Generalization Performance of Margin Multi-category Classifiers

Khadija Musayeva

UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Generalization Performance of Margin Multi-category Classifiers

THÈSE

présentée et soutenue publiquement le 23 septembre 2019

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention informatique)

par

Khadija Musayeva

Composition du jury

Président :	Anne Boyer	Professeur, Université de Lorraine
Rapporteurs :	Sana Louhichi	Professeur, Université de Grenoble Alpes
	Younès Bennani	Professeur, Université Paris 13
Examinateurs :	Anne Boyer	Professeur, Université de Lorraine
	Myriam Maumy-Bertrand	Maître de Conférences, Université de Strasbourg
Directeurs de thèse :	Yann Guermeur	Directeur de Recherche CNRS, Nancy
	Fabien Lauer	Maître de Conférences, Université de Lorraine

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Mis en page avec la classe thesul.

À ma mère

Contents

Résumé

Cette thèse s'intéresse à la théorie de la discrimination multi-classe. Nos travaux ont un intérêt particulier dans le contexte de la classification extrême. Celle-ci est de plus en plus rencontrée à l'ère du "*big data*" et consiste à associer un objet à une (ou plusieurs, dans les tâches à étiquettes multiples) catégorie parmi des dizaines de milliers, voire des millions [2]. Ainsi, il devient essentiel d'établir des garanties sur les performances qui "passent à l'échelle" en ce qui concerne le nombre de catégories. Une autre application possible de nos travaux concerne la sélection de modèle. Dans cette thèse, nous nous concentrons sur la dérivation de garanties sur les performances sous des hypothèses minimales sur le modèle de prédiction aussi bien que les données.

Nos travaux sont basés sur la théorie statistique de l'apprentissage fondée par Vapnik et Chervonenkis à la fin des années soixante [88]. Cette théorie fournit un cadre pour l'analyse qualitative et quantitative des performances en généralisation des modèles prédictifs. Les modèles prédictifs considérés ici sont des classifieurs multi-classes à marge. Ceux-ci incluent la plupart des classifieurs les plus populaires comme les réseaux de neurones [4], les machines à vecteurs supports [20] ou les plus proches voisins [49]. Un classifieur multi-classe à marge implémente un ensemble de fonctions à valeurs vectorielles, avec une composante à valeur réelle par catégorie. Un tel classifieur associe un objet à la catégorie pour laquelle la sortie de la fonction composante correspondante est la plus grande. Les sorties à valeurs réelles permettent de déterminer la performance d'un classifieur à partir des différences de fonctions composantes. Pour ces classifieurs, nous nous intéressons aux bornes sur la probabilité de mauvais classement et à leurs dépendances au nombre de catégories C, à la taille de l'échantillon m et au paramètre de marge γ. Ces bornes reposent d'une part sur la performance empirique du classifieur et d'autre part sur la notion de *mesure de capacité*.

La mesure de capacité d'un classifieur contrôle le taux de convergence uniforme de la performance empirique vers la performance en généralisation [88]. Plus la capacité d'un classifieur est grande, plus la convergence sera lente. Dans cette thèse, nous travaillons avec des mesures de capacité sensibles à l'échelle : la complexité de Rademacher, les nombres de recouvrement et

les dimensions combinatoires. Ces mesures sont très liées : il est possible de borner les unes à partir des autres. En particulier, la technique du chaînage et les bornes combinatoires rendent possibles les liens suivants :

$$\text{complexité de Rademacher} \xrightarrow{\text{chaînage}} \text{nombres de recouvrement} \tag{1}$$

$$\text{nombre de recouvrement} \xrightarrow{\text{borne combinatoire}} \text{dimension combinatoire} \tag{2}$$

A partir de ces relations, il est aussi possible de lier la complexité de Rademacher à la dimension combinatoire en passant par les nombres de recouvrement. Une relation réciproque existe également.

Un type particulier de résultat, une *décomposition* de mesure de capacité, borne la capacité d'une classe composite à partir des capacités de ses classes composantes. Dans le cadre de la discrimination multi-classe, ce type de borne permet d'estimer une mesure de capacité multi-classe à partir d'un ensemble de capacités bi-classes, et ainsi de rendre explicite la dépendance au nombre de catégories. Les relations entre les mesures de capacité et leurs décompositions sont au cœur de la théorie des bornes sur l'erreur de généralisation dans ce contexte.

La performance en classification d'un modèle est évaluée à partir d'une fonction de perte. Que ce soit dans le cas binaire ou multi-classe, lorsque la fonction de perte est l'indicatrice de mauvais classement classique, la convergence uniforme est bien étudiée. Mais lorsque la fonction de perte possède une certaine régularité, comme la continuité lipschitzienne, certaines questions restent ouvertes. Dans cette thèse, nous nous concentrons sur les fonction de perte à marge lipschitziennes. Indépendamment de la fonction de perte utilisée, les bornes de généralisation de base impliquent deux types de mesures de capacité : les nombres de recouvrement et la complexité de Rademacher.

Les premiers résultats sur les nombres de recouvrement pour le problème classique de *Glivenko-Cantelli* remontent aux travaux de Pollard [68]. Bartlett et Long [12] ont amélioré le résultat de Pollard concernant les classes de fonctions à valeurs réelles et fourni l'estimation de la complexité de l'échantillon, c'est-à-dire la taille de l'échantillon suffisante pour que la convergence uniforme ait lieu. Dans le cas multi-classe, lorsque la fonction de perte est lipschitzienne, contrôler la déviation uniforme directement par un nombre de recouvrement est une question ouverte.

Contrairement aux nombres de recouvrement, la complexité de Rademacher fut introduite relativement récemment en théorie de l'apprentissage. Pour les classifieurs multi-classes à marge, Koltchinskii et Panchenko [47] fournissent une borne sur l'erreur dans laquelle ils décomposent la complexité de Rademacher pour obtenir une dépendance explicite au nombre de catégories. Kuznetsov et al. [53], ainsi que Maurer [57], améliorent la dépendance quadratique de ces travaux

en une dépendance linéaire. Plus précisément, dans le cas général où les fonctions composantes sont indépendantes, leur résultat de décomposition donne le schéma :

$$\text{complexité de Rademacher (multi-classe)} \xrightarrow{\text{décomposition}} \sum_{k=1}^{C} k\text{-ème complexité de Rademacher} \tag{3}$$

En fait, avec la complexité de Rademacher, il existe plusieurs options pour établir la dépendance à C, puisqu'à la vue du chemin

$$\text{complexité de Rademacher} \xrightarrow{\text{chaînage}} \text{nombres de recouvrement} \xrightarrow{\text{borne combinatoire}} \text{dimension combinatoire} \tag{4}$$

la décomposition peut être reléguée aux niveaux suivants.

Plusieurs résultats de décomposition existent pour les nombres de recouvrement. Les décompositions de Zhang [93] et Duan [23] concernent des nombres de recouvrement basés sur une métrique spécifique, alors que celle de Guermeur [37] est la généralisation de [23] à toute métrique L_p. Cette décomposition d'un nombre de recouvrement multi-classe conduit à un produit de nombres de recouvrement bi-classes :

$$\text{nombre de recouvrement (multi-classe)} \xrightarrow{\text{décomposition}} \prod_{k=1}^{C} k\text{-ème nombre de recouvrement} \tag{5}$$

Guermeur [37] a démontré qu'en combinant cette décomposition avec la borne combinatoire de Mendelson et Vershynin [61], il est possible d'améliorer la dépendance au nombre de catégories jusqu'à une dépendance sous-linéaire.

En suivant toujours le chemin (4), une autre possibilité consiste à décomposer au dernier niveau : celui de la dimension combinatoire. La dimension combinatoire d'intérêt dans cette thèse est la "*fat-shattering dimension*". Les résultats de décomposition pour cette mesure de capacité sont dus à Bartlett [7] et Duan [23]. Le résultat de Duan concerne la fat-shattering dimension d'un produit de classes de fonctions composantes indépendantes et son extension au cas multi-classe est directe. Cependant, l'application de ce résultat dans le chaînage conduit à une borne avec une dépendance (super) linéaire au nombre de catégories. En fait, la décomposition de la fat-shattering dimension dans le contexte de la discrimination multi-classe et son impact sur les bornes de généralisation ne sont pas encore bien étudiés.

Dans cette thèse, notre but est de répondre à ces questions ouvertes et d'améliorer les garanties existantes. En particulier, dans le cadre du chemin (4), nous cherchons à améliorer la dépendance au nombre de catégories par rapport aux résultats de l'état de l'art mentionnés ci-dessus. Une part importante de ce travail concerne l'analyse de l'équilibre qui existe entre les dépendances à C et à m. Les principales contributions sont basées sur le raisonnement suivant.

Les nombres de recouvrement sont des quantités liées aux espaces métriques. Ainsi, la manière dont nous bornons ces mesures de capacité avec une borne combinatoire, ainsi que les résultats de décomposition, dépendent de la métrique utilisée. Dans cette thèse, nous travaillons principalement avec le logarithme des nombres de recouvrement, correspondant à l'*entropie métrique*. Pour cette quantité, la relation (5) donne un schéma impliquant une somme :

$$\text{entropie métrique (multi-classe) à l'échelle } \epsilon \xrightarrow{\text{décomposition}} \sum_{k=1}^{C} k\text{-ème entropie métrique à l'échelle } \epsilon' \quad (6)$$

L'entropie métrique croît lorsque son échelle décroît. Sous différentes métriques, les échelles ϵ' des entropies métriques bi-classes sont altérées de différentes manières par rapport au nombre de catégories. Plus la métrique est "forte", plus l'échelle est grande, et donc plus l'entropie métrique est petite. En particulier, dans le cas extrême de la métrique L_∞, la dépendance à C de l'échelle ϵ' disparaît. Une fois décomposée, nous relions chaque entropie métrique bi-classe à la fat-shattering dimension via une borne combinatoire :

$$k\text{-ème entropie métrique à l'échelle } \epsilon' \xrightarrow{\text{borne combinatoire}} k\text{-ème fat-shattering dimension} \quad (7)$$

Pour chaque choix de métrique, nous avons une borne combinatoire différente. Ainsi, à ce niveau, le but est d'utiliser au mieux l'influence de la métrique sur la décomposition et la borne combinatoire. L'influence du choix de la métrique se propage ensuite au travers du chaînage pour impacter les dépendances à C, m et γ de la complexité de Rademacher dans le cadre du schéma (4).

