2}$. Assume that the conclusion of the lemma fails. We claim that

$$\begin{cases} \forall k \in [\![2; M - M_T]\!], \quad \beta_k \leqslant \max\left\{\frac{2(M-M_T)-k+1}{(M-M_T)^3}, \frac{1}{2^k}\right\} \\ \forall k \in [\![2; M_T]\!], \quad \beta'_k \leqslant \max\left\{\frac{2M_T-k+1}{M_T^3}, \frac{1}{2^k}\right\} \end{cases}.$$

Indeed, assume that $\beta_k > \max\left\{\frac{2(M-M_T)-k+1}{(M-M_T)^3}, \frac{1}{2^k}\right\}$ for some $k \in [\![2; M - M_T]\!]$ and let k_0 be the smallest such index. By construction, $\beta_{k_0} > \max\left\{\frac{1}{2}\beta_{k_0-1}, \frac{1}{(M-M_T)^2}\right\}$ (even for $k_0 = 2$ and $k_0 = M - M_T$), so that

$$\begin{cases} \mathbb{P}\{T \geqslant M_T + k_0\} = \beta_{k_0} > \max\left\{\frac{1}{2}\beta_{k_0-1}, \frac{1}{M^2}\right\} \\ \mathbb{P}\{T \leqslant M_T + k_0 - 2\} = 1 - \mathbb{P}\{T \geqslant M_T + k_0 - 1\} = 1 - \beta_{k_0-1} \end{cases}.$$

Since $\beta_{k_0-1} \geqslant \beta_{k_0} > \frac{1}{M^2}$ and $\beta_{k_0-1} \leqslant \beta_1 \leqslant \frac{1}{2}$, so that $\beta_{k_0-1} \in \left[\frac{1}{M^2}, \frac{1}{2}\right]$, this implies that the conclusion of the lemma would hold with α being $M_T + k_0 - 1$ and $\beta = \beta_{k_0-1}$, which contradicts the assumption that the conclusion of the lemma fails. The inequality $\beta'_k \leqslant \max\left\{\frac{2M_T-k+1}{M_T^3}, \frac{1}{2^k}\right\}$ can be proved in a symmetrical way. As a consequence, upper bounding the maxima by the corresponding sums gives:

$$\begin{aligned}
\mathrm{Var}\left[T\right] =& \ \mathrm{Var}\left[T - M_T\right] \\
\leqslant& \ \mathbb{E}\left[\left(T - M_T\right)^2\right] \\
=& \ \sum_{t=0}^{+\infty} \mathbb{P}\left\{\left(T - M_T\right)^2 > t\right\} \\
=& \ \sum_{t=0}^{+\infty} \left(\mathbb{P}\left\{T > M_T + \sqrt{t}\right\} + \mathbb{P}\left\{T < M_T - \sqrt{t}\right\}\right) \\
=& \ \sum_{k=1}^{M-M_T} (2k-1)\,\beta_k + \sum_{k=1}^{M_T} (2k-1)\,\beta'_k \\
<& \ 2\sum_{k=1}^{+\infty} \frac{2k-1}{2^k} + \sum_{k=2}^{M-M_T} (2k-1)\,\frac{2\left(M - M_T\right) - k + 1}{\left(M - M_T\right)^3} + \sum_{k=2}^{M_T} (2k-1)\,\frac{2M_T - k + 1}{M_T^3} \\
\leqslant& \ 6 + 2 \max_{\Delta \in \mathbb{N}\setminus\{1,2\}} \frac{(8\Delta + 11)\,(\Delta - 1)}{6\Delta^2} \\
<& \ 9.
\end{aligned}$$

This is in contradiction with the hypothesis that $\mathrm{Var}\left[T\right] \geqslant 9$ and thus concludes the proof. □

Proposition 1. *Let $\mathcal{T} = \{t_i : 1 \leqslant i \leqslant n\}$ be a finite set and $\mathbf{t}_n = (t_i)_{1\leqslant i\leqslant n}$. Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ into $[\![0; M_{\mathcal{F}}]\!]$ with $M_{\mathcal{F}} \geqslant 2$. Suppose that $\mathcal{F}$ is of cardinality at least 2 and $\epsilon \in \mathbb{R}_+^*$ is such that $\mathcal{F}$ is ϵ-separated in the metric $d_{2,\mathbf{t}_n}$. Then there exists $i_0 \in [\![1; n]\!]$ such that*

$$Var\left[f\left(t_{i_0}\right)\right] \geqslant \frac{\epsilon^2}{4}.$$

Proof. Let us endow $\mathcal{F}$ with the uniform (counting) measure. Then, the separation assumption on $\mathcal{F}$ can be used to derive a lower bound on $\mathbb{E}\left[d^2_{2,\mathbf{t}_n}\left(f, f'\right)\right]$. Indeed, with probability $1 - |\mathcal{F}|^{-1}$ we have $f \neq f'$ and, whenever this event occurs, $d_{2,\mathbf{t}_n}\left(f, f'\right) \geqslant \epsilon$. As a consequence,

$$\begin{aligned}
\mathbb{E}\left[d^2_{2,\mathbf{t}_n}\left(f, f'\right)\right] \geqslant& \ \left(1 - |\mathcal{F}|^{-1}\right)\epsilon^2 \\
\geqslant& \ \frac{\epsilon^2}{2}.
\end{aligned}$$

Furthermore,

$$\begin{aligned}\mathbb{E}\left[d^2_{2,\mathbf{t}_n}\left(f,f'\right)\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left(f\left(t_i\right)-f'\left(t_i\right)\right)^2\right]\\ &= \frac{2}{n}\sum_{i=1}^{n}\mathrm{Var}\left[f\left(t_i\right)\right].\end{aligned}$$

Thus, there exists $i_0 \in [\![1;n]\!]$ such that

$$\mathrm{Var}\left[f\left(t_{i_0}\right)\right] \geqslant \frac{1}{2}\mathbb{E}\left[d^2_{2,\mathbf{t}_n}\left(f,f'\right)\right] \geqslant \frac{\epsilon^2}{4}.$$

□

Lemma 17. *Let $\mathcal{T}=\{t_i:\ 1\leqslant i\leqslant n\}$ be a finite set and $\mathbf{t}_n=(t_i)_{1\leqslant i\leqslant n}$. Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ into $[\![0;M_{\mathcal{F}}]\!]$ with $M_{\mathcal{F}}\geqslant 2$. Suppose that $\mathcal{F}$ is of cardinality at least 2 and is 6-separated in the metric $d_{2,\mathbf{t}_n}$. Then there exist an index $i_0\in[\![1;n]\!]$ and either $(\alpha,\beta)\in[\![1;M_{\mathcal{F}}-1]\!]\times\left[\frac{1}{M^2_{\mathcal{F}}},\frac{1}{2}\right]$ such that*

$$\begin{cases}\left|\{f\in\mathcal{F}:\ f\left(t_{i_0}\right)\geqslant\alpha+1\}\right| \geqslant \max\left\{\frac{1}{2}\beta,\frac{1}{M^2_{\mathcal{F}}}\right\}|\mathcal{F}|\\ \left|\{f\in\mathcal{F}:\ f\left(t_{i_0}\right)\leqslant\alpha-1\}\right| \geqslant (1-\beta)\,|\mathcal{F}|\end{cases}$$

or $(\alpha',\beta')\in[\![1;M_{\mathcal{F}}-1]\!]\times\left[\frac{1}{M^2_{\mathcal{F}}},\frac{1}{2}\right]$ such that

$$\begin{cases}\left|\{f\in\mathcal{F}:\ f\left(t_{i_0}\right)\geqslant\alpha'+1\}\right| \geqslant (1-\beta')\,|\mathcal{F}|\\ \left|\{f\in\mathcal{F}:\ f\left(t_{i_0}\right)\leqslant\alpha'-1\}\right| \geqslant \max\left\{\frac{1}{2}\beta',\frac{1}{M^2_{\mathcal{F}}}\right\}|\mathcal{F}|\end{cases}.$$

Proof. According to Proposition 1, there exists $i_0\in[\![1;n]\!]$ such that

$$\mathrm{Var}\left[f\left(t_{i_0}\right)\right] \geqslant \frac{1}{2}\mathbb{E}\left[d^2_{2,\mathbf{t}_n}\left(f,f'\right)\right] \geqslant 9.$$

This implies that the random variable $f\left(t_{i_0}\right)$ satisfies the hypotheses of Lemma 16, and the conclusion then springs from the application of this lemma. □

Lemma 17 will be used in the proof of the combinatorial result involving the margin Natarajan dimension: Lemma 9. However, we established it in this section, because its proof can be easily simplified to produce the following variant, appropriate for the margin Graph dimension.

Lemma 18. *Let $\mathcal{T} = \{t_i : 1 \leqslant i \leqslant n\}$ be a finite set and $\mathbf{t}_n = (t_i)_{1 \leqslant i \leqslant n}$. Suppose that $\mathcal{F} \subset \mathbb{Z}^{\mathcal{T}}$ is of cardinality at least 2 and is 5-separated in the metric $d_{2,\mathbf{t}_n}$. Then there exist an index $i_0 \in [\![1; n]\!]$ and either $(\alpha, \beta) \in \mathbb{Z} \times \left(0, \frac{1}{2}\right]$ such that*

$$\begin{cases} |\{f \in \mathcal{F} : \ f(t_{i_0}) \geqslant \alpha + 1\}| \geqslant \frac{1}{2}\beta |\mathcal{F}| \\ |\{f \in \mathcal{F} : \ f(t_{i_0}) \leqslant \alpha - 1\}| \geqslant (1 - \beta) |\mathcal{F}| \end{cases}$$

or $(\alpha', \beta') \in \mathbb{Z} \times \left(0, \frac{1}{2}\right]$ such that

$$\begin{cases} |\{f \in \mathcal{F} : \ f(t_{i_0}) \geqslant \alpha' + 1\}| \geqslant (1 - \beta') |\mathcal{F}| \\ |\{f \in \mathcal{F} : \ f(t_{i_0}) \leqslant \alpha' - 1\}| \geqslant \frac{1}{2}\beta' |\mathcal{F}| \end{cases}.$$

The following lemma is the basic combinatorial result underlying Lemma 7.

Lemma 19. *Let $\mathcal{F}$ be a function class defined as in Definition 8. For every $s_{\mathcal{Z}^n} = \{z_i = (x_i, y_i) : 1 \leqslant i \leqslant n\} \subset \mathcal{Z}$, $\gamma \in (0, 1]$ and $\eta \in \left(0, \frac{\gamma}{2}\right]$, if $\mathcal{F}_\gamma^{(\eta)}\Big|_{s_{\mathcal{Z}^n}}$ is 5-separated in the metric $d_{2,\mathbf{z}_n}$, then*

$$\left|\mathcal{F}_\gamma^{(\eta)}\Big|_{s_{\mathcal{Z}^n}}\right| \leqslant \Sigma^{\frac{3}{2}}, \tag{54}$$

where $\Sigma = \sum_{u=0}^{d_G} \binom{n}{u} M_\gamma^u$ with $M_\gamma = \left\lfloor \frac{\gamma}{\eta} \right\rfloor$ and $d_G = S\text{-}G\text{-}dim\left(\mathcal{F}^{(\eta)}\right)$.

Proof. The principle of the proof is the one of the proof of Lemma 15. Two of the three main formulas still apply: Inequalities (47) and (50). For $\left|\mathcal{F}_\gamma^{(\eta)}\Big|_{s_{\mathcal{Z}^n}}\right| \geqslant 2$, the incidence of the change of metric is concentrated in the derivation of the 2-separating tree of $\tilde{\mathcal{F}}$ (or equivalently $\mathcal{F}_\gamma^{(\eta)}\Big|_{s_{\mathcal{Z}^n}}$), and thus the lower bound on $\ell\left(\tilde{\mathcal{F}}\right)$. Since the inner nodes $\bar{\mathcal{F}}$ are such that the corresponding sets $\bar{\mathcal{F}}_\gamma^{(\eta)}\Big|_{s_{\mathcal{Z}^n}}$ are 5-separated in the metric $d_{2,\mathbf{z}_n}$, then according to Lemma 18, for each of these nodes, we can ensure that there exists $\beta \in \left(0, \frac{1}{2}\right]$ such that the two sons $\bar{\mathcal{F}}_+$ and $\bar{\mathcal{F}}_-$ of $\bar{\mathcal{F}}$ verify either $\left|\bar{\mathcal{F}}_+\right| \geqslant (1 - \beta)\left|\bar{\mathcal{F}}\right|$ and $\left|\bar{\mathcal{F}}_-\right| \geqslant \frac{1}{2}\beta\left|\bar{\mathcal{F}}\right|$ or vice versa (in place of (48)). As a consequence, the counterpart of (51) is:

$$\ell\left(\bar{\mathcal{F}}\right) \geqslant \left|\bar{\mathcal{F}}\right|^{\frac{2}{3}}. \tag{55}$$

Once more, the proof is an induction on the depth of the node. Inequality (55) is obviously true for the leaves (which are of cardinality 1). Suppose now that it is true for the two sons of an inner node. Then,

$$\begin{aligned} \ell\left(\bar{\mathcal{F}}\right) &= \ell\left(\bar{\mathcal{F}}_+\right) + \ell\left(\bar{\mathcal{F}}_-\right) \\ &\geqslant \left[(1 - \beta)^{\frac{2}{3}} + \left(\frac{\beta}{2}\right)^{\frac{2}{3}}\right] \left|\bar{\mathcal{F}}\right|^{\frac{2}{3}} \\ &\geqslant \left|\bar{\mathcal{F}}\right|^{\frac{2}{3}}. \end{aligned}$$

Note that the value of the constant, $\frac{2}{3}$, can be obtained by maximizing K over $(0,1]$ subject to $\left(\frac{1}{2}\right)^K \left[1+\left(\frac{1}{2}\right)^K\right] \geqslant 1$. Finally, combining Inequalities (47), (50) and (55) produces (54) by transitivity. □

The following lemma, a slight improvement of Lemma 13 in Mendelson and Vershynin (2003), implements the probabilistic extraction principle.

Lemma 20. *Let $\mathcal{T} = \{t_i : 1 \leqslant i \leqslant n\}$ be a finite set, $\mathbf{t}_n = (t_i)_{1\leqslant i\leqslant n}$ and $M_{\mathcal{F}} \in \mathbb{R}_+^*$. Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ into $[0, M_{\mathcal{F}}]$ with finite cardinality $|\mathcal{F}| \geqslant 2$. Assume that for some $\epsilon \in (0, M_{\mathcal{F}}]$, $\mathcal{F}$ is ϵ-separated with respect to the metric $d_{2,\mathbf{t}_n}$, and let*

$$r = \frac{\ln(|\mathcal{F}|)}{K_e \epsilon^4}$$

with

$$K_e = \frac{3}{112 M_{\mathcal{F}}^4}.$$

Then, there exists a subvector $\mathbf{t}'_q$ of $\mathbf{t}_n$ of size $q \leqslant r$ such that $\mathcal{F}$ is $\frac{\epsilon}{2}$-separated with respect to the metric $d_{2,\mathbf{t}'_q}$.

Proof. This proof uses an abuse of notation that will be repeated in the sequel: the symbol $\mathbb{P}$ designates different probability measures, some of which implicitly defined. We first note that the statement is trivially true for $r \geqslant n$ (it suffices to set $\mathbf{t}'_q = \mathbf{t}_n$). Thus, we proceed under the hypothesis $r \in [1, n)$. Let us set $\mathcal{F} = \{f_j : 1 \leqslant j \leqslant |\mathcal{F}|\}$ and $\mathcal{D}_{\mathcal{F}} = \{f_j - f_{j'} : 1 \leqslant j < j' \leqslant |\mathcal{F}|\}$. The set $\mathcal{D}_{\mathcal{F}}$ has cardinality $|\mathcal{D}_{\mathcal{F}}| < \frac{1}{2}|\mathcal{F}|^2$. Let $(\epsilon_i)_{1\leqslant i\leqslant n}$ be a sequence of n independent Bernoulli random variables with common expectation $\mu = \frac{r}{2n}$. Then, by application of the ϵ-separation property, for every $\delta_f \in \mathcal{D}_{\mathcal{F}}$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \epsilon_i \delta_f(t_i)^2 < \frac{\epsilon^2\mu}{2}\right) \leqslant \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n (\mu - \epsilon_i)\,\delta_f(t_i)^2 > \frac{\epsilon^2\mu}{2}\right). \tag{56}$$

Since by construction, for every $i \in [\![1; n]\!]$, $\mathbb{E}\left[(\mu-\epsilon_i)\,\delta_f(t_i)^2\right] = 0$ and $|\mu - \epsilon_i|\,\delta_f(t_i)^2 \leqslant M_{\mathcal{F}}^2(1-\mu) < M_{\mathcal{F}}^2$ with probability one, the right-hand side of (56) can be bounded from above thanks to Bernstein's inequality. Given that

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[(\mu-\epsilon_i)^2\,\delta_f(t_i)^4\right] \leqslant M_{\mathcal{F}}^4 \mu(1-\mu) < M_{\mathcal{F}}^4 \mu,$$

we obtain

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\delta_f\left(t_i\right)^2<\frac{\epsilon^2\mu}{2}\right) &\leqslant \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\mu-\epsilon_i\right)\delta_f\left(t_i\right)^2>\frac{\epsilon^2\mu}{2}\right)\\
&\leqslant \exp\left(-\frac{3\mu n\epsilon^4}{4\left(6M_{\mathcal{F}}^4+M_{\mathcal{F}}^2\epsilon^2\right)}\right)\\
&\leqslant \exp\left(-\frac{3r\epsilon^4}{56M_{\mathcal{F}}^4}\right)\\
&= |\mathcal{F}|^{-2}.
\end{aligned}$$

Therefore, given the assumption on r, applying the union bound provides us with:

$$\begin{aligned}
\mathbb{P}\left(\exists\delta_f\in\mathcal{D}_{\mathcal{F}}:\ \left(\frac{1}{r}\sum_{i=1}^{n}\epsilon_i\delta_f\left(t_i\right)^2\right)^{\frac{1}{2}}<\frac{\epsilon}{2}\right) &= \mathbb{P}\left(\exists\delta_f\in\mathcal{D}_{\mathcal{F}}:\ \frac{1}{n}\sum_{i=1}^{n}\epsilon_i\delta_f\left(t_i\right)^2<\frac{\epsilon^2\mu}{2}\right)\\
&\leqslant \sum_{\delta_f\in\mathcal{D}_{\mathcal{F}}}\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\delta_f\left(t_i\right)^2<\frac{\epsilon^2\mu}{2}\right)\\
&\leqslant |\mathcal{D}_{\mathcal{F}}|\cdot|\mathcal{F}|^{-2}\\
&< \frac{1}{2}. \qquad (57)
\end{aligned}$$

Moreover, if $\mathcal{S}_1$ is the random set $\{i\in[\![1;n]\!]:\ \epsilon_i=1\}$, then by Markov's inequality,

$$\mathbb{P}\left(|S_1|>r\right)=\mathbb{P}\left(\sum_{i=1}^{n}\epsilon_i>r\right)\leqslant\frac{1}{2}. \qquad (58)$$

Combining (57) and (58) by means of the union bound provides us with

$$\mathbb{P}\left\{\left(\exists\delta_f\in\mathcal{D}_{\mathcal{F}}:\ \left(\frac{1}{r}\sum_{i=1}^{n}\epsilon_i\delta_f\left(t_i\right)^2\right)^{\frac{1}{2}}<\frac{\epsilon}{2}\right)\ \text{or}\ \left(|S_1|>r\right)\right\}<1$$

which implies that

$$\mathbb{P}\left\{\left(\forall\delta_f\in\mathcal{D}_{\mathcal{F}}:\ \|\delta_f\|_{L_2\left(\mu_{(t_i)_{i\in\mathcal{S}_1}}\right)}\geqslant\frac{\epsilon}{2}\right)\ \text{and}\ \left(|S_1|\leqslant r\right)\right\}>0.$$

This translates into the fact that there exists a subvector $\mathbf{t}'_q$ of $\mathbf{t}_n$ of size $q\leqslant r$ such that the class $\mathcal{F}$ is $\frac{\epsilon}{2}$-separated with respect to the metric $d_{2,\mathbf{t}'_q}$, i.e., our claim. □

The proof of Lemma 7 is the following one.