Si la décomposition est reléguée au niveau de la fat-shattering dimension, alors elle peut être réalisée grâce à la relation réciproque : il est possible de borner la fat-shattering dimension en fonction des autres mesures de capacité. En fait, il est plus aisé de dériver ces relations réciproques puisqu'elles s'obtiennent simplement à partir des définitions des mesures de capacité (dans le cas de l'entropie métrique, cela est vrai pour une métrique spécifique, la métrique L_∞) :

$$\text{fat-shattering dimension} \xrightarrow{\text{définitions}} \begin{cases} \text{entropie métrique} \\ \text{complexité de Rademacher} \end{cases}$$

Dans le cas de l'entropie métrique, il est ensuite possible d'obtenir une décomposition de la fat-shattering dimension par application des schémas (6) et (7). En se basant sur cette chaîne d'inégalités, on remarque que le choix de la métrique impacte la décomposition de la fat-shattering dimension. Il est aussi possible de relier la fat-shattering dimension à la complexité de Rademacher en combinant leurs définitions. Cela permet par la suite d'utiliser des bornes efficaces sur la complexité de Rademacher pour certains classifieurs spécifiques. A la vue du

schéma (4), les résultats obtenus au niveau de la fat-shattering dimension se propagent via la borne combinatoire et le chaînage. Dans cette thèse, nous étudions tous ces cheminements et niveaux de décomposition en suivant le plan ci-dessous.

Dans le Chapitre 1, nous introduisons le cadre théorique de nos travaux. Plus précisément, nous décrivons le cadre probabiliste de l'apprentissage, donnons les définitions formelles des classifieurs multi-classes à marge, des fonctions de perte, des performances empiriques et en généralisation et des mesures de capacité discutées ci-dessus. Quelques connections entre les théorèmes limites classiques et ces mesures sont aussi présentées.

Le Chapitre 2 s'attache ensuite à borner la probabilité de déviation uniforme entre la performance empirique et celle en généralisation en fonction d'un nombre de recouvrement. La taille de l'échantillon permettant d'obtenir la convergence uniforme est aussi ici estimée. Ce travail correspond à l'extension au cas multi-classe des travaux de Bartlett et Long [12].

Le Chapitre 3 donne une revue de la littérature sur les bornes de généralisation impliquant une complexité de Rademacher. Les résultats de décomposition pour cette mesure de capacité y sont ici mis en valeur et discutés au regard de l'impact qu'ils ont sur les dépendances aux paramètres d'intérêt.

Le Chapitre 4 relie la complexité de Rademacher à l'entropie métrique et considère la décomposition de cette dernière. En particulier, nous étudions comment le choix de la métrique influence le chaînage en termes de dépendance aux paramètres d'intérêt. Nous montrons ici que dans le cas extrême de la métrique L_∞, il est possible d'améliorer la dépendance à C par rapport à l'état de l'art, cependant au dépend d'une légère détérioration de la dépendance à la taille de l'échantillon m. Ensuite, une nouvelle borne combinatoire est dérivée et nous montrons comment celle-ci permet d'améliorer la dépendance à C sans modifier celles à m et au paramètre de marge γ.

Le Chapitre 5 se concentre quant-à lui sur la décomposition au dernier niveau : celui de la fat-shattering dimension. Deux types de bornes sur cette mesure de capacité sont obtenus. En premier lieu, nous relions la fat-shattering dimension à la complexité de Rademacher et discutons les bornes sur cette dernière dédiées à certaines familles de classifieurs spécifiques. Dans un second temps, nous décomposons la fat-shattering dimension en passant par l'entropie métrique. Une nouvelle borne combinatoire est dérivée à partir de cette décomposition pour permettre d'améliorer encore la dépendance à C par rapport aux résultats du Chapitre 4 basés sur une décomposition au niveau des entropies métriques. Cependant, cette amélioration conduit à une légère détérioration de la dépendance à m.

Le dernier chapitre résume l'ensemble de nos travaux et conclue en particulier que dans

le contexte du schéma (4), il existe une interaction entre les dépendances à C et à m. Bien qu'il soit possible d'amliorer progressivement la dépendance à C en décomposant au niveau des entropies métriques ou de la fat-shattering dimension, cela implique une légère baisse du taux de convergence par rapport à m. Enfin, la thèse se conclut par quelques directions de recherche envisageables à la suite de ces travaux.

Introduction

This thesis deals with the theory of multi-category pattern classification. Our work is especially relevant in the context of extreme classification. Extreme classification is a phenomenon of the big data era where the goal is to assign an object to one (or to multiple, in the multi-labelling tasks) of tens of thousands, possibly millions of categories [2]. Thus, it is of essential interest to derive performance guarantees that scale well with the number of categories. Another possible application of our work is in the model selection procedures. In the present thesis, we focus on deriving performance guarantees under minimal assumptions regarding the predictive model, as well as, the data.

Our work is based on the statistical learning theory founded by Vapnik and Chervonenkis in the late sixties [88]. This theory provides a framework for qualitative and quantitave analysis of generalization performance of predictive models. The predictive models considered here are multi-category margin classifiers. These include most well-known classifiers such as neural networks [4], support-vector machines [20] and nearest neighbours [49]. A multi-category margin classifier implements a set of vector-valued functions, with one real-valued component per category. Such a classifier assigns a pattern to the category for which the output of the corresponding function is the highest. The real-valued outputs allow one to assess the classification performance of a classifier based on the differences of component functions. For these classifiers, we are interested in the bounds on the probability of misclassification with explicit dependencies on the number C of categories, the sample size m and the margin parameter γ. These bounds rely on the empirical performance of a classifier as well as on the notion of *capacity measure*.

The capacity measure of a classifier controls the rate of uniform convergence of the empirical performance to the generalization one [88]. The higher the capacity of the classifier, the slower the convergence. In this thesis, we deal with scale-sensitive capacity measures and consider the following ones: the Rademacher complexity, the covering numbers and combinatorial dimensions. These measures are closely related: we can bound one in terms of another. Particularly, the

following relationships are possible thanks to the chaining method and a combinatorial bound:

$$\text{Rademacher complexity} \xrightarrow{\text{chaining}} \text{covering numbers} \tag{1}$$

$$\text{covering number} \xrightarrow{\text{combinatorial bound}} \text{combinatorial dimension} \tag{2}$$

From these relationships one can see that the Rademacher complexity can be related to the combinatorial dimension through the covering number. The converse relationships also exist.

A specific kind of result, a *decomposition* of capacity measure, upper bounds the capacity of a composite class in terms of that of component classes. In the multi-category classification setting, this kind of bound allows one to estimate the multi-class capacity measure from a set of bi-class ones thus making explicit the dependency on the number of classes. The relationships between the capacity measures as well as their decompositions are at the core of the theory of error bounds.

The classification performance of a model is assessed based on the use of a loss function. Be it in the binary or multi-category classification setting, when the loss function used is the classical indicator loss function, the uniform convergence problem is well studied. On the other hand, when the loss function possesses some regularity, for instance, the *Lipschitz continuity*, there still remain open questions. In this thesis, we focus on the Lipschitz continuous *margin loss* functions. Irrespective of the loss function used, the basic generalization bounds involve two types of capacity measures: the covering number and the Rademacher complexity.

Covering number result for the classical *Glivenko-Cantelli* problem can be traced back to the work of Pollard [68]. Bartlett and Long [12] improved the covering number result of Pollard concerning the classes of real-valued functions, and provided the sample complexity estimate, i.e., the sample size sufficient for the uniform convergence to take place. In the multi-category case, when the loss function used is a Lipschitz continuous margin loss function, controlling the uniform deviation directly by a covering number is an open question.

Unlike the covering number, the Rademacher complexity has been introduced into the learning theory relatively recently. For multi-category margin classifiers, Koltchinskii and Panchenko [47] provide an error bound where they decompose the Rademacher complexity to obtain an explicit dependency on the number of categories. Kuznetsov et al. [53], and Maurer [57] improve the quadratic dependency of the previous authors to a linear one. Indeed, in the general case with independent component functions, their decomposition result yields:

$$\text{(multi-class) Rademacher complexity} \xrightarrow{\text{decomposition}} \sum_{k=1}^{C} k\text{-th Rademacher complexity}. \tag{3}$$

In fact, with the Rademacher complexity one has several options to elaborate the dependency on C, since, in view of the pathway

$$\text{Rademacher complexity} \xrightarrow{\text{chaining}} \text{covering numbers} \xrightarrow{\text{combinatorial bound}} \text{combinatorial dimension}, \tag{4}$$

the decomposition could be postponed to the subsequent levels.

Several decomposition results exist for covering numbers. The decompositions of Zhang [93] and Duan [23] concern covering numbers with specific metrics, while that of Guermeur [37] is the generalization of [23] to all L_p-metrics. This decomposition of a multi-class covering number produces a product of bi-class ones:

$$\text{(multi-class) covering number} \xrightarrow{\text{decomposition}} \prod_{k=1}^{C} k\text{-th covering number}. \tag{5}$$

Guermeur [37] demonstrated that combining this decomposition with the combinatorial bound of Mendelson and Vershynin [61], one can improve the dependency on the number of classes to a sub-linear one.

Still following the pathway (4), another possibility is to postpone the decomposition to the last level: that of a combinatorial dimension. The combinatorial dimension of interest in this thesis is the fat-shattering dimension. The decomposition results for this capacity measure are due to Bartlett [7] and Duan [23]. Duan's result concerns the fat-shattering dimension of a product of independent function classes and its extension to the multi-class setting is straightforward. However, the application of this result in the chaining leads to a bound with a (super) linear dependency on the number of categories. In fact, the decomposition of the fat-shattering dimension in the context of multi-category classification and its impact on the error bounds are not well studied.

In this thesis, our goal is to fill in these gaps, and provide improved learning guarantees. Particularly, in the context of the pathway (4), our goal is to improve the dependency on the number of categories over the aforementioned state-of-the-art results. An important part of this work is the analysis of the trade-off that exists between the dependencies on C and m. The main contributions are based on the following reasoning.

The covering number is a quantity related to metric spaces. Thus, the way we upper bound this capacity measure, i.e., a combinatorial bound, as well as the decomposition result depend on the metric used. In this thesis, we mainly work with the logarithm of the covering number which is the *metric entropy*. For this quantity, the relationship (5) turns into the following schema

involving a sum:

$$\text{(multi-class) entropy at scale } \epsilon \xrightarrow{\text{decomposition}} \sum_{k=1}^{C} k\text{-th entropy at scale } \epsilon'. \tag{6}$$

The metric entropy grows as its scale decreases. Under different metrics, the scales ϵ' of bi-class metric entropies are altered in different ways with respect to the number of categories. The "stronger" the metric the larger the scale of the metric entropy, consequently, the smaller the metric entropy. Particularly, in the extreme case of the L_∞-metric, the dependency on C in the scales ϵ vanishes. Once decomposed, we relate each bi-class metric entropy to the fat-shattering dimension through a combinatorial bound:

$$k\text{-th entropy at scale } \epsilon' \xrightarrow{\text{combinatorial bound}} k\text{-th fat-shattering dimension} \tag{7}$$

Now, for each choice of metric we have a different combinatorial bound. Thus, at this level, our aim is to make the best use of the influence of the metric on the decomposition and the combinatorial bound. The result of this choice then propagates through the chaining bound affecting the dependencies on C, m and γ of the Rademacher complexity as per schema 4.

If we postpone the decomposition to the level of the fat-shattering dimension, then, it can be realized thanks to the following converse relationships: one can upper bound the fat-shattering dimension in terms of the other two measures. In fact, deriving the converse relationships is simpler because they are obtained based on the definitions of these capacity measures (in the case of the metric entropy it is true for a specific metric, the L_∞-metric):

$$\text{fat-shattering dimension} \xrightarrow{\text{definitions}} \begin{cases} \text{metric entropy} \\ \text{Rademacher complexity} \end{cases}$$

In the case of the metric entropy, one can obtain a decomposition of the fat-shattering dimension following the schemas (6) and (7). Based on this chain of bounds, one can see that the choice of the metric influences the decomposition of the fat-shattering dimension. Similarly, one can relate the fat-shattering dimension to a Rademacher complexity based on the interplay between their definitions. We can then make use of the efficient upper-bounds on the Rademacher complexity for specific classifiers. In view of schema (4), the results obtained at the level of the fat-shattering dimension then propagates through the combinatorial bound and the chaining method. In this thesis, we study all these pathways and the levels of decomposition with the following outline.

In Chapter 1, we introduce the theoretical framework of our work. More precisely, we describe the probabilistic setting of learning, give formal definitions of margin multi-category classifiers,

loss function, empirical and generalization performances. We outline the connections between the classical limit theorems and the capacity measures.

In Chapter 2, we upper bound the probability of the uniform deviation between the empirical performance and generalization one in terms of a covering number. We then estimate the sample size required for the uniform convergence. This work corresponds to the multi-class generalization of that of Bartlett and Long [12].

In Chapter 3, we give a literature review of generalization bounds involving a Rademacher complexity. We highlight the decomposition results for this capacity measure and discuss the impact they have on the dependencies on the basic parameters.

Chapter 4 relates the Rademacher complexity to the metric entropy and considers the decomposition of the latter capacity measure. Particularly, we study how the choice of the metric influences the chaining method in terms of the dependencies on the basic parameters. We show that in the extreme case of the L_∞-metric, one can obtain a better than the state-of-the-art dependency on C, by slightly worsening that on the sample size m. Then, we derive a new combinatorial bound which, when applied in the chaining, improves the dependency on C over the state of the art, while not worsening those on m and the margin parameter γ.

In Chapter 5, we focus on the decomposition at the last level: that of the fat-shattering dimension. We obtain two kinds of upper bounds on this capacity measure. First, we relate the fat-shattering dimension to the Rademacher complexity and focus on the upper bounds on this capacity measure for specific families of classifiers. Second, we decompose the fat-shattering dimension by relating it to the metric entropy and estimating the last quantity. We derive a new combinatorial bound based on this decomposition of the fat-shattering dimension. We demonstrate that when applied in the chaining, this result improves the dependency on C over those obtained by the decomposition at the level of the metric entropy (the results of Chapter 4). However, this is achieved at the cost of slightly deteriorating the dependency on m.

The final chapter summarizes our work. Particularly, this work concludes that in the context of schema (4), there is an interaction between the dependencies on C and m: although one can progressively improve the dependency on C through the decomposition of the metric entropy or the fat-shattering dimension, this slightly deteriorates the one on the sample size. We conclude the thesis by giving possible directions for future research.

Chapter 1

Preliminaries

This chapter deals with the mathematical framework of the present thesis. In Section 1.1, we introduce the notations used throughout. Section 1.2 describes the probabilistic setting. In Section 1.3, we give the formal definitions of margin multi-category classifiers, loss functions, empirical and generalization performances. The capacity measures used in this work are introduced in Section 1.4. Section 1.5 discusses the crucial relations between the statistical limit theorems and capacity measures.

1.1 Notation

$\mathbb{R}_+$ denotes the set of strictly positive reals, and $\mathbb{N}^*$ the set $\mathbb{N} \setminus \{0\}$. For any $t = (t_i)_{1 \leqslant i \leqslant d} \in \mathbb{R}^d$, $\|t\|_p = \left(\sum_{i=1}^d |t_i|^p \right)^{\frac{1}{p}}$ with $p \in \mathbb{R}_+$, and $\|t\|_\infty = \max_{1 \leqslant i \leqslant d} |t_i|$. l_2 is the space of all sequences $t = (t_i)_{i \geqslant 1}$, $t_i \in \mathbb{R}$, such that $\sum_i |t_i|^2 < \infty$. $[\![i, j]\!]$ stands for the set of integers from i to j. $\mathbb{1}_A$ is the indicator function of an event A such that $\mathbb{1}_A = 1$ if A occurs and 0 otherwise. δ_{ij} denotes the Kronecker delta function such that $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. $\lfloor t \rfloor$ is the greatest integer less than or equal to t, $\lceil t \rceil$ is the smallest integer greater than or equal to t. $h \circ f$ denotes the composition of functions h and f. We distinguish the sample size m from the generic notation n which stands for a number of points in a set that need not be a realization of a random sample.

1.2 Probabilistic Framework

We denote the description space by $\mathcal{X}$, and $\mathcal{Y} = [\![1, C]\!]$ with $C > 2$ stands for the finite set of categories. We assume that $\mathcal{X}$ is a Polish space, that is, a separable completely metrizable topological space. Let $\mathcal{A}_\mathcal{X}$ and $\mathcal{A}_\mathcal{Y}$ be σ-algebras on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let $(\mathcal{X} \times \mathcal{Y}, \mathcal{A}_\mathcal{X} \otimes \mathcal{A}_\mathcal{Y})$ be a measurable product space, and let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Assume that $Z = (X, Y) \in \mathcal{Z}$ is a ran-

dom pair with the law P on $\mathcal{A}_\mathcal{X} \otimes \mathcal{A}_\mathcal{Y}$. P completely characterizes the pattern recognition problem. In the context of learning, the only available information about P is in an m-sample $\mathbf{Z}_m = (Z_i)_{1 \leqslant i \leqslant m} = ((X_i, Y_i))_{1 \leqslant i \leqslant m}$, a sequence of m independent copies of Z (distributed according to P^m).

In the sequel, $(\mathcal{T}, \mathcal{A}_\mathcal{T})$ denotes the generic measurable space with probability measure $P_\mathcal{T}$ on it. The empirical measure associated with $P_\mathcal{T}$ is defined as follows.

Definition 1 (Empirical measure) *For any $t \in \mathcal{T}$, let δ_t be the Dirac measure on $\mathcal{A}_\mathcal{T}$ such that*

$$\forall A \in \mathcal{A}_\mathcal{T}, \quad \delta_t(A) = \begin{cases} 1, \text{if } t \in A; \\ 0, \text{otherwise.} \end{cases}$$

Let a random variable T be an identity map on $(\mathcal{T}, \mathcal{A}_\mathcal{T}, P_\mathcal{T})$. Let $(T_i)_{1 \leqslant i \leqslant n}$ be a sequence of n copies of T. The empirical measure P_n is the linear combination of Dirac measures concentrated on T_i:

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{T_i}.$$

To sidestep the complications that might arise from the measurability of a supremum of an uncountable set, we need the following property.

Definition 2 (Image admissible Suslin [24]) *A measurable space $(\mathcal{T}, \mathcal{A}_\mathcal{T})$ is called a* Suslin *space if there is a Borel-measurable map from some Polish space onto $\mathcal{T}$. A set $\mathcal{F}$ of real-valued functions on $(\mathcal{T}, \mathcal{A}_\mathcal{T})$ is* image admissible Suslin *if $(\mathcal{T}, \mathcal{A}_\mathcal{T})$ is a Suslin space and there exists a Suslin space $(\mathcal{F}', \mathcal{A}_\mathcal{F})$ and a mapping F from this space onto $\mathcal{F}$ such that $(t, f') \mapsto F(f')(t)$ is measurable on $(\mathcal{T} \times \mathcal{F}', \mathcal{A}_\mathcal{T} \otimes \mathcal{A}_\mathcal{F})$.*

In the following, $\mathcal{F}$ stands for a set of $(\mathcal{A}_\mathcal{T}, \mathcal{B}_\mathbb{R})$-measurable functions $f : \mathcal{T} \to \mathbb{R}$:

$$\forall B \in \mathcal{B}_\mathbb{R}, \quad f^{-1}[B] = \{t \in \mathcal{T} \ : \ f(t) \in B\} \in \mathcal{A}_\mathcal{T}.$$

We assume that $\mathcal{F}$ and all classes of real-valued functions considered in the thesis are image admissible Suslin.

1.3 Margin Multi-category Classifiers

We consider margin multi-category classifiers that take their decisions based on a score per category and focus on those that implement classes of functions with values in a hypercube of

$\mathbb{R}^C$. Most well-known classifiers, such as neural networks [4, 10], multi-category support vector machines [35, 22], and nearest neighbors [49] are margin multi-category classifiers, since they satisfy the following definition.

Definition 3 (Margin multi-category classifiers [37]) *Let $\mathcal{G} = \prod_{k=1}^{C} \mathcal{G}_k$ be a class of functions from $\mathcal{X}$ to $[-M_{\mathcal{G}}, M_{\mathcal{G}}]^C$ with $M_{\mathcal{G}} \in [1, +\infty)$. For each $g = (g_k)_{1 \leqslant k \leqslant C} \in \mathcal{G}$, dr_g is a* multi-category margin classifier *such that for all $x \in \mathcal{X}$, $dr_g(x) = \operatorname{argmax}_{1 \leqslant k \leqslant C} g_k(x)$, breaking ties with a dummy category $*$.*

Given $g \in \mathcal{G}$, dr_g misclassifies (x, y) if $dr_g(x) \neq y$. The goal of the learning process is to minimize the probability of misclassification $P\left(dr_g(X) \neq Y\right)$ over $\mathcal{G}$. To characterize the classification performance of margin classifiers, we introduce the following functions that make use of the real-valued outputs of the component functions of $g \in \mathcal{G}$. This definition could be traced back to the work of Koltchinskii and Panchenko [47].