Proof. Let us consider any vector $\mathbf{z}_n \in \mathcal{Z}^n$ and let $s_{\mathcal{Z}^n} = \{z_i : 1 \leqslant i \leqslant n\}$ be the smallest subset of $\mathcal{Z}$ containing all the components of $\mathbf{z}_n$. Note that its cardinality can be strictly inferior to n, in case that $\mathbf{z}_n$ has two identical components. By defintion,

$$\mathcal{M}\left(\epsilon, \mathcal{F}_\gamma, d_{2,\mathbf{z}_n}\right) = \mathcal{M}\left(\epsilon, \mathcal{F}_\gamma|_{s_{\mathcal{Z}^n}}, d_{2,\mathbf{z}_n}\right).$$

Let $\tilde{\mathcal{F}}$ be a subset of $\mathcal{F}$ of cardinality $\mathcal{M}\left(\epsilon, \mathcal{F}_\gamma, d_{2,\mathbf{z}_n}\right)$ such that $\tilde{\mathcal{F}}_\gamma\Big|_{s_{\mathcal{Z}^n}}$ is ϵ-separated with respect to the metric $d_{2,\mathbf{z}_n}$ and in bijection with $\tilde{\mathcal{F}}$. By construction, $\tilde{\mathcal{F}}_\gamma\Big|_{s_{\mathcal{Z}^n}}$ satisfies the hypotheses of Lemma 20, with $M_{\mathcal{F}} = \gamma$. Consequently, applying the lemma establishes that there exists a subvector $\mathbf{z}'_q$ of $\mathbf{z}_n$ of size

$$q \leqslant \frac{\ln\left(\left|\tilde{\mathcal{F}}\right|\right)}{K_\gamma \epsilon^4} \tag{59}$$

with $K_\gamma = \frac{3}{112\gamma^4}$ such that the class $\tilde{\mathcal{F}}_\gamma\Big|_{s_{\mathcal{Z}^n}}$ is also $\frac{\epsilon}{2}$-separated with respect to the metric $d_{2,\mathbf{z}'_q}$. As a consequence, denoting $s_{\mathcal{Z}^q} = \{z'_i = (x'_i, y'_i) : 1 \leqslant i \leqslant q\}$ ($|s_{\mathcal{Z}^q}| \leqslant q$), it appears that $\left|\tilde{\mathcal{F}}_\gamma\Big|_{s_{\mathcal{Z}^n}}\right| = \left|\tilde{\mathcal{F}}_\gamma\Big|_{s_{\mathcal{Z}^q}}\right|$ (and thus $\left|\tilde{\mathcal{F}}_\gamma\Big|_{s_{\mathcal{Z}^q}}\right| = \left|\tilde{\mathcal{F}}\right|$). Applying (41) to the function class $\tilde{\mathcal{F}}_\gamma\Big|_{s_{\mathcal{Z}^q}}$ with $N = 5$ and the corresponding largest possible value for η, $\frac{\epsilon}{12}$, it appears that $\tilde{\mathcal{F}}_\gamma^{(\frac{\epsilon}{12})}\Big|_{s_{\mathcal{Z}^q}}$ is a set of cardinality $\left|\tilde{\mathcal{F}}\right|$ which is 5-separated with respect to the metric $d_{2,\mathbf{z}'_q}$. Thus, Lemma 19 can be applied to $\tilde{\mathcal{F}}$ as follows:

$$\begin{aligned}
\left|\tilde{\mathcal{F}}\right| &= \left|\tilde{\mathcal{F}}_\gamma^{(\frac{\epsilon}{12})}\Big|_{s_{\mathcal{Z}^q}}\right| \\
&\leqslant \left(\sum_{u=0}^{d_G} \binom{q}{u} \left(\frac{12\gamma}{\epsilon}\right)^u\right)^{\frac{3}{2}} \\
&\leqslant \left(\frac{12\gamma e q}{d_G \epsilon}\right)^{\frac{3d_G}{2}},
\end{aligned} \tag{60}$$

where $d_G = \text{S-G-dim}\left(\tilde{\mathcal{F}}^{(\frac{\epsilon}{12})}\right)$. A substitution of the upper bound on q provided by (59) into (60) gives:

$$\left|\tilde{\mathcal{F}}\right| \leqslant \left(K\left(\frac{\gamma}{\epsilon}\right)^5 \frac{\ln\left(\left|\tilde{\mathcal{F}}\right|\right)}{d_G}\right)^{\frac{3}{2}d_G}$$

with $K = 448e$. In order to upper bound $\ln\left(\left|\tilde{\mathcal{F}}\right|^{\frac{1}{d_G}}\right)$, we resort once more to (40), this time with $u_0 = 1$. Thus,

$$\left|\tilde{\mathcal{F}}\right|^{\frac{1}{d_G}} \leqslant \ln^{\frac{3}{2}}\left(\left|\tilde{\mathcal{F}}\right|^{\frac{1}{d_G}}\right)\left(K\left(\frac{\gamma}{\epsilon}\right)^5\right)^{\frac{3}{2}}$$

and $\left|\tilde{\mathcal{F}}\right| = \mathcal{M}\left(\epsilon, \mathcal{F}_\gamma, d_{2,\mathbf{z}_n}\right)$ imply that

$$\begin{aligned}\mathcal{M}\left(\epsilon, \mathcal{F}_\gamma, d_{2,\mathbf{z}_n}\right) &\leqslant \left(2K\left(\frac{\gamma}{\epsilon}\right)^5\right)^{\frac{12}{5}d_G} \\ &\leqslant \left(\frac{5\gamma}{\epsilon}\right)^{12d_G}. \end{aligned} \tag{61}$$

Since $\tilde{\mathcal{F}}^{(\frac{\epsilon}{12})} \subset \mathcal{F}^{(\frac{\epsilon}{12})}$, by application of Formula (45),

$$\begin{aligned}\text{S-G-dim}\left(\tilde{\mathcal{F}}^{(\frac{\epsilon}{12})}\right) &\leqslant \text{S-G-dim}\left(\mathcal{F}^{(\frac{\epsilon}{12})}\right) \\ &\leqslant d_G\left(\frac{\epsilon}{24}\right). \end{aligned} \tag{62}$$

By substitution of (62) into (61), we obtain that for every vector $\mathbf{z}_n \in \mathcal{Z}^n$,

$$\mathcal{M}\left(\epsilon, \mathcal{F}_\gamma, d_{2,\mathbf{z}_n}\right) \leqslant \left(\frac{5\gamma}{\epsilon}\right)^{12d_G\left(\frac{\epsilon}{24}\right)}. \tag{63}$$

At last, (63) implies (12) since its right-hand side does not depend on $\mathbf{z}_n$. □

B.4 Margin Natarajan Dimension - Uniform Convergence Norm

The proof of Lemma 8 is essentially that of Lemma 6, with the main differences being concentrated in the basic combinatorial result (the counterpart of Lemma 15). Thus, we only highlight these differences. The first one is that the number $s\left(\bar{\mathcal{F}}^{(\eta)}\right)$ of pairs $(s'_{\mathcal{Z}^u}, \mathbf{b}'_u)$ strongly G-shattered by $\bar{\mathcal{F}}^{(\eta)}$ is replaced with the number $s'\left(\bar{\mathcal{F}}^{(\eta)}\right)$ of triplets $(s'_{\mathcal{Z}^u}, \mathbf{b}'_u, \mathbf{c}'_u)$ strongly N-shattered by $\bar{\mathcal{F}}^{(\eta)}$. Let $d_N = \text{S-N-dim}\left(\mathcal{F}^{(\eta)}\right)$. Once more, under the hypothesis $d_N \geqslant 1$, combinatorics provides us with:

$$s'\left(\mathcal{F}^{(\eta)}\right) \leqslant \sum_{u=1}^{d_N} \binom{n}{u} M_\gamma^u \left(C-1\right)^u = \Sigma' - 1. \tag{64}$$

In order to obtain the counterpart of (50), i.e.,

$$\begin{aligned}s'\left(\bar{\mathcal{F}}^{(\eta)}\right) &\geqslant s'\left(\bar{\mathcal{F}}_+^{(\eta)}\right) + s'\left(\bar{\mathcal{F}}_-^{(\eta)}\right) + 1 \\ &\geqslant \ell\left(\bar{\mathcal{F}}\right) - 1, \end{aligned} \tag{65}$$

(49) must be replaced with

$$\begin{cases} \forall f_+ \in \bar{\mathcal{F}}_+, \quad f_+^{(\eta)}\left(z_{i_0}\right) - b_{i_0} \geqslant 1 \\ \forall f_- \in \bar{\mathcal{F}}_-, \quad f_-^{(\eta)}\left(x_{i_0}, c_{i_0}\right) + b_{i_0} \geqslant 1 \end{cases}. \tag{66}$$