Definition 4 (Class $\mathcal{F}_{\mathcal{G}}$ of margin functions) *Let $\mathcal{G}$ be as in Definition 3. For any $g \in \mathcal{G}$, the margin function $f_g : \mathcal{Z} \to [-M_{\mathcal{G}}, M_{\mathcal{G}}]$ is defined as follows:*

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

We define $\mathcal{F}_{\mathcal{G}} = \{f_g : g \in \mathcal{G}\}$.

Then, the misclassification probability is equal to $P\left(f_g(X, Y) \leqslant 0\right)$. Using the standard $\{0, 1\}$-loss function the goal of learning can be reformulated based on the following definition.

Definition 5 (Risk L) *Let $\mathcal{G}$ be as in Definition 3. Let $\phi : \mathbb{R} \to \{0, 1\}$ be the standard indicator loss function defined as $\phi(t) = \mathbb{1}_{\{t \leqslant 0\}}$. For any $g \in \mathcal{G}$, its risk $L(g)$ is*

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

Clearly, to minimize the risk over $\mathcal{G}$ requires the knowledge of the distribution P. As P is unknown, one way to address this problem is to use the following simple induction principle [87]: instead of the risk L, minimize its empirical version L_m evaluated on the basis of $\mathbf{Z}_m$. We obtain L_m by replacing the law P in Definition 5 with its empirical counterpart P_m.

Definition 6 (Empirical risk L_m) *Let $\mathcal{G}$ be as in Definition 3 and ϕ be as in Definition 5. Let P_m be the empirical measure supported on $\mathbf{Z}_m$. Then, the empirical risk of any $g \in \mathcal{G}$ is defined as:*

$$L_m(g) = \int_{\mathcal{Z}} \phi\left(f_g(z)\right) dP_m(z) = \frac{1}{m} \sum_{i=1}^{m} \phi\left(f_g(Z_i)\right).$$

Note that the standard indicator loss function has no sensitivity to the values of f_g except for their signs. The loss functions introduced below, on the other hand, allow us to capture the nature of margin classification, and derive generalization bounds characterized through the margin parameter.

Definition 7 (Margin loss functions) *For any $\gamma \in (0,1]$, define the margin indicator loss function $\bar{\phi}_\gamma : \mathbb{R} \to \{0,1\}$ as*

$$\bar{\phi}_\gamma(t) = \mathbb{1}_{\{t \leqslant \gamma\}},$$

and the truncated hinge loss function $\phi_\gamma : \mathbb{R} \to [0,1]$ as

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

Clearly, for a fixed $\gamma \in (0,1]$, $\bar{\phi}_\gamma$ dominates ϕ_γ which in its turn dominates the standard loss function. Moreover, ϕ_γ is Lipschitz continuous with constant $\frac{1}{\gamma}$. We observe that when ϕ_γ is applied to f_g, the values of the latter strictly above γ and below zero become irrelevant to the estimation of the classification accuracy. Taking benefit from this fact, we introduce truncated margin functions $f_{g,\gamma}$ by restricting the codomain of f_g to $[0,\gamma]$ for all $g \in \mathcal{G}$.

Definition 8 (Class $\mathcal{F}_{\mathcal{G},\gamma}$ of truncated margin functions [37]) *Let $\mathcal{F}_\mathcal{G}$ be a class of functions satisfying Definition 4. Fix $\gamma \in (0,1]$. For any $f_g \in \mathcal{F}_\mathcal{G}$, we define $f_{g,\gamma} : \mathcal{Z} \to [0,\gamma]$ as*

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

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

A margin risk and its empirical counterpart are obtained by replacing ϕ by ϕ_γ in Definitions 5 and 6. In these risks, we will use $f_{g,\gamma}$ instead of f_g. The latter substitution does not affect the margin risk, but it leads to more efficient results due to the restricted codomain.

Definition 9 (Margin risk L_γ, Empirical margin risk $L_{\gamma,m}$ [37]) *For any $g \in \mathcal{G}$, its margin risk is defined as*

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

and its empirical margin risk as

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

Note that the margin risk does not represent the probability of error. Indeed, based on the relationship between the standard loss and the truncated hinge loss functions addressed above

we have that for all $g \in \mathcal{G}$ and for all $\gamma \in (0, 1]$, $L(g) \leqslant L_\gamma(g)$. A similar relationship holds for the empirical versions of these risks.

The question of the convergence of the empirical risk to the true one uniformly over the function class has been studied by Vapnik et Chervonenkis [88]. The conditions under which this happens is based on the notion of capacity measure which we introduce below.

1.4 Capacity Measures

This section gives the definitions of the capacity measures used in the thesis: the covering/packing numbers, the Rademacher complexity and the fat-shattering dimension. We also give several crucial results used in the sequel.

1.4.1 Covering/Packing Numbers

The first capacity measure—and historically the oldest one—we introduce is close to the notion of size of a set. An ϵ-cover of a set in a (pseudo-)metric space is a set of balls of radius ϵ that completely covers this set. The idea to characterize the "largeness" of sets in terms of their minimal covers emerged in the works of Soviet mathematicians in the late 40s. Much work has been dedicated to the study of this notion by Babenko [6], Kolmogorov [44], Vitushkin [90] and Tikhomirov [84] in the 50s. These studies have been set forth in a systematic manner by Kolmogorov and Tikhomirov [45]. The following definitions are due to them.

Definition 10 (Covering numbers, metric entropy, packing numbers [45]) *Let $(\mathcal{T}, \rho)$ be a (pseudo-)metric space. Let $\mathcal{B}_\epsilon(t) = \{t' \in \mathcal{T} : \rho(t', t) < \epsilon\}$ be the open ball of radius $\epsilon > 0$ centered at $t \in \mathcal{T}$. For any $\mathcal{T}' \subset \mathcal{T}$, if there is a set $\bar{\mathcal{T}} \subset \mathcal{T}$ such that*

$$\mathcal{T}' \subset \bigcup_{t \in \bar{\mathcal{T}}} \mathcal{B}_\epsilon(t), \tag{1.1}$$

then $\{\mathcal{B}_\epsilon(t) : t \in \bar{\mathcal{T}}\}$ is an ϵ-cover of $\mathcal{T}'$ and $\bar{\mathcal{T}}$ is an ϵ-net of $\mathcal{T}'$. The ϵ-covering number $\mathcal{N}^{ext}(\epsilon, \mathcal{T}', \rho)$ of $\mathcal{T}'$ is the cardinality of a minimal set $\bar{\mathcal{T}}$ for which (1.1) is true. If an ϵ-net belongs to $\mathcal{T}'$ then the corresponding covering number is a proper one denoted by $\mathcal{N}(\epsilon, \mathcal{T}', \rho)$. The (base 2 or base e) logarithm of a covering number is called metric entropy.

$\mathcal{T}'' \subset \mathcal{T}'$ is ϵ-separated *with respect to the metric ρ if for any two distinct elements $t_1, t_2 \in \mathcal{T}''$, $\rho(t_1, t_2) \geqslant \epsilon$. The ϵ-packing number $\mathcal{M}(\epsilon, \mathcal{T}', \rho)$ of $\mathcal{T}'$ is the maximal cardinality of its ϵ-separated subsets.*

An ϵ-net of a set $\mathcal{T}'$ can also be thought of as its ϵ-approximation with respect to the metric ρ, in the sense that for any element t in $\mathcal{T}'$ there exists t' in its ϵ-net with $\rho(t, t') < \epsilon$.

The following relationship follows from the triangle inequality.

Lemma 1 (After Theorem 4 in [45]) *For any $\epsilon > 0$,*

$$\mathcal{N}^{ext}\left(\epsilon, \mathcal{T}', \rho\right) \leqslant \mathcal{N}\left(\epsilon, \mathcal{T}', \rho\right) \leqslant \mathcal{N}^{ext}\left(\frac{\epsilon}{2}, \mathcal{T}', \rho\right). \tag{1.2}$$

The covering and packing numbers are related to each other as follows.

Lemma 2 (After Theorem 2 in [45]) *For any $\epsilon > 0$,*

$$\mathcal{M}\left(2\epsilon, \mathcal{T}', \rho\right) \leqslant \mathcal{N}^{ext}\left(\epsilon, \mathcal{T}', \rho\right) \leqslant \mathcal{N}\left(\epsilon, \mathcal{T}', \rho\right) \leqslant \mathcal{M}\left(\epsilon, \mathcal{T}', \rho\right). \tag{1.3}$$

From Definition 10, it is clear that the size of an ϵ-net grows (more presicely, does not decrease) as $\epsilon \to 0$. If for any $\epsilon > 0$, there is a finite ϵ-net in $\mathcal{T}$ for its subset $\mathcal{T}'$, then $\mathcal{T}'$ is called a totally bounded set. In an Euclidean space, the definition of total boundedness coincides with that of boundedness: a bounded set is a set that can be completely included in a ball. From Definition 3 it follows that any component class $\mathcal{G}_k$ is totally bounded.

To specify the metrics used for covering/packing numbers, we first introduce the following space of (equivalence classes of) functions.

Definition 11 (Set $L_p(\mathcal{T})$) *Let $\mathcal{F}$ denote the set of all real-valued measurable functions on $(\mathcal{T}, \mathcal{A}_{\mathcal{T}}, P_{\mathcal{T}})$. For all $p \in \mathbb{N}^*$, $L_p(\mathcal{T})$ is the space of p-integrable functions defined as*

$$L_p(\mathcal{T}) = \left\{ f \in \mathcal{F} \,:\, \|f\|_{L_p} = \left(\int_{\mathcal{T}} |f(t)|^p \, dP_{\mathcal{T}}(t) \right)^{\frac{1}{p}} < \infty \right\},$$

and $L_\infty(\mathcal{T})$ is

$$L_\infty(\mathcal{T}) = \left\{ f \in \mathcal{F} \,:\, \|f\|_{L_\infty} = \operatorname*{ess\,sup}_{t \in \mathcal{T}} |f(t)| < \infty \right\}.$$

The following relationship holds between the L_p-norms.