This calls for an additional application of the pigeonhole principle in the derivation of the class $\bar{\mathcal{F}}_-$, so that the lower bound on $\left|\bar{\mathcal{F}}_-\right|$ provided by (48) is replaced with

$$\left|\bar{\mathcal{F}}_-\right| \geqslant \frac{\left|\bar{\mathcal{F}}\right|}{3M_\gamma\left(C-1\right)n}.$$

This implies that the counterpart of (51) is

$$\ell\left(\bar{\mathcal{F}}\right) \geqslant \left|\bar{\mathcal{F}}\right|^{\frac{1}{\log_2\left(3M_\gamma\sqrt{C-1}n\right)}}. \tag{67}$$

Once more, it is proved by induction on the depth of the node. Inequality (67) is obviously true for the leaves (which are of cardinality 1). Suppose now that it is true for the two sons of an inner node. Then,

$$\begin{aligned}
\ell\left(\bar{\mathcal{F}}\right) =& \ \ell\left(\bar{\mathcal{F}}_+\right) + \ell\left(\bar{\mathcal{F}}_-\right) \\
\geqslant& \ \left(\frac{\left|\bar{\mathcal{F}}\right|}{3M_\gamma n}\right)^{\frac{1}{\log_2\left(3M_\gamma\sqrt{C-1}n\right)}} + \left(\frac{\left|\bar{\mathcal{F}}\right|}{3M_\gamma\left(C-1\right)n}\right)^{\frac{1}{\log_2\left(3M_\gamma\sqrt{C-1}n\right)}} \\
=& \ \frac{1}{2}\left(\left(\sqrt{C-1}\right)^{\frac{1}{\log_2\left(3M_\gamma\sqrt{C-1}n\right)}} + \left(\sqrt{C-1}\right)^{-\frac{1}{\log_2\left(3M_\gamma\sqrt{C-1}n\right)}}\right)\left|\bar{\mathcal{F}}\right|^{\frac{1}{\log_2\left(3M_\gamma\sqrt{C-1}n\right)}} \\
\geqslant& \ \frac{1}{2}\min_{t\in\mathbb{R}_+^*}\left(t+\frac{1}{t}\right)\left|\bar{\mathcal{F}}\right|^{\frac{1}{\log_2\left(3M_\gamma\sqrt{C-1}n\right)}} \\
=& \ \left|\bar{\mathcal{F}}\right|^{\frac{1}{\log_2\left(3M_\gamma\sqrt{C-1}n\right)}}.
\end{aligned}$$

Combining Inequalities (64), (65) and (67), the counterpart of Inequality (46) is

$$\left|\mathcal{F}_\gamma^{(\eta)}\Big|_{s_{\mathcal{Z}^n}}\right| \leqslant \left(3M_\gamma\sqrt{C-1}n\right)^{\log_2(\Sigma')}.$$

B.5 Margin Natarajan Dimension - L_2-Norm

The main difference between the proof of Lemma 9 and the proof of Lemma 7 is located in the small deviation principle (Lemma 17 replaces Lemma 18). Since the consequences of this change appear in the derivation of the basic combinatorial result, we provide this latter result with its full proof.

Lemma 21. *Let $\mathcal{F}$ be a function class defined as in Definition 8. For every $s_{\mathcal{Z}^n} = \{z_i = (x_i, y_i) : 1 \leqslant i \leqslant n\} \subset \mathcal{Z}$, $\gamma \in (0,1]$ and $\eta \in \left(0, \frac{\gamma}{2}\right]$, if $\mathcal{F}_\gamma^{(\eta)}\Big|_{s_{\mathcal{Z}^n}}$ is 6-separated in the metric $d_{2,\mathbf{z}_n}$, then*

$$\left|\mathcal{F}_\gamma^{(\eta)}\Big|_{s_{\mathcal{Z}^n}}\right| \leqslant \left(\Sigma'\right)^{\log_2\left(\frac{M_\gamma^2}{\sqrt{M_\gamma^2-2}}\sqrt{C-1}\right)} \leqslant \left(\Sigma'\right)^{\frac{1}{2}\log_2\left(2M_\gamma^2(C-1)\right)} \tag{68}$$

where $\Sigma' = \sum_{u=0}^{d_N}\binom{n}{u}M_\gamma^u\left(C-1\right)^u$ with $M_\gamma = \left\lfloor\frac{\gamma}{\eta}\right\rfloor$ and $d_N = S\text{-}N\text{-}dim\left(\mathcal{F}^{(\eta)}\right)$.

Proof. Inequality (68) is trivially true for $\left|\mathcal{F}_\gamma^{(\eta)}\Big|_{s_{\mathcal{Z}^n}}\right| = 1$. Indeed, the minimal value of its right-hand side, corresponding to $d_N = 0$, is 1. Thus, the rest of the proof makes use of the restriction $\left|\mathcal{F}_\gamma^{(\eta)}\Big|_{s_{\mathcal{Z}^n}}\right| \geqslant 2$. A direct consequence is that according to Lemma 14, $d_N \geqslant 1$. A subset of $s_{\mathcal{Z}^n}$ of cardinality $u \in [\![1; n]\!]$ is denoted by $s'_{\mathcal{Z}^u} = \{z'_i : 1 \leqslant i \leqslant u\}$, with the convention

$$\forall\, (i,j) : 1 \leqslant i < j \leqslant u,\ \left(z'_i, z'_j\right) = (z_v, z_w) \Longrightarrow 1 \leqslant v < w \leqslant n.$$

For every subset $\bar{\mathcal{F}}$ of $\mathcal{F}$, denote by $s'\left(\bar{\mathcal{F}}^{(\eta)}\right)$ the number of triplets $(s'_{\mathcal{Z}^u}, \mathbf{b}'_u, \mathbf{c}'_u)$ with $s'_{\mathcal{Z}^u} \subset s_{\mathcal{Z}^n}$, $\mathbf{b}'_u \in [\![1; M_\gamma - 1]\!]^u$ and $\mathbf{c}'_u \in \mathcal{Y}^u$ (with $\forall i \in [\![1; u]\!]$, $c'_i \neq y'_i$) strongly N-shattered by $\bar{\mathcal{F}}^{(\eta)}$ (the convention above has been introduced to avoid handling duplicates). Since $d_N \geqslant 1$, Inequality (64) provides us once more with an upper bound on $s'\left(\mathcal{F}^{(\eta)}\right)$. In order to derive a lower bound on the same quantity, we also build a 2-separating tree of $\tilde{\mathcal{F}}$. Let $\bar{\mathcal{F}}$ be one of its nodes such that $\left|\bar{\mathcal{F}}\right| \geqslant 2$ (inner node). Its two sons, $\bar{\mathcal{F}}_+$ and $\bar{\mathcal{F}}_-$, are built by application of Lemma 17 and the pigeonhole principle. According to Lemma 17, we can ensure that there exist an index $i_0 \in [\![1; n]\!]$, $(\alpha, \beta) \in [\![1; M_\gamma - 1]\!] \times \left[\frac{1}{M_\gamma^2}, \frac{1}{2}\right]$ and two subsets $\bar{\mathcal{F}}_+$ and $\hat{\mathcal{F}}_-$ of $\bar{\mathcal{F}}$ verifying either $\left|\bar{\mathcal{F}}_+\right| \geqslant (1-\beta)\left|\bar{\mathcal{F}}\right|$ and $\left|\hat{\mathcal{F}}_-\right| \geqslant \max\left\{\frac{1}{2}\beta, \frac{1}{M_\gamma^2}\right\}\left|\bar{\mathcal{F}}\right|$ or vice versa, such that

$$\begin{cases} \forall f_+ \in \bar{\mathcal{F}}_+, \quad f_{+,\gamma}^{(\eta)}\left(z_{i_0}\right) \geqslant \alpha + 1 \\ \forall f_- \in \hat{\mathcal{F}}_-, \quad f_{-,\gamma}^{(\eta)}\left(z_{i_0}\right) \leqslant \alpha - 1 \end{cases}.$$

Setting $b_{i_0} = \alpha$, it springs from Lemma 14 that

$$\begin{cases} \forall f_+ \in \bar{\mathcal{F}}_+, \quad f_+^{(\eta)}\left(z_{i_0}\right) - b_{i_0} \geqslant 1 \\ \forall f_- \in \hat{\mathcal{F}}_-, \quad \max_{k \neq y_{i_0}} f_-^{(\eta)}\left(x_{i_0}, k\right) + b_{i_0} \geqslant 1 \end{cases},$$

i.e., (49) is obtained with $\bar{\mathcal{F}}_-$ replaced with $\hat{\mathcal{F}}_-$. There comes the application of the pigeonhole principle, to obtain Formula (66). The derivation of the corresponding function classes is as follows. There exists $c_{i_0} \in \mathcal{Y} \setminus \{y_{i_0}\}$ such that among the functions f_- in $\hat{\mathcal{F}}_-$, at least $\left\lceil \frac{\left|\hat{\mathcal{F}}_-\right|}{C-1} \right\rceil$ of them satisfy $c_{i_0} \in \operatorname{argmax}_{k \neq y_{i_0}} f_-\left(x_{i_0}, k\right)$. We choose $\bar{\mathcal{F}}_-$ to be any such subset of $\hat{\mathcal{F}}_-$. With Formula (66) at hand, Inequality (65) is also available. Thus, finishing the proof still boils down to deriving a lower bound on $\ell\left(\bar{\mathcal{F}}\right)$. The originality rests on the fact that two cases must be considered, to take into account the two sources of asymmetry between the cardinalities of $\bar{\mathcal{F}}_+$ and $\bar{\mathcal{F}}_-$. Indeed, we have either $\left|\bar{\mathcal{F}}_+\right| \geqslant \max\left\{\frac{1}{2}\beta, \frac{1}{M_\gamma^2}\right\}\left|\bar{\mathcal{F}}\right|$ and $\left|\bar{\mathcal{F}}_-\right| \geqslant \frac{1-\beta}{C-1}\left|\bar{\mathcal{F}}\right|$ or $\left|\bar{\mathcal{F}}_+\right| \geqslant (1-\beta)\left|\bar{\mathcal{F}}\right|$ and $\left|\bar{\mathcal{F}}_-\right| \geqslant \frac{1}{C-1}\max\left\{\frac{1}{2}\beta, \frac{1}{M_\gamma^2}\right\}\left|\bar{\mathcal{F}}\right|$. The

induction hypothesis is this time:

$$\ell\left(\bar{\mathcal{F}}\right) \geqslant \left|\bar{\mathcal{F}}\right|^{\frac{1}{\log_2\left(\frac{M_\gamma^2}{\sqrt{M_\gamma^2-2}}\sqrt{C-1}\right)}}. \tag{69}$$