Lemma 3 *Let $p, q \in [1, \infty]$ with $p \leqslant q$. Then, for any $f \in L_q(\mathcal{T})$,*

$$\|f\|_{L_p} \leqslant \|f\|_{L_q}. \tag{1.4}$$

Replacing $P_{\mathcal{T}}$ with an empirical measure P_n supported on a sequence $\mathbf{t}_n = (t_i)_{1 \leqslant i \leqslant n} \in \mathcal{T}^n$ gives rise to an empirical (semi-)norm.

Definition 12 (Empirical (semi-)norm $\|\cdot\|_{L_p(\mathbf{t}_n)}$) *Let $\mathbf{t}_n = (t_i)_{1 \leqslant i \leqslant n} \in \mathcal{T}^n$ and $p \in \mathbb{N}^* \cup \infty$. For any $f \in \mathcal{F}$, we define its empirical (semi-)norm $\|f\|_{L_p(\mathbf{t}_n)}$ supported on $\mathbf{t}_n$ as*

$$\forall p \in \mathbb{N}^*,\ \|f\|_{L_p(\mathbf{t}_n)} = \left(\frac{1}{n} \sum_{i=1}^{n} |f(t_i)|^p \right)^{\frac{1}{p}}$$

and

$$\|f\|_{L_\infty(\mathbf{t}_n)} = \max_{1 \leqslant i \leqslant n} |f(t_i)|.$$

Hereafter, the L_p-norm will stand for the empirical norm defined as above. This norm induces the following empirical (pseudo-)metrics.

Definition 13 (Empirical (pseudo-)metric $d_{p,\mathbf{t}_n}$) *Let $\mathbf{t}_n = (t_i)_{1 \leqslant i \leqslant n} \in \mathcal{T}^n$. For any $f, f' \in \mathcal{F}$, the empirical pseudo-metric $d_{p,\mathbf{t}_n}$ is defined as*

$$\forall p \in \mathbb{N}^*,\ d_{p,\mathbf{t}_n}(f, f') = \left(\frac{1}{n} \sum_{i=1}^{n} \left| f(t_i) - f'(t_i) \right|^p \right)^{\frac{1}{p}}$$

and

$$d_{\infty,\mathbf{t}_n}(f, f') = \max_{1 \leqslant i \leqslant n} \left| f(t_i) - f'(t_i) \right|.$$

Clearly, the relationship (1.4) carries over to the empirical metrics. This implies that for any $\epsilon > 0$ and $p, q \in [1, \infty]$ with $p \leqslant q$,

$$\mathcal{N}\left(\epsilon, \mathcal{F}, d_{p,\mathbf{t}_n}\right) \leqslant \mathcal{N}\left(\epsilon, \mathcal{F}, d_{q,\mathbf{t}_n}\right), \tag{1.5}$$

since an ϵ-net of $\mathcal{F}$ with respect to the $d_{q,\mathbf{t}_n}$ distance is also an ϵ-net with respect to the $d_{p,\mathbf{t}_n}$ distance.

Definition 14 (Restriction $\mathcal{F}|_{\mathcal{T}'}$) *Let $\mathcal{T}' \subseteq \mathcal{T}$. Then $\mathcal{F}|_{\mathcal{T}'}$ is the set of functions in $\mathcal{F}$ restricted to the domain $\mathcal{T}'$.*

For any $\mathbf{t}_n \in \mathcal{T}'^n$, we have that

$$\begin{cases} \mathcal{N}\left(\epsilon, \mathcal{F}, d_{p,\mathbf{t}_n}\right) = \mathcal{N}\left(\epsilon, \mathcal{F}|_{\mathcal{T}'}, d_{p,\mathbf{t}_n}\right) \\ \mathcal{M}\left(\epsilon, \mathcal{F}, d_{p,\mathbf{t}_n}\right) = \mathcal{M}\left(\epsilon, \mathcal{F}|_{\mathcal{T}'}, d_{p,\mathbf{t}_n}\right) \end{cases}. \tag{1.6}$$

Under the empirical metrics, one can define the uniform version of covering numbers as

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

and similarly for packing numbers.

1.4.2 Combinatorial Dimensions

Roughly speaking, a combinatorial dimension of a set functions is the maximum number of points on which the functions output in all possible ways around fixed levels. The first combinatorial dimension, called the *VC-dimension*, originated in the seminal work of Vapnik and Chervonenkis [88, 87]. This dimension concerns a set of $\{0, 1\}$-valued functions. The VC-dimension has been generalized to classes of real-valued functions in several ways. The first such generalization is due to Pollard [69]. By introducing the scale parameter, Kearns and Schapire generalized the dimension of Pollard to a scale-sensitive one called the fat-shattering dimension.

Definition 15 (Fat-shattering dimension [43]) *Let $\gamma \in \mathbb{R}_+$. A non-empty subset $\mathcal{T}'$ of $\mathcal{T}$ is said to be* fat-shattered *or* γ-shattered *by $\mathcal{F}$ if there is a level function $v : \mathcal{T}' \to \mathbb{R}$ such that for any subset $\mathcal{T}'' \subseteq \mathcal{T}'$, there is a function f in $\mathcal{F}$ satisfying*

$$\begin{cases} \forall t \in \mathcal{T}'', \quad f(t) \geqslant v(t) + \gamma, \\ \forall t \in \mathcal{T}' \setminus \mathcal{T}'', \quad f(t) \leqslant v(t) - \gamma. \end{cases}$$

The fat-shattering dimension *of $\mathcal{F}$ at scale γ, γ-dim$(\mathcal{F})$, is the maximal cardinality of a subset of $\mathcal{T}$ γ-shattered by $\mathcal{F}$. If $\mathcal{F}$ γ-shatters sets of unbounded finite sizes, then γ-dim$(\mathcal{F}) = \infty$.*

The strong dimension is used as an auxilary quantity in the proofs of results involving the fat-shattering dimension. This dimension concerns classes of integer-valued functions.

Definition 16 (Strong dimension [1]) *Let $\mathcal{F}'$ be a set of functions from $\mathcal{T}$ to a finite set B of integers. A non-empty subset $\mathcal{T}'$ of $\mathcal{T}$ is said to be* strongly shattered *by $\mathcal{F}'$ if there is a level function $v : \mathcal{T}' \to B$ such that for any $\mathcal{T}'' \subseteq \mathcal{T}'$, there is a function $f' \in \mathcal{F}'$ satisfying*

$$\begin{cases} \forall t \in \mathcal{T}'', \quad f(t) \geqslant v(t) + 1, \\ \forall t \in \mathcal{T}' \setminus \mathcal{T}'', \quad f(t) \leqslant v(t) - 1. \end{cases}$$

The strong dimension *of $\mathcal{F}'$ is the maximal cardinality of a subset of $\mathcal{T}'$ strongly shattered by $\mathcal{F}'$. If $\mathcal{F}$ strongly shatters sets of unbounded finite sizes, then its strong dimension is infinity.*

The generalization of the VC-dimension to classes of $\{0, 1, \ldots, n\}$-valued functions, with finite n, is due to Natarajan [65]. For such classes of functions, Ben-David et al. [15] provide a family of dimensions called Ψ-dimensions unifying the generalizations of the VC-dimension. It includes the Natarajan and the graph dimensions [25, 65] as special cases. Guermeur [34, 37] extended the Ψ-dimensions to the scale-sensitive setting, i.e., to classes of real-valued functions. We give the definitions of the scale sensitive versions of the Natarajan and the graph dimensions below. In these definitions, $\gamma \in \mathbb{R}_+$ and $\mathcal{F}$ stands for a set of real-valued functions on $\mathcal{Z}$.

Definition 17 (Margin Natarajan dimension [37]) *A subset $\mathcal{Z}' = \{(x_i, y_i)\}_{i=1}^n$ of $\mathcal{Z}$ is said to be γ-N shattered by $\mathcal{F}$, if there is a level function $v : \mathcal{Z}' \to \mathbb{R}$ and a vector $(c_i)_{1 \leqslant i \leqslant n} \in \mathcal{Y}^n$ satisfying $c_i \neq y_i$ for all i, such that for any $I' \subseteq I = \{1, \ldots, n\}$, there is a function $f \in \mathcal{F}$ satisfying*

$$\begin{cases} \forall i \in I', \quad f(x_i, y_i) \geqslant \gamma + v(z_i), \\ \forall i \in I \setminus I', \quad f(x_i, c_i) \geqslant \gamma - v(z_i). \end{cases}$$

The margin Natarajan dimension of $\mathcal{F}$, γ-N-dim$(\mathcal{F})$, is the maximal cardinality of a subset of $\mathcal{Z}$ γ-N-shattered by $\mathcal{F}$. If $\mathcal{F}$ γ-N-shatters sets of unbounded finite sizes, then γ-N-dim$(\mathcal{F}) = \infty$.

Definition 18 (Margin graph dimension [37]) *A subset $\mathcal{Z}' = \{(x_i, y_i)\}_{i=1}^n$ of $\mathcal{Z}$ is said to be γ-G shattered by $\mathcal{F}$, if there is a level function $v : \mathcal{Z}' \to \mathbb{R}$ such that for any $I' \subseteq I = \{1, \ldots, n\}$, there is a function $f \in \mathcal{F}$ satisfying*

$$\begin{cases} \forall i \in I', \quad f(x_i, y_i) \geqslant \gamma + v(z_i), \\ \forall i \in I \setminus I', \quad \max_{k \neq y_i} f(x_i, k) \geqslant \gamma - v(z_i). \end{cases}$$

The margin graph dimension of $\mathcal{F}$, γ-G-$dim\,(\mathcal{F})$, is the maximal cardinality of a subset of $\mathcal{Z}$ γ-G-shattered by $\mathcal{F}$. If $\mathcal{F}$ γ-G-shatters sets of unbounded finite sizes, then γ-G-$dim\,(\mathcal{F}) = \infty$.