Once more, it is obviously true for the leaves. We prove it for the first case (the other one is treated in the same way). The computation makes use of the following analytical result:

$$\forall K \in (0,1], \quad \operatorname*{argmin}_{\beta \in \left[\frac{1}{M_\gamma^2}, \frac{1}{2}\right]} \left\{ \left(\frac{1-\beta}{C-1}\right)^K + \left(\max\left\{\frac{\beta}{2}, \frac{1}{M_\gamma^2}\right\}\right)^K \right\} = \frac{2}{M_\gamma^2}.$$

Then,

$$\begin{aligned}
\ell\left(\bar{\mathcal{F}}\right) =& \ \ell\left(\bar{\mathcal{F}}_+\right) + \ell\left(\bar{\mathcal{F}}_-\right) \\
\geqslant& \left(\left(\frac{1}{M_\gamma^2}\right)^{\frac{1}{\log_2\left(\frac{M_\gamma^2}{\sqrt{M_\gamma^2-2}}\sqrt{C-1}\right)}} + \left(\frac{M_\gamma^2-2}{M_\gamma^2(C-1)}\right)^{\frac{1}{\log_2\left(\frac{M_\gamma^2}{\sqrt{M_\gamma^2-2}}\sqrt{C-1}\right)}} \right) \left|\bar{\mathcal{F}}\right|^{\frac{1}{\log_2\left(\frac{M_\gamma^2}{\sqrt{M_\gamma^2-2}}\sqrt{C-1}\right)}} \\
=& \ \frac{1}{2}\left(\left(\sqrt{\frac{C-1}{M_\gamma^2-2}}\right)^{\frac{1}{\log_2\left(\frac{M_\gamma^2}{\sqrt{M_\gamma^2-2}}\sqrt{C-1}\right)}} + \left(\sqrt{\frac{M_\gamma^2-2}{C-1}}\right)^{\frac{1}{\log_2\left(\frac{M_\gamma^2}{\sqrt{M_\gamma^2-2}}\sqrt{C-1}\right)}} \right) \\
& \times \left|\bar{\mathcal{F}}\right|^{\frac{1}{\log_2\left(\frac{M_\gamma^2}{\sqrt{M_\gamma^2-2}}\sqrt{C-1}\right)}} \\
\geqslant& \ \frac{1}{2}\min_{t\in\mathbb{R}_+^*}\left(t+\frac{1}{t}\right)\left|\bar{\mathcal{F}}\right|^{\frac{1}{\log_2\left(\frac{M_\gamma^2}{\sqrt{M_\gamma^2-2}}\sqrt{C-1}\right)}} \\
=& \ \left|\bar{\mathcal{F}}\right|^{\frac{1}{\log_2\left(\frac{M_\gamma^2}{\sqrt{M_\gamma^2-2}}\sqrt{C-1}\right)}}.
\end{aligned}$$

Combining Inequalities (64), (65) and (69) produces by transitivity:

$$\left|\tilde{\mathcal{F}}\right| \leqslant \left(\Sigma'\right)^{\log_2\left(\frac{M_\gamma^2}{\sqrt{M_\gamma^2-2}}\sqrt{C-1}\right)}.$$

The left-hand side inequality of Formula (68) has been established. To conclude the proof, it suffices to remember that $M_\gamma \geqslant 2$. □

The proof of Lemma 9 is the following one.

Proof. The beginning of the proof is identical to the beginning of the proof of Lemma 7 up to the use of Formula (41), which is now done with $N = 6$ (instead of $N = 5$). The reason

for this change is to satisfy the hypotheses of the basic combinatorial result, Lemma 21, which replaces Lemma 19.

$$\begin{aligned}\left|\tilde{\mathcal{F}}\right| &= \left|\tilde{\mathcal{F}}_{\gamma}^{\left(\frac{\epsilon}{14}\right)}\Big|_{s_{\mathbb{Z}^q}}\right| \\ &\leqslant \left(\sum_{u=0}^{d_N}\binom{q}{u}\left(\frac{14\gamma}{\epsilon}\right)^u (C-1)^u\right)^{\frac{1}{2}\log_2\left(2\left(\frac{14\gamma}{\epsilon}\right)^2(C-1)\right)} \\ &\leqslant \left(\frac{14\gamma\,(C-1)\,eq}{d_N\epsilon}\right)^{\frac{1}{2}\log_2\left(2\left(\frac{14\gamma}{\epsilon}\right)^2(C-1)\right)d_N}, \end{aligned} \tag{70}$$

where $d_N = \text{S-N-dim}\left(\tilde{\mathcal{F}}^{\left(\frac{\epsilon}{14}\right)}\right)$. A substitution of the upper bound on q provided by (59) into (70) gives:

$$\left|\tilde{\mathcal{F}}\right| \leqslant \left(K\,(C-1)\left(\frac{\gamma}{\epsilon}\right)^5 \frac{\ln\left(\left|\tilde{\mathcal{F}}\right|\right)}{d_N}\right)^{\frac{1}{2}\log_2\left(2\left(\frac{14\gamma}{\epsilon}\right)^2(C-1)\right)d_N}$$

with $K = \frac{1568}{3}e$. In order to upper bound $\ln\left(\left|\tilde{\mathcal{F}}\right|^{\frac{1}{d_N}}\right)$, we resort once more to (40), with $u_0 = \frac{1}{4}\log_2\left(2\left(\frac{14\gamma}{\epsilon}\right)^2(C-1)\right)$. Thus,

$$\ln\left(\left|\tilde{\mathcal{F}}\right|^{\frac{1}{d_N}}\right) \leqslant \frac{1}{2}\log_2\left(2\left(\frac{14\gamma}{\epsilon}\right)^2(C-1)\right)\left|\tilde{\mathcal{F}}\right|^{\frac{1}{\log_2\left(2\left(\frac{14\gamma}{\epsilon}\right)^2(C-1)\right)d_N}}$$

implies that

$$\left|\tilde{\mathcal{F}}\right| \leqslant \left(\frac{1}{2}\log_2\left(2\left(\frac{14\gamma}{\epsilon}\right)^2(C-1)\right)K\,(C-1)\left(\frac{\gamma}{\epsilon}\right)^5\right)^{\log_2\left(2\left(\frac{14\gamma}{\epsilon}\right)^2(C-1)\right)d_N}.$$

Since

$$\begin{aligned} 2\left(\frac{14\gamma}{\epsilon}\right)^2(C-1) > 16 &\Longrightarrow \log_2\left(2\left(\frac{14\gamma}{\epsilon}\right)^2(C-1)\right) < \left(2\left(\frac{14\gamma}{\epsilon}\right)^2(C-1)\right)^{\frac{1}{2}} \\ &\Longrightarrow \log_2\left(2\left(\frac{14\gamma}{\epsilon}\right)^2(C-1)\right) < \left(\frac{K}{2}(C-1)\left(\frac{\gamma}{\epsilon}\right)^5\right)^{\frac{1}{2}}, \end{aligned}$$

the upper bound on $\left|\tilde{\mathcal{F}}\right|$ simplifies into

$$\mathcal{M}\left(\epsilon, \mathcal{F}_\gamma, d_{2,\mathbf{z}_n}\right) = \left|\tilde{\mathcal{F}}\right| \leqslant \left((C-1)\left(\frac{4\gamma}{\epsilon}\right)^5\right)^{\frac{3}{2}\log_2\left(2\left(\frac{14\gamma}{\epsilon}\right)^2(C-1)\right)d_N}. \tag{71}$$

Since $\tilde{\mathcal{F}}^{\left(\frac{\epsilon}{14}\right)} \subset \mathcal{F}^{\left(\frac{\epsilon}{14}\right)}$, by application of Formula (45),

$$\begin{aligned} \text{S-N-dim}\left(\tilde{\mathcal{F}}^{\left(\frac{\epsilon}{14}\right)}\right) &\leqslant \text{S-N-dim}\left(\mathcal{F}^{\left(\frac{\epsilon}{14}\right)}\right) \\ &\leqslant d_N\left(\frac{\epsilon}{28}\right). \end{aligned} \tag{72}$$

A substitution of (72) into (71) produces an upper bound on $\mathcal{M}\left(\epsilon, \mathcal{F}_\gamma, d_{2,\mathbf{z}_n}\right)$ which does not depend on $\mathbf{z}_n$, thus concluding the proof. □

C Proofs of the Structural Results

This appendix gathers the proofs of the upper bounds on γ-N-dim $(\rho_{\mathcal{G}})$. The proof of the first of them, Lemma 10, makes use of three partial results which are now stated.

C.1 Technical Lemmas

Proposition 2. *Let $\mathcal{F}$ be a function class defined as in Definition 8. Suppose that for $\gamma \in \mathbb{R}_+^*$, the subset $\bar{\mathcal{F}}$ of $\mathcal{F}$ γ-N-shatters the triplet $\left(\{(x_i, y_i) : 1 \leqslant i \leqslant n\}, \mathbf{b}_n, \mathbf{c}_n\right)$. Then $\bar{\mathcal{F}}$ also γ-N-shatters another triplet, $\left(\{(x_i, y_i') : 1 \leqslant i \leqslant n\}, \mathbf{b}_n', \mathbf{c}_n'\right)$, derived from the first one as follows:*

$$\forall i \in [\![1; n]\!], \begin{cases} \text{if } y_i < c_i,\ y_i' = y_i,\ b_i' = b_i,\ c_i' = c_i \\ \text{if } y_i > c_i,\ y_i' = c_i,\ b_i' = -b_i,\ c_i' = y_i \end{cases}.$$

As a consequence, the derivation of an upper bound on γ-N-dim$(\mathcal{F})$ can make use of a stronger hypothesis on $(\mathbf{y}_n, \mathbf{c}_n)$: $\forall i \in [\![1; n]\!]$, $y_i < c_i$, provided that the hypothesis of non-negativity of the biases b_i is relaxed.

Proof. Without loss of generality, we assume that $\bar{\mathcal{F}}$ is of minimal cardinality 2^n and set accordingly $\bar{\mathcal{F}} = \{f_{\mathbf{s}_n} : \mathbf{s}_n \in \{-1, 1\}^n\}$. Consider the following bijection on $\{-1, 1\}^n$:

$$\begin{array}{rccc} B: & \{-1,1\}^n & \longrightarrow & \{-1,1\}^n \\ & \mathbf{s}_n & \mapsto & \mathbf{s}_n' \end{array}$$

$$\forall i \in [\![1; n]\!], \begin{cases} \text{if } y_i < c_i,\ s_i' = s_i \\ \text{if } y_i > c_i,\ s_i' = -s_i \end{cases}.$$

Then,

$$\forall i \in [\![1; n]\!], \begin{cases} \text{if } s_i' = 1,\ f_{\mathbf{s}_n'}(x_i, y_i) - b_i \geqslant \gamma \\ \text{if } s_i' = -1,\ f_{\mathbf{s}_n'}(x_i, c_i) + b_i \geqslant \gamma \end{cases} \Longrightarrow \forall i \in [\![1; n]\!], \begin{cases} \text{if } s_i = 1,\ f_{\mathbf{s}_n'}(x_i, y_i') - b_i' \geqslant \gamma \\ \text{if } s_i = -1,\ f_{\mathbf{s}_n'}(x_i, c_i') + b_i' \geqslant \gamma \end{cases}.$$

By definition, the triplet $(s'_{\mathcal{Z}^n}, \mathbf{b}'_n, \mathbf{c}'_n)$ is γ-N-shattered by $\bar{\mathcal{F}}$, which concludes the proof. $\square$

Proposition 3 is an extension of Proposition 1.4 in Talagrand (2003) holding for the L_p-norms with $p \in \mathbb{N} \setminus \{0,1\}$ (instead of simply $p = 2$), that explicits the value of the constant.

Proposition 3. *Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{T}$. For every $\gamma \in \mathbb{R}_+^*$ satisfying γ-dim $(\mathcal{F}) > 0$, $n \in [\![1; \gamma\text{-dim}\,(\mathcal{F})]\!]$ and $p \in \mathbb{N} \setminus \{0,1\}$,*

$$n \leqslant K_p \log_2 \left(\mathcal{M}_p\left(\gamma, \mathcal{F}, n\right)\right)$$

with $K_p = \left(\frac{2^p}{2^{p-1}-1}\right)^2$.

Proof. Suppose that for $\gamma \in \mathbb{R}_+^*$, the subset $s_{\mathcal{T}^n} = \{t_i : 1 \leqslant i \leqslant n\}$ of $\mathcal{T}$ is γ-shattered by $\mathcal{F}$ and $\mathbf{b}_n = (b_i)_{1 \leqslant i \leqslant n} \in \mathbb{R}^n$ is a witness to this shattering. By definition, there exists a subset $\bar{\mathcal{F}} = \{f_{\mathbf{s}_n} :\ \mathbf{s}_n \in \{-1,1\}^n\}$ of $\mathcal{F}$ satisfying

$$\forall \mathbf{s}_n \in \{-1,1\}^n\,,\ \forall i \in [\![1;n]\!],\ \ s_i\left(f_{\mathbf{s}_n}\left(t_i\right) - b_i\right) \geqslant \gamma. \tag{73}$$

Let $\mathbf{t}_n = (t_i)_{1 \leqslant i \leqslant n}$. To prove the proposition, it suffices to establish that

$$n \leqslant K_p \log_2 \left(\mathcal{M}\left(\gamma, \bar{\mathcal{F}}, d_{p,\mathbf{t}_n}\right)\right). \tag{74}$$

For $(\mathbf{s}_n, \mathbf{s}'_n) \in \left(\{-1,1\}^n\right)^2$, let $\mathcal{S}\left(\mathbf{s}_n, \mathbf{s}'_n\right)$ be the subset of $[\![1;n]\!]$ defined by:

$$\mathcal{S}\left(\mathbf{s}_n, \mathbf{s}'_n\right) = \left\{i \in [\![1;n]\!] :\ s_i \neq s'_i\right\}.$$

Then, making use of (73), we obtain that

$$\begin{aligned} d_{p,\mathbf{t}_n}\left(f_{\mathbf{s}_n}, f_{\mathbf{s}'_n}\right) &\geqslant \left(\frac{1}{n}\left|\mathcal{S}\left(\mathbf{s}_n, \mathbf{s}'_n\right)\right|(2\gamma)^p\right)^{\frac{1}{p}} \\ &= 2\gamma\left(\frac{d_H\left(\mathbf{s}_n, \mathbf{s}'_n\right)}{n}\right)^{\frac{1}{p}}, \end{aligned}$$

where d_H stands for the Hamming distance. Thus, a sufficient condition for $d_{p,\mathbf{t}_n}\left(f_{\mathbf{s}_n}, f_{\mathbf{s}'_n}\right) \geqslant \gamma$ is $d_H\left(\mathbf{s}_n, \mathbf{s}'_n\right) \geqslant \left\lceil \left(\frac{1}{2}\right)^p n \right\rceil$. As a consequence, to prove (74), it suffices to establish that there is a subset of the set of vertices of the hypercube Q_n of cardinality $\left\lceil 2^{\frac{n}{K_p}} \right\rceil$ which is $\left\lceil \left(\frac{1}{2}\right)^p n \right\rceil$-separated with respect to the Hamming distance (the separation is well-defined since $\left\lceil 2^{\frac{n}{K_p}} \right\rceil \geqslant 2$). To that end, a probabilistic approach similar to that of the proof of Lemma 20 is implemented. For $q \in [\![2; 2^n]\!]$, let $\boldsymbol{\epsilon}_{q,n} = (\epsilon_{j,i})_{1 \leqslant j \leqslant q, 1 \leqslant i \leqslant n}$ be a Bernoulli

random matrix (its entries $\epsilon_{j,i}$ are independent Bernoulli random variables with common expectation $\frac{1}{2}$). Then, by application of the union bound,

$$\mathbb{P}\left(\exists\left(j,j'\right)\in[\![1;q]\!]^2 : 1\leqslant j<j'\leqslant q \text{ and } \sum_{i=1}^{n}\mathbb{1}_{\left\{\epsilon_{j,i}\neq\epsilon_{j',i}\right\}}<\left(\frac{1}{2}\right)^p n\right)$$
$$\leqslant\binom{q}{2}\mathbb{P}\left(\sum_{i=1}^{n}\epsilon_i>n\left(1-\left(\frac{1}{2}\right)^p\right)\right),$$

where $(\epsilon_i)_{1\leqslant i\leqslant n}$ is a Bernoulli random vector. To upper bound the tail probability on the right-hand side, we resort to Hoeffding's inequality, which gives

$$\mathbb{P}\left(\sum_{i=1}^{n}\epsilon_i-\frac{n}{2}>\frac{n}{2}\left(1-\left(\frac{1}{2}\right)^{p-1}\right)\right)\leqslant\exp\left(-\frac{n}{2}\left(1-\left(\frac{1}{2}\right)^{p-1}\right)^2\right).$$

By transitivity, this implies that a sufficient condition for

$$\mathbb{P}\left(\exists\left(j,j'\right)\in[\![1;q]\!]^2 : 1\leqslant j<j'\leqslant q \text{ and } \sum_{i=1}^{n}\mathbb{1}_{\left\{\epsilon_{j,i}\neq\epsilon_{j',i}\right\}}<\left(\frac{1}{2}\right)^p n\right)<1$$

is

$$\binom{q}{2}\exp\left(-\frac{n}{2}\left(1-\left(\frac{1}{2}\right)^{p-1}\right)^2\right)<1$$

and consequently

$$q\leqslant\left\lceil 2^{\frac{n}{K_p}}\right\rceil,$$

which is precisely the value announced and thus concludes the proof. □

The transition between covering and packing numbers is provided by a well-known equivalence.

Lemma 22. *Let $(\mathcal{E},\rho)$ be a pseudo-metric space. For every totally bounded set $\mathcal{E}'\subset\mathcal{E}$ and $\epsilon\in\mathbb{R}_+^*$, $\mathcal{M}\left(2\epsilon,\mathcal{E}',\rho\right)\leqslant\mathcal{N}^{int}\left(\epsilon,\mathcal{E}',\rho\right)\leqslant\mathcal{M}\left(\epsilon,\mathcal{E}',\rho\right)$.*

C.2 Margin Natarajan Dimension of $\rho_{\mathcal{G}}$

The proof of Lemma 10 is the following one.