According to Proposition 1 in [37], the fat-shattering dimension of the margin class $\mathcal{F}_\mathcal{G}$ dominates the margin Natarajan and the margin graph dimensions:

$$\forall \gamma \in (0, M_\mathcal{G}], \quad \gamma\text{-}N\text{-dim}\,(\mathcal{F}_\mathcal{G}) \leqslant \gamma\text{-}G\text{-dim}\,(\mathcal{F}_\mathcal{G}) \leqslant \gamma\text{-dim}\,(\mathcal{F}_\mathcal{G})\,. \tag{1.7}$$

1.4.3 Rademacher/Gaussian Complexity

The last capacity measure considered in this thesis is closely related to the notion of stochastic process. Let $\mathcal{S}$ be some set. A *stochastic process* is an indexed collection $\mathscr{G} = \{\mathscr{G}_s \ : \ s \in \mathcal{S}\}$ of random variables defined on the same probability space $(\mathcal{T}, \mathcal{A}_\mathcal{T}, P_\mathcal{T})$ such that for any $s \in \mathcal{S}$, $\mathscr{G}_s : \mathcal{T} \to \mathbb{R}$ is $(\mathcal{A}_\mathcal{T}, \mathbb{B}_\mathbb{R})$-measurable. In the learning theory, $\mathcal{S}$ is usually a set of functions, and one deals with a particular type of stochastic process: an empirical process. In the following definitions, T is a random variable taking values in $(\mathcal{T}, \mathcal{A}_\mathcal{T}, P_\mathcal{T})$ and $\mathbf{T}_n = (T_i)_{1 \leqslant i \leqslant n}$ is a sequence of n independent copies of T.

Definition 19 (Empirical process [54, 86, 68]) *Let $L_1(\mathcal{T})$ be as in Definition 11. The (centered) empirical process $\mathscr{G}_{n,\mathcal{F}}$ indexed by a set $\mathcal{F} \subset L_1(\mathcal{T})$ is as follows:*

$$\mathscr{G}_{n,\mathcal{F}} = \left\{ \frac{1}{n} \sum_{i=1}^n f(T_i) - \mathbb{E}\left[f(T)\right] : f \in \mathcal{F} \right\}.$$

For a fixed $\gamma \in (0, 1]$, the set $\{L_\gamma(g) - L_{\gamma,m}(g) : g \in \mathcal{G}\}$ of deviations central to this thesis is an empirical process. Whether the classical limit theorems hold for empirical processes is the major question in the theory of empirical processes. It has been shown by Vapnik and Chervonenkis [88], Giné and Zinn [30], that these theorems can be proven for the "symmetrized" versions of these processes. This is based on the use of a Rademacher variable σ, symmetric Bernoulli random variable taking values 1 and -1 with equal probability, and a standard Gaussian random variable. This calls for the definition of the Rademacher and Gaussian processes. Let $\boldsymbol{\sigma}_n = (\sigma_i)_{1 \leqslant i \leqslant n}$ be a sequence of n independent copies of a Rademacher variable σ, and let $\boldsymbol{o}_n = (o_i)_{1 \leqslant i \leqslant n}$ be an

orthogaussian sequence, i.e., a sequence of n independent copies of a standard Gaussian random variable.

Definition 20 (Rademacher process) *The Rademacher process $\mathscr{R}_n$ indexed by $\mathcal{F}$ is an empirical process conditioned on $\mathbf{T}_n$:*

$$\mathscr{R}_{n,\mathcal{F}} = \left\{ \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(T_i) : f \in \mathcal{F} \right\}.$$

In this definition, replacing the Rademacher sequence with an orthogaussian one yields a Gaussian process. The use of suprema of these processes as capacity measures of the class by which they are indexed is relatively recent. This work was started by Koltchinskii [46], Bartlett and co-authors [8] and Bartlett and Mendelson [13].

Definition 21 (Rademacher complexity, Gaussian complexity) *The empirical Rademacher complexity of $\mathcal{F}$ given $\mathbf{T}_n$ is defined as*

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

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

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

The empirical Gaussian complexity $\hat{G}_n(\mathcal{F})$, and the Gaussian complexity $G_n(\mathcal{F})$ are defined in the same way by substituting $\boldsymbol{o}_n$ for $\boldsymbol{\sigma}_n$.

One way to interpret the Rademacher complexity of a class $\mathcal{F}$ is to think of it as the degree of imitation of $\mathcal{F}$ of the Rademacher noise $\boldsymbol{\sigma}_n$ (likewise for the Gaussian complexity). For some $M_{\mathcal{F}} > 0$, let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ to $[-M_{\mathcal{F}}, M_{\mathcal{F}}]$. Let $\mathbf{s}_n = (s_i)_{1 \leqslant i \leqslant n} \in \{-1, 1\}^n$. If for any $\mathbf{s}_n$, there exists a function $f \in \mathcal{F}$ such that $s_i f(t_i) = M_{\mathcal{F}}$, then the class $\mathcal{F}$ agrees well with the Rademacher noise $\boldsymbol{\sigma}_n$. Note also that according to Jensen's inequality and by definition of a Rademacher variable,

$$\hat{R}_n(\mathcal{F}) \geqslant \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\sigma_i f(t_i)\right] = 0.$$

Several properties of the Rademacher complexity (the proofs of which are gathered in Appendix A) will be useful in the sequel. The first important result is the *contraction* principle due to Talagrand: it allows one to switch from the Rademacher complexity of the composition of a Lipschitz function with a class to the Rademacher complexity of this class.

Lemma 4 (After Theorem 4.12 in [54]) *Let $\psi : \mathbb{R} \to \mathbb{R}$ be an L-Lipschitz function and let $\psi \circ \mathcal{F} = \{\psi \circ f : f \in \mathcal{F}\}$. Then,*

$$\hat{R}_n(\psi \circ \mathcal{F}) \leqslant L\hat{R}_n(\mathcal{F}). \tag{1.8}$$

The Rademacher complexity does not change when one takes the convex hull of a class.

Lemma 5 *Let $conv(\mathcal{F}) = \left\{\sum_{j=1}^{N} \alpha_j f_j : N \in \mathbb{N}^*, \alpha_j \geqslant 0, \sum_{j=1}^{N} \alpha_j = 1, f_j \in \mathcal{F}\right\}$. Then,*

$$\hat{R}_n(\mathcal{F}) = \hat{R}_n(conv(\mathcal{F})).$$

Remark 1 *The classical definition of the Rademacher complexity involves the supremum of the absolute value of the sum $\sum_{i=1}^{n} \sigma_i f(T_i)$ (see, for instance, [13]). In this case, thanks to the presence of the absolute value, one has $\hat{R}_n(\mathcal{F}) = \hat{R}_n(conv(\mathcal{F})) = \hat{R}_n(absconv(\mathcal{F}))$ where*

$$absconv(\mathcal{F}) = \left\{\sum_{j=1}^{N} \alpha_j f_j : N \in \mathbb{N}^*, \alpha_j \in \mathbb{R}, \sum_{j=1}^{N} |\alpha_j| \leqslant 1, f_j \in \mathcal{F}\right\}.$$

These two definitions agree if for any f in $\mathcal{F}$, $-f$ is also in $\mathcal{F}$.

The Gaussian complexity dominates the Rademacher complexity as follows.

Lemma 6 (After Lemma 3.2.10 in [83]) *For all $n \in \mathbb{N}$,*

$$\hat{R}_n(\mathcal{F}) \leqslant \sqrt{\frac{\pi}{2}}\hat{G}_n(\mathcal{F})$$

1.5 Limit Theorems and Capacity Measures

The seminal work of Vapnik and Chervonenkis [88] demonstrated that there is a connection between the law of large numbers and a capacity measure of the class. More precisely, in Theorem 4 in [88], the authors give the necessary and sufficient conditions for the uniform convergence of the empirical risk to the true one in terms of the growth rate of the "metric entropy". The connections between the classical limit theorems and the capacity measures have been later developed by Koltchinskii [48], Gine and Zinn [30], Talagrand [80], Dudley [26] and Alon et al. [1]. In this section, we give the formal definitions of classes that satisfy the strong law of large numbers and the central limit theorem, i.e., Glivenko-Cantelli and Donsker classes, respectively, and the necessary and sufficient conditions for these classes.

Let T be a real-valued random variable with a law $P_{\mathcal{T}}$, and $(T_i)_{1 \leqslant i \leqslant n}$ a sequence of n independent copies of T. The Glivenko-Cantelli theorem states that the empirical distribution

function $F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \delta_{T_i}((-\infty, t])$ converges almost surely to the true distribution function $F(t) = P_{\mathcal{T}}(T \leqslant t)$ uniformly over $\mathbb{R}$. The collection $\{(-\infty, t] \ : \ t \in \mathbb{R}\}$ of all half-lines is said to be a Glivenko-Cantelli class. This concept was generalized to function classes in the following manner.

Definition 22 (Glivenko-Cantelli class [27]) *Let $L_1(\mathcal{T})$ be as in Definition 11. Then, $\mathcal{F} \subset L_1(\mathcal{T})$ is called a Glivenko-Cantelli class if and only if for all $\epsilon > 0$,*

$$\lim_{n\to\infty} P_{\mathcal{T}}^{\infty}\left(\sup_{m\geqslant n}\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^{m} f(t_i) - \mathbb{E}\left[f(T)\right]\right| > \epsilon\right) = 0.$$

If $\mathcal{F}$ is a Glivenko-Cantelli class for any probability measure $P_{\mathcal{T}}$ on $\mathcal{A}_{\mathcal{T}}$, then it is called a universal *Glivenko-Cantelli class. $\mathcal{F}$ is called a* uniform *Glivenko-Cantelli class if and only if for all $\epsilon > 0$,*

$$\lim_{n\to\infty}\sup_{P_{\mathcal{T}}} P_{\mathcal{T}}^{\infty}\left\{\sup_{m>n}\sup_{f\in\mathcal{F}}\left|\frac{1}{m}\sum_{i=1}^{m} f(t_i) - \mathbb{E}[f(T)]\right| > \epsilon\right\} = 0.$$

For the empirical process $\{\sqrt{n}\,(F_n(t) - F(t)) \ : \ t \in \mathbb{R}\}$, the multivariate central limit theorem asserts that the law of any finite subset $\{\sqrt{n}\,(F_n(t_i) - F(t_i)) : 1 \leqslant i \leqslant l\}$ with $l \in \mathbb{N}^*$ converges to the zero-mean Gaussian distribution with a covariance function $\min(F(t_i), F(t_j)) - F(t_i)F(t_j)$. The convergence in distribution of an empirical process indexed by a function class has been addressed by Donsker [26]. The following definition of a Donsker class is from van der Vaart and Wellner [86].

Definition 23 (Donsker class [86]) *Let $L_2(\mathcal{T})$ be as in Definition 11. Let $\mathscr{G}_{\mathcal{F}}$ be a zero-mean Gaussian process indexed by $\mathcal{F} \subset L_2(\mathcal{T})$ with a covariance function*

$$\mathbb{E}[f_1(T)f_2(T)] - \mathbb{E}[f_1(T)]\mathbb{E}[f_2(T)],$$

for any $f_1, f_2 \in \mathcal{F}$. Let $\mathscr{G}_{n,\mathcal{F}}$ be an empirical process indexed by $\mathcal{F}$ as in Definition 19. $\mathcal{F}$ is called a $P_{\mathcal{T}}$-Donsker class if and only if $\mathscr{G}_{\mathcal{F}}$ is a tight Borel-measurable element of the space of all uniformly bounded functions from $\mathcal{F}$ to $\mathbb{R}$ and $\mathscr{G}_{n,\mathcal{F}}$ converges in distribution to $\mathscr{G}_{\mathcal{F}}$. $\mathcal{F}$ is called a universal Donsker class if $\mathcal{F}$ is $P_{\mathcal{T}}$-Donsker for any any probability measure $P_{\mathcal{T}}$ on $\mathcal{A}_{\mathcal{T}}$.

For the uniform convergence of the empirical risk to the true one, it is assumed that a class is a Glivenko-Cantelli class. According to Slutsky's lemma (Example 1.4.7 in [86]), on the other hand, all Donsker classes are Glivenko-Cantelli, and thus much more happens for such classes (i.e., faster rate of convergence). Whether a (uniformly bounded) class $\mathcal{F}$ is a Glivenko-Cantelli or Donsker class depends on the growth-rate of the metric entropy of $\mathcal{F}$.

Theorem 1 (After Theorem 6 [27]) *Let $L_\infty(\mathcal{T})$ be as in Definition 11. A class $\mathcal{F} \subset L_\infty(\mathcal{T})$ is a uniform Glivenko-Cantelli class if and only if for any $\epsilon > 0$ and for any $p \in \mathbb{N}^* \cup \{\infty\}$,*

$$\lim_{n\to\infty} \frac{\ln \mathcal{N}_p(\epsilon, \mathcal{F}, n)}{n} = 0. \tag{1.9}$$

Similarly, the following theorem gives the necessary and sufficient conditions for Donsker class.

Theorem 2 (Dudley [26]) *Let $L_\infty(\mathcal{T})$ be as in Definition 11. A class $\mathcal{F} \subset L_\infty(\mathcal{T})$ is a universal Donsker class if*

$$\int_0^\infty \sup_{n\in\mathbb{N}^*} \sqrt{\ln \mathcal{N}_2(\epsilon, \mathcal{F}, n)} d\epsilon < \infty. \tag{1.10}$$

If $\mathcal{F}$ is a universal Donsker class, then there exists a constant $K > 0$, such that for any $\epsilon > 0$,

$$\sup_{n\in\mathbb{N}^*} \ln \mathcal{N}_2(\epsilon, \mathcal{F}, n) \leqslant \frac{K}{\epsilon^2}.$$

Based on the bounds which relate the covering number (or metric entropy) of a class to one of its combinatorial dimension, which we call *combinatorial* bounds, the Glivenko-Cantelli and Donsker classes can be characterized in terms of their combinatorial dimensions. The first such bound related the growth function of $\{0,1\}$-valued function classes to their VC-dimensions [88, 74, 87]. The result relating the L_1-norm metric entropy of such classes to their VC-dimensions is due to Haussler [40]; this was generalized to L_p-norms with $p \in \mathbb{N}^*$ by Van der Vaart and Wellner [86]. These results give sufficient conditions both for Glivenko-Cantelli and Donsker theorems: the classes with finite VC-dimension, the *VC-classes*, are both Glivenko-Cantelli and Donsker. For classes of real-valued functions, on the other hand, the combinatorial bound of Alon et al. [1] involves a scale-sensitive generalization of the VC-dimension: the fat-shattering dimension.

Theorem 3 (After Theorem 2.5 in [1]) *Let $L_\infty(\mathcal{T})$ be as in Definition 11. A class $\mathcal{F} \subset L_\infty(\mathcal{T})$ is a uniform Glivenko-Cantelli class if and only if the fat-shattering dimension of $\mathcal{F}$ is finite at every scale ϵ.*

Due to the fact that the fat-shattering dimension is scale-sensitive, the finiteness of this dimension is not sufficient for Donsker theorem as it is the case for the VC-dimension. In fact, it is the growth rate of the fat-shattering dimension that determines whether a class obeys Donsker theorem. As it was noted by Mendelson [59], if the fat-shattering dimension of a class is $O(\epsilon^{-\alpha})$ for $0 < \alpha < 2$, then it is a Donsker class.

In this thesis, we make the assumption that the component classes $\mathcal{G}_k$ are uniform Glivenko-Cantelli, i.e., their fat-shattering dimensions are finite. For the results of Chapter 4 and Chapter 5, a stronger assumption is made: the fat-shattering dimensions of the classes $\mathcal{G}_k$ grow no faster than polynomially with the inverse of their scales.

Chapter 2

Controlling Uniform Convergence by a Covering Number

In this chapter, our goal is to control the uniform deviation of the risk and the empirical margin risk in terms of an L_1-norm covering number.

Inspired by the seminal work of Vapnik and Chervonenkis [88], Bartlett [7] established an error bound for margin bi-category classifiers on the basis of the margin indicator loss function. The capacity measure appearing in his error bound is the L_∞-norm covering number. This work has been extended to the multi-category setting by Guermeur [34, 36]. Our work is based on the use of the truncated hinge loss function. For this loss function, the uniform deviation of interest can be handled based on Pollard's method [68], as well as the result of Bartlett and Long [12]. Pollard derived the rate of convergence for the classical Glivenko-Cantelli problem based on the L_1-norm approximation of the set. He also extended it to classes of real-valued functions. For these classes, based on Pollard's method, Bartlett and Long [12] derived a faster rate of convergence. They translated this result into a sample complexity estimate, i.e., given $\epsilon, \delta \in (0, 1)$, a minimum sample size needed for the uniform deviation to be at most ϵ with probability at least $1 - \delta$. This result was based on the L_1-norm generalization of the bound of Alon et al. [1], Lemma 8 in [12]. Apart from ϵ and δ, this sample complexity estimate also depends on the fat-shattering dimension of the function class.

In Section 2.1, we generalize their uniform convergence result, Lemma 10 combined with Lemma 11 in [12], to the multi-category setting. In Section 2.2, we derive several sample complexity estimates using the L_1 and L_2-norm combinatorial bounds [12, 36, 61]. These bounds include the one that depend on the sample size and the ones that do not, i.e., *dimension-free* bounds. Our estimates obtained based on the L_1-norm combinatorial bounds match with that of

Bartlett and Long. Particularly, we demonstrate that using the dimension-free L_1-norm bound does not improve the sample size estimate. The dimension-free L_2-norm bound of Mendelson and Vershynin [61], on the other hand, improves the dependency on ϵ at the cost of deteriorating the scale parameter of the fat-shattering dimension.

2.1 L_1-norm Covering Number Bound

Given a fixed $\gamma \in (0, 1]$ and a fixed $\epsilon \in (0, 1)$, our goal is to upper bound the probability

$$P^m \left(\sup_{g \in \mathcal{G}} \left(L(g) - L_{\gamma,m}(g) \right) > \epsilon \right) \tag{2.1}$$

in terms of the L_1-norm covering number of $\mathcal{F}_{\mathcal{G},\gamma}$. This corresponds to a multi-category extension of Lemma 10 combined with Lemma 11 of Bartlett and Long [12]. We derive the following bound where the scale of the covering number involves the margin parameter γ due to the use of a margin loss function.

Theorem 4 *Fix $\epsilon \in (0, 1)$ and $\gamma \in (0, 1]$. Then for $m > \frac{2}{\epsilon^2}$,*

$$P^m \left(\sup_{g \in \mathcal{G}} \left(L\left(g\right) - L_{\gamma,m}\left(g\right) \right) > \epsilon \right) \leqslant 2\mathcal{N}_1 \left(\frac{\epsilon\gamma}{8}, \mathcal{F}_{\mathcal{G},\gamma}, 2m \right) \exp\left(-\frac{m\epsilon^2}{32} \right). \tag{2.2}$$

Proof The proof consists of the following steps: i) apply the symmetrization technique of Vapnik and Chervonenkis [88] which allows one to switch from the original problem (deviation between the margin risk and the empirical one) to the symmetrized one (deviation between two empirical margin risks); ii) thanks to step (i), switch to a finite class, i.e., to an ϵ-approximation of the class $\mathcal{F}_{\mathcal{G},\gamma}$; iii) perform the second symmetrization to introduce a Rademacher sequence which allows one to condition on the sample (the distribution of which we do not know) and work with a Rademacher sequence instead; and iv) thanks to steps (ii)-(iii), apply the union bound and a concentration inequality which gives Inequality (2.2).

First Symmetrization The idea of the first symmetrization is to use an independent copy $\mathbf{Z}'_m$ of $\mathbf{Z}_m$ which is usually referred to as a "ghost sample" and bound the tail probability in terms of the probability of events involving the empirical risks only. As it has been pointed out in Chapter 1, the truncated hinge loss function dominates the standard indicator loss function and thus, for any $\gamma \in (0, 1]$ and for any $g \in \mathcal{G}$,

$$L(g) \leqslant L_\gamma(g).$$

Then, it follows that

$$P^m\left(\sup_{g\in\mathcal{G}}\left(L(g)-L_{\gamma,m}(g)\right)>\epsilon\right)\leqslant P^m\left(\sup_{g\in\mathcal{G}}\left(L_\gamma(g)-L_{\gamma,m}(g)\right)>\epsilon\right).$$

This allows us to make use of Lemma 35 in Appendix C. Applying it to the right-hand side gives:

$$\begin{aligned}P^m\left(\sup_{g\in\mathcal{G}}\left(L(g)-L_{\gamma,m}(g)\right)>\epsilon\right)&\leqslant 2P^{2m}\left\{\sup_{g\in\mathcal{G}}\left(\frac{1}{m}\sum_{i=1}^m\phi_\gamma\left(f_{g,\gamma}\left(Z_i'\right)\right)-\frac{1}{m}\sum_{i=1}^m\phi_\gamma\left(f_{g,\gamma}\left(Z_i\right)\right)\right)\geqslant\frac{\epsilon}{2}\right\}\\&=2\int_{\mathcal{Z}^{2m}}\mathbb{1}\left\{\sup_{g\in\mathcal{G}}\left(\frac{1}{m}\sum_{i=1}^m\phi_\gamma\left(f_{g,\gamma}\left(z_i'\right)\right)-\frac{1}{m}\sum_{i=1}^m\phi_\gamma\left(f_{g,\gamma}\left(z_i\right)\right)\right)\geqslant\frac{\epsilon}{2}\right\}dP^{2m}(\mathbf{z}_{2m}),\end{aligned}\tag{2.3}$$

where $\mathbf{z}_{2m}$ is the concatenation of $\mathbf{z}_m=(z_i)_{1\leqslant i\leqslant m}\in\mathcal{Z}^m$ and $\mathbf{z}'_m=(z'_i)_{1\leqslant i\leqslant m}\in\mathcal{Z}^m$.