Proof. Suppose that for $\gamma\in(0,M_{\mathcal{G}}]$, the triplet $(s_{\mathcal{Z}^n},\mathbf{b}_n,\mathbf{c}_n)$ is γ-N-shattered by $\rho_{\mathcal{G}}$. According to Proposition 2, in order to upper bound n, i.e., γ-N-dim $(\rho_{\mathcal{G}})$, one can assume that for every $i\in[\![1;n]\!]$, $y_i<c_i$, and the biases can be negative. Let $\bar{\mathcal{G}}=\{g^{\mathbf{s}_n}:\ \mathbf{s}_n\in\{-1,1\}^n\}$

be a subset of $\mathcal{G}$ (of minimal cardinality) such that $\rho_{\bar{\mathcal{G}}} = \{\rho_{g^{\mathbf{s}_n}} : \mathbf{s}_n \in \{-1,1\}^n\}$ γ-N-shatters $(s_{\mathcal{Z}^n}, \mathbf{b}_n, \mathbf{c}_n)$. For every pair $(k,l) \in [\![1;C]\!]^2$ satisfying $k < l$, let $\mathcal{S}_{k,l}$ be the subset of $[\![1;n]\!]$ defined as follows:

$$\mathcal{S}_{k,l} = \{i \in [\![1;n]\!] : \ y_i = k \text{ and } c_i = l\}$$

and let $n_{k,l} \in [\![0;n]\!]$ be its cardinality. By construction, $\mathcal{P} = \{\mathcal{S}_{k,l} : \ n_{k,l} > 0\}$ is a partition of $[\![1;n]\!]$. For every vector $\mathbf{s}_n = (s_i)_{1 \leqslant i \leqslant n} \in \{-1,1\}^n$, the function $g^{\mathbf{s}_n}$ satisfies:

$$\forall i \in [\![1;n]\!], \quad \begin{cases} \text{if } s_i = 1, \ \rho_{g^{\mathbf{s}_n}}(x_i, y_i) - b_i \geqslant \gamma \\ \text{if } s_i = -1, \ \rho_{g^{\mathbf{s}_n}}(x_i, c_i) + b_i \geqslant \gamma \end{cases}.$$

For a fixed $\mathcal{S}_{k,l} \in \mathcal{P}$, this implies that

$$\forall i \in \mathcal{S}_{k,l}, \ \ s_i \left(\frac{1}{2} \left(g_k^{\mathbf{s}_n}(x_i) - g_l^{\mathbf{s}_n}(x_i) \right) - b_i \right) \geqslant \gamma.$$

Let $\mathcal{D}_{\mathcal{G},k,l} = \left\{ \frac{1}{2}(g_k - g_l) : \ g \in \mathcal{G} \right\}$. By definition, we have established that its subset $\left\{ \frac{1}{2}\left(g_k^{\mathbf{s}_n} - g_l^{\mathbf{s}_n}\right) : \ \mathbf{s}_n \in \{-1,1\}^n \right\}$ γ-shatters a set of cardinality $n_{k,l}$, with the consequence that

$$n_{k,l} \leqslant \gamma\text{-dim}\left(\mathcal{D}_{\mathcal{G},k,l}\right).$$

Summing over all the elements of the partition $\mathcal{P}$ gives

$$\gamma\text{-N-dim}\left(\rho_{\mathcal{G}}\right) \leqslant \sum_{k=1}^{C-1} \sum_{l=k+1}^{C} \gamma\text{-dim}\left(\mathcal{D}_{\mathcal{G},k,l}\right). \tag{75}$$

Inequality (75) implies (15) since by definition, $\mathcal{D}_{\mathcal{G}} = \bigcup_{1 \leqslant k < l \leqslant C} \mathcal{D}_{\mathcal{G},k,l}$. The second upper bound on the γ-dimensions of the classes $\mathcal{D}_{\mathcal{G},k,l}$ is obtained by application of the standard strategy outlined in Section 3.1. Applying in sequence Proposition 3 and Lemma 22 (left-hand side inequality) gives:

$$\begin{aligned} \gamma\text{-dim}\left(\mathcal{D}_{\mathcal{G},k,l}\right) &\leqslant 16 \log_2 \left(\mathcal{M}_2 \left(\gamma, \mathcal{D}_{\mathcal{G},k,l}, \gamma\text{-dim}\left(\mathcal{D}_{\mathcal{G},k,l}\right) \right) \right) \\ &\leqslant 16 \log_2 \left(\mathcal{N}_2^{\text{int}} \left(\frac{\gamma}{2}, \mathcal{D}_{\mathcal{G},k,l}, \gamma\text{-dim}\left(\mathcal{D}_{\mathcal{G},k,l}\right) \right) \right). \end{aligned}$$

An optimization of the proof of Lemma 2 in the degenerate case $C = 2$ produces

$$\mathcal{N}_2^{\text{int}} \left(\frac{\gamma}{2}, \mathcal{D}_{\mathcal{G},k,l}, \gamma\text{-dim}\left(\mathcal{D}_{\mathcal{G},k,l}\right) \right) \leqslant \mathcal{N}_2^{\text{int}} \left(\frac{\gamma}{2}, \mathcal{G}_k, \gamma\text{-dim}\left(\mathcal{D}_{\mathcal{G},k,l}\right) \right) \times \mathcal{N}_2^{\text{int}} \left(\frac{\gamma}{2}, \mathcal{G}_l, \gamma\text{-dim}\left(\mathcal{D}_{\mathcal{G},k,l}\right) \right).$$

To upper bound the covering numbers in the right-hand side, it suffices to apply Lemma 22 (right-hand side inequality) and Theorem 1 in Mendelson and Vershynin (2003) (with the

optimized constants of Lemma 7). This produces:

$$\begin{aligned}\mathcal{N}_2^{\text{int}}\left(\frac{\gamma}{2}, \mathcal{G}_k, \gamma\text{-dim}\left(\mathcal{D}_{\mathcal{G},k,l}\right)\right) &\leqslant \mathcal{M}_2\left(\frac{\gamma}{2}, \mathcal{G}_k, \gamma\text{-dim}\left(\mathcal{D}_{\mathcal{G},k,l}\right)\right)\\ &\leqslant \left(\frac{20M_{\mathcal{G}}}{\gamma}\right)^{12d_k\left(\frac{\gamma}{48}\right)},\end{aligned}$$

where $d_k(\epsilon) = \epsilon\text{-dim}(\mathcal{G}_k)$. The second upper bound on the γ-dimensions of the classes $\mathcal{D}_{\mathcal{G},k,l}$ has been obtained. By substitution into (75), we get

$$\begin{aligned}\gamma\text{-N-dim}\left(\rho_{\mathcal{G}}\right) &\leqslant 192\left(C-1\right)\log_2\left(\frac{20M_{\mathcal{G}}}{\gamma}\right)\sum_{k=1}^{C}\left(\frac{\gamma}{48}\right)\text{-dim}\left(\mathcal{G}_k\right)\\ &\leqslant 384\binom{C}{2}\log_2\left(\frac{20M_{\mathcal{G}}}{\gamma}\right)\left(\frac{\gamma}{48}\right)\text{-dim}\left(\mathcal{G}_0\right).\end{aligned}$$

□

C.3 Margin Natarajan Dimension of $\rho_{\mathcal{H}_\Lambda}$

The proof of Lemma 11 reuses the notations of the proof of Lemma 10, with $\mathcal{G}$ set equal to $\mathcal{H}_\Lambda$.

Proof. By application of Khintchine inequality, every set $\mathcal{S}_{k,l}$ can be split into two subsets $\mathcal{S}_{k,l}^+$ and $\mathcal{S}_{k,l}^-$ such that

$$\left\|\sum_{i\in\mathcal{S}_{k,l}^+}\kappa_{x_i} - \sum_{i\in\mathcal{S}_{k,l}^-}\kappa_{x_i}\right\|_{\mathbf{H}_\kappa} \leqslant \sqrt{n_{k,l}}\Lambda_{\mathcal{X}}. \tag{76}$$

Let the vector $\mathbf{s}_n \in \{-1,1\}^n$ be defined as follows:

$$\forall i \in [\![1;n]\!],\ \begin{cases} i\in\mathcal{S}_{k,l}^+ \Longrightarrow s_i = 1\\ i\in\mathcal{S}_{k,l}^- \Longrightarrow s_i = -1\end{cases}.$$

For every $\mathcal{S}_{k,l}\in\mathcal{P}$ and $i\in\mathcal{S}_{k,l}$, applying the reproducing property gives

$$\begin{cases}\text{if } s_i = 1,\ \frac{1}{2}\left\langle h_k^{\mathbf{s}_n} - h_l^{\mathbf{s}_n}, \kappa_{x_i}\right\rangle_{\mathbf{H}_\kappa} - b_i \geqslant \gamma \text{ and } \frac{1}{2}\left\langle h_l^{-\mathbf{s}_n} - h_k^{-\mathbf{s}_n}, \kappa_{x_i}\right\rangle_{\mathbf{H}_\kappa} + b_i \geqslant \gamma\\ \text{if } s_i = -1,\ \frac{1}{2}\left\langle h_l^{\mathbf{s}_n} - h_k^{\mathbf{s}_n}, \kappa_{x_i}\right\rangle_{\mathbf{H}_\kappa} + b_i \geqslant \gamma \text{ and } \frac{1}{2}\left\langle h_k^{-\mathbf{s}_n} - h_l^{-\mathbf{s}_n}, \kappa_{x_i}\right\rangle_{\mathbf{H}_\kappa} - b_i \geqslant \gamma\end{cases}. \tag{77}$$

For every $k\in[\![1;C]\!]$, let $h_k^\delta = \frac{1}{2}\left(h_k^{\mathbf{s}_n} - h_k^{-\mathbf{s}_n}\right)$. Then (77) produces by summation:

$$\forall i\in\mathcal{S}_{k,l},\ \begin{cases} i\in\mathcal{S}_{k,l}^+ \Longrightarrow \frac{1}{2}\left\langle h_k^\delta - h_l^\delta, \kappa_{x_i}\right\rangle_{\mathbf{H}_\kappa} \geqslant \gamma\\ i\in\mathcal{S}_{k,l}^- \Longrightarrow \frac{1}{2}\left\langle h_k^\delta - h_l^\delta, -\kappa_{x_i}\right\rangle_{\mathbf{H}_\kappa} \geqslant \gamma\end{cases}. \tag{78}$$

By summation over $i \in \mathcal{S}_{k,l}$, it results from (78) that:

$$\frac{1}{2}\left\langle h_k^\delta - h_l^\delta, \sum_{i\in\mathcal{S}_{k,l}^+} \kappa_{x_i} - \sum_{i\in\mathcal{S}_{k,l}^-} \kappa_{x_i} \right\rangle_{\mathbf{H}_\kappa} \geqslant n_{k,l}\gamma. \tag{79}$$

Applying the Cauchy-Schwarz inequality to (79) yields

$$\frac{1}{2}\left\|h_k^\delta - h_l^\delta\right\|_{\mathbf{H}_\kappa} \left\| \sum_{i\in\mathcal{S}_{k,l}^+} \kappa_{x_i} - \sum_{i\in\mathcal{S}_{k,l}^-} \kappa_{x_i} \right\|_{\mathbf{H}_\kappa} \geqslant n_{k,l}\gamma. \tag{80}$$

By substitution of (76) into (80),

$$\forall \mathcal{S}_{k,l} \in \mathcal{P},\ \ n_{k,l} \leqslant \left(\frac{\frac{1}{2}\left\|h_k^\delta - h_l^\delta\right\|_{\mathbf{H}_\kappa} \Lambda_{\mathcal{X}}}{\gamma} \right)^2.$$

By summation over all the elements of the partition $\mathcal{P}$,

$$n \leqslant \left(\frac{\Lambda_{\mathcal{X}}}{2\gamma}\right)^2 \sum_{1\leqslant k<l\leqslant C} \left\|h_k^\delta - h_l^\delta\right\|_{\mathbf{H}_\kappa}^2. \tag{81}$$

To upper bound the sum in the right-hand side of (81), we first note that

$$\left\|h_k^\delta - h_l^\delta\right\|_{\mathbf{H}_\kappa}^2 \leqslant \frac{1}{2}\left(\left\|h_k^{\mathbf{s}_n} - h_l^{\mathbf{s}_n}\right\|_{\mathbf{H}_\kappa}^2 + \left\|h_k^{-\mathbf{s}_n} - h_l^{-\mathbf{s}_n}\right\|_{\mathbf{H}_\kappa}^2\right),$$

with the consequence that

$$n \leqslant \left(\frac{\Lambda_{\mathcal{X}}}{2\gamma}\right)^2 \frac{1}{2} \sum_{1\leqslant k<l\leqslant C} \left(\left\|h_k^{\mathbf{s}_n} - h_l^{\mathbf{s}_n}\right\|_{\mathbf{H}_\kappa}^2 + \left\|h_k^{-\mathbf{s}_n} - h_l^{-\mathbf{s}_n}\right\|_{\mathbf{H}_\kappa}^2\right). \tag{82}$$

Now, since by hypothesis, $\sum_{k=1}^C h_k = 0_{\mathbf{H}_\kappa}$,

$$\begin{aligned}
\sum_{1\leqslant k<l\leqslant C} \left\|h_k^{\mathbf{s}_n} - h_l^{\mathbf{s}_n}\right\|_{\mathbf{H}_\kappa}^2 &= (C-1)\sum_{k=1}^C \left\|h_k^{\mathbf{s}_n}\right\|_{\mathbf{H}_\kappa}^2 - 2\sum_{1\leqslant k<l\leqslant C} \left\langle h_k^{\mathbf{s}_n}, h_l^{\mathbf{s}_n}\right\rangle_{\mathbf{H}_\kappa} \\
&= C\sum_{k=1}^C \left\|h_k^{\mathbf{s}_n}\right\|_{\mathbf{H}_\kappa}^2 - \sum_{k=1}^C\sum_{l=1}^C \left\langle h_k^{\mathbf{s}_n}, h_l^{\mathbf{s}_n}\right\rangle_{\mathbf{H}_\kappa} \\
&= C\sum_{k=1}^C \left\|h_k^{\mathbf{s}_n}\right\|_{\mathbf{H}_\kappa}^2 - \left\|\sum_{k=1}^C h_k^{\mathbf{s}_n}\right\|_{\mathbf{H}_\kappa}^2 \\
&= C\sum_{k=1}^C \left\|h_k^{\mathbf{s}_n}\right\|_{\mathbf{H}_\kappa}^2 \\
&= C\left\|h^{\mathbf{s}_n}\right\|_{\mathbf{H}_{\kappa,C}}^2 \\
&\leqslant C\Lambda^2.
\end{aligned} \tag{83}$$

Obviously, (83) also provides an upper bonud on $\sum_{1\leqslant k<l\leqslant C} \left\|h_k^{-\mathbf{s}_n} - h_l^{-\mathbf{s}_n}\right\|_{\mathbf{H}_\kappa}^2$. Thus, a substitution into (82) concludes the proof. □

D Upper Bound on the Rademacher Complexity

The proof of Theorem 3 is the following one.

Proof. In all three cases, the starting point is Inequality (28).

First case: $d_{\mathcal{G},\gamma} \in (0,2)$

This case is the only one for which the entropy integral exists. Setting for every $j \in \mathbb{N}$, $h(j) = \gamma 2^{-\frac{2}{2-d_{\mathcal{G},\gamma}} j}$, we obtain

$$R_m(\rho_{\mathcal{G},\gamma}) \leqslant 24\left(1+2^{\frac{2}{2-d_{\mathcal{G},\gamma}}}\right)\sqrt{\frac{F_1(C)}{m}}\gamma^{1-\frac{d_{\mathcal{G},\gamma}}{2}}\int_0^{\frac{1}{2}}\ln\left((C-1)\left(4\epsilon^{-\frac{2}{2-d_{\mathcal{G},\gamma}}}\right)^5\right)d\epsilon.$$

Let $I(C)$ denote the integral. Then,

$$\begin{aligned} I(C) &= \int_0^{\frac{1}{2}}\ln\left((C-1)\left(4\epsilon^{-\frac{2}{2-d_{\mathcal{G},\gamma}}}\right)^5\right)d\epsilon \\ &= \frac{1}{2}\left(\ln\left((C-1)4^5\right)+10\frac{1+\ln(2)}{2-d_{\mathcal{G},\gamma}}\right). \end{aligned}$$

Second case: $d_{\mathcal{G},\gamma} = 2$

$$R_m(\rho_{\mathcal{G},\gamma}) \leqslant h(N) + \frac{8}{3}\sqrt{\frac{F_1(C)}{m}}\sum_{j\in\mathcal{J}}\frac{h(j)+h(j-1)}{h(j)}\log_2\left((C-1)\left(\frac{4\gamma}{h(j)}\right)^5\right).$$

We set $N = \left\lceil \log_2\left(\frac{\sqrt{m}}{\log_2(m)}\right)\right\rceil$ and $h(j) = \gamma\frac{\log_2(m)}{\sqrt{m}}2^{N-j}$. Note that $m \geqslant 2$ ensures that $N \geqslant 1$. Then,

$$\begin{aligned} R_m(\rho_{\mathcal{G},\gamma}) &\leqslant \gamma\frac{\log_2(m)}{\sqrt{m}} + 8\sqrt{\frac{F_1(C)}{m}}\sum_{j=1}^{N}\log_2\left((C-1)\left(\frac{4\gamma}{h(j)}\right)^5\right) \\ &\leqslant \gamma\frac{\log_2(m)}{\sqrt{m}} + 8\sqrt{\frac{F_1(C)}{m}}\left\lceil\log_2\left(\frac{\sqrt{m}}{\log_2(m)}\right)\right\rceil\log_2\left((C-1)\left(4\frac{\sqrt{m}}{\log_2(m)}\right)^5\right). \end{aligned}$$

Third case: $d_{\mathcal{G},\gamma} > 2$

For $N = \left\lceil \frac{d_{\mathcal{G},\gamma}-2}{2d_{\mathcal{G},\gamma}}\log_2\left(\frac{m}{\log_2(m)}\right)\right\rceil$, let us set $h(j) = \gamma\left(\frac{\log_2(m)}{m}\right)^{\frac{1}{d_{\mathcal{G},\gamma}}}2^{\frac{2}{d_{\mathcal{G},\gamma}-2}(-j+N)}$. We then get

$$R_m(\rho_{\mathcal{G},\gamma}) \leqslant \gamma\left(\frac{\log_2(m)}{m}\right)^{\frac{1}{d_{\mathcal{G},\gamma}}}\left[1+\frac{8}{3}\left(1+2^{\frac{2}{d_{\mathcal{G},\gamma}-2}}\right)\left(\frac{1}{\gamma}\right)^{\frac{d_{\mathcal{G},\gamma}}{2}}\sqrt{\frac{F_1(C)}{\log_2(m)}}S_N\right]$$

with

$$\begin{aligned} S_N &= \sum_{j=1}^{N} 2^{j-N} \log_2 \left((C-1) \left(\frac{4\gamma}{h(j)} \right)^5 \right) \\ &\leqslant \log_2 \left((C-1) \left(\frac{4\gamma}{h(N)} \right)^5 \right) \sum_{j=1}^{N} 2^{j-N} \\ &\leqslant 2 \log_2 \left((C-1) \left(4 \left(\frac{m}{\log_2(m)} \right)^{\frac{1}{d_{\mathcal{G},\gamma}}} \right)^5 \right). \end{aligned}$$

□