Denote the integral by I. To keep the notation simple, for the rest of the proof, let $\phi'_g=\phi_\gamma(f_{g,\gamma})$.

Maximal Inequality At this point, we approximate $\mathcal{F}_{\mathcal{G},\gamma}$ by its finite cover with respect to $d_{1,\mathbf{z}_{2m}}$. Let $\bar{\mathcal{G}}$ be a subset of $\mathcal{G}$ so that $\mathcal{F}_{\bar{\mathcal{G}},\gamma}$ is an $\frac{\epsilon\gamma}{8}$-net of $\mathcal{F}_{\mathcal{G},\gamma}$ of minimal cardinality $\mathcal{N}\left(\frac{\epsilon\gamma}{8},\mathcal{F}_{\mathcal{G},\gamma},d_{1,\mathbf{z}_{2m}}\right)$:

$$\forall g\in\mathcal{G},\ \exists\bar{g}\in\bar{\mathcal{G}},\quad\frac{1}{2m}\sum_{i=1}^m\left(\left|f_{\bar{g},\gamma}\left(z_i\right)-f_{g,\gamma}\left(z_i\right)\right|+\left|f_{\bar{g},\gamma}\left(z_i'\right)-f_{g,\gamma}\left(z_i'\right)\right|\right)<\frac{\epsilon\gamma}{8}.$$

Using the $\frac{1}{\gamma}$-Lipschitz property of ϕ_γ, we get:

$$\begin{aligned}\frac{1}{2m}\sum_{i=1}^m\left(\left|\phi'_{\bar{g}}\left(z_i\right)-\phi'_g\left(z_i\right)\right|+\left|\phi'_{\bar{g}}\left(z_i'\right)-\phi'_g\left(z_i'\right)\right|\right)&\leqslant\frac{1}{2m\gamma}\sum_{i=1}^m\left(\left|f_{\bar{g}}\left(z_i\right)-f_g\left(z_i\right)\right|+\left|f_{\bar{g}}\left(z_i'\right)-f_g\left(z_i'\right)\right|\right)\\&<\frac{\epsilon}{8}.\end{aligned}$$

On the other hand,

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^m\left(\phi'_{\bar{g}}(z_i)-\phi'_{\bar{g}}\left(z_i'\right)+\phi'_g\left(z_i'\right)-\phi'_g(z_i)\right)&\leqslant\frac{1}{m}\sum_{i=1}^m\left(\left|\phi'_{\bar{g}}\left(z_i\right)-\phi'_g\left(z_i\right)\right|+\left|\phi'_{\bar{g}}\left(z_i'\right)-\phi'_g\left(z_i'\right)\right|\right)\\&<\frac{\epsilon}{4}.\end{aligned}$$

It follows that

$$\frac{1}{m}\sum_{i=1}^m\left(\phi'_g\left(z_i'\right)-\phi'_g(z_i)\right)\geqslant\frac{\epsilon}{2}\implies\frac{1}{m}\sum_{i=1}^m\left(\phi'_{\bar{g}}\left(z_i'\right)-\phi'_{\bar{g}}\left(z_i\right)\right)>\frac{\epsilon}{4}.\tag{2.4}$$

Thus,

$$I\leqslant\int_{\mathcal{Z}^{2m}}\mathbb{1}\left\{\max_{\bar{g}\in\bar{\mathcal{G}}}\frac{1}{m}\sum_{i=1}^m\left(\phi'_{\bar{g}}\left(z_i'\right)-\phi'_{\bar{g}}(z_i)\right)>\frac{\epsilon}{4}\right\}dP^{2m}(\mathbf{z}_{2m}).$$

In the subsequent steps, we bound the right-hand side thanks to the introduction of a Rademacher sequence.

Second Symmetrization Since Z_i' and Z_i are independent copies of Z, for any measurable function s on $\mathcal{Z}$, $s(Z_i') - s(Z_i)$ is distributed in the same way as $s(Z_i) - s(Z_i')$ for any $i \in [\![1, m]\!]$. Thus, $s(Z_i') - s(Z_i)$ is a symmetric random variable. It implies that for any $\mathbf{s}_m = (s_i)_{1 \leqslant i \leqslant m} \in \{-1, 1\}^m$,

$$I \leqslant \int_{\mathcal{Z}^{2m}} \mathbb{1}\left\{ \max_{\bar{g} \in \bar{\mathcal{G}}} \frac{1}{m} \sum_{i=1}^{m} s_i \left(\phi'_{\bar{g}}(z_i') - \phi'_{\bar{g}}(z_i) \right) > \frac{\epsilon}{4} \right\} dP^{2m}(\mathbf{z}_{2m}).$$

It follows that

$$\begin{aligned} I &\leqslant \frac{1}{2^m} \sum_{\mathbf{s}_m \in \{-1,1\}^m} \int_{\mathcal{Z}^{2m}} \mathbb{1}\left\{ \max_{\bar{g} \in \bar{\mathcal{G}}} \frac{1}{m} \sum_{i=1}^{m} s_i \left(\phi'_{\bar{g}}(z_i') - \phi'_{\bar{g}}(z_i) \right) > \frac{\epsilon}{4} \right\} dP^{2m}(\mathbf{z}_{2m}) \\ &= \mathbb{E}_{\boldsymbol{\sigma}_m} \left[\int_{\mathcal{Z}^{2m}} \mathbb{1}\left\{ \max_{\bar{g} \in \bar{\mathcal{G}}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \left(\phi'_{\bar{g}}(z_i') - \phi'_{\bar{g}}(z_i) \right) > \frac{\epsilon}{4} \right\} dP^{2m}(\mathbf{z}_{2m}) \right], \end{aligned}$$

where $\boldsymbol{\sigma}_m$ is a Rademacher sequence. Here, Tonelli's theorem applies and we can change the order of integration:

$$\begin{aligned} I &\leqslant \int_{\mathcal{Z}^{2m}} \mathbb{E}_{\boldsymbol{\sigma}_m} \left[\mathbb{1}\left\{ \max_{\bar{g} \in \bar{\mathcal{G}}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \left(\phi'_{\bar{g}}(z_i') - \phi'_{\bar{g}}(z_i) \right) > \frac{\epsilon}{4} \right\} \right] dP^{2m}(\mathbf{z}_{2m}) \\ &= \int_{\mathcal{Z}^{2m}} P_{\boldsymbol{\sigma}_m} \left(\max_{\bar{g} \in \bar{\mathcal{G}}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \left(\phi'_{\bar{g}}(z_i') - \phi'_{\bar{g}}(z_i) \right) > \frac{\epsilon}{4} \right) dP^{2m}(\mathbf{z}_{2m}). \end{aligned} \tag{2.5}$$

Concentration inequality Now, the integrand on the right-hand side of (2.5) calls for the application of the union-bound:

$$I \leqslant \int_{\mathcal{Z}^{2m}} \sum_{\bar{g} \in \bar{\mathcal{G}}} P_{\boldsymbol{\sigma}_m} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i \left(\phi'_{\bar{g}}(z_i') - \phi'_{\bar{g}}(z_i) \right) > \frac{\epsilon}{4} \right) dP^{2m}(\mathbf{z}_{2m}). \tag{2.6}$$

For a fixed $\mathbf{z}_{2m} \in \mathcal{Z}^{2m}$ and for a fixed $\bar{g} \in \bar{\mathcal{G}}$, we can bound the tail probability

$$P_{\boldsymbol{\sigma}_m} \left\{ \frac{1}{m} \sum_{i=1}^{m} \sigma_i \left(\phi'_{\bar{g}}(z_i') - \phi'_{\bar{g}}(z_i) \right) > \frac{\epsilon}{4} \right\}$$

using Hoeffding's inequality (B.6) in Appendix B. Note that

$$\forall i \in [\![1, m]\!], \quad \sigma_i \left(\phi'_{\bar{g}}(z_i') - \phi'_{\bar{g}}(z_i) \right) \in [-1, 1],$$

almost surely. Thus,

$$P_{\boldsymbol{\sigma}_m} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i \left(\phi'_{\bar{g}}(z_i') - \phi'_{\bar{g}}(z_i) \right) > \frac{\epsilon}{4} \right) \leqslant \exp\left(-\frac{m\epsilon^2}{32} \right).$$

Substituting the last bound into (2.6) and taking into account that the cardinality of $\bar{\mathcal{G}}$ is $\mathcal{N}\left(\frac{\epsilon\gamma}{8}, \mathcal{F}_{\mathcal{G},\gamma}, d_{1,\mathbf{z}_{2m}}\right)$, we obtain:

$$\begin{aligned} I &\leqslant \exp\left(-\frac{m\epsilon^2}{32}\right) \int_{\mathcal{Z}^{2m}} \mathcal{N}\left(\frac{\epsilon\gamma}{8}, \mathcal{F}_{\mathcal{G},\gamma}, d_{1,\mathbf{z}_{2m}}\right) dP^{2m}\left(\mathbf{z}_{2m}\right) \\ &\leqslant \exp\left(-\frac{m\epsilon^2}{32}\right) \mathcal{N}_1\left(\frac{\epsilon\gamma}{8}, \mathcal{F}_{\mathcal{G},\gamma}, 2m\right). \end{aligned} \tag{2.7}$$

Taking this bound into account in (2.3) concludes the proof. ■

In the following, we translate our result, Theorem 4, into a sample complexity estimate, i.e., the minimum sample size required for the right-hand side of (2.1) to be at most $\delta \in (0,1)$.

2.2 Sample Complexity

We derive several sample complexity estimates based on the L_1 and L_2-norm combinatorial bounds. Our starting point is the following L_1-norm combinatorial bound obtained using Lemmas 9-11 in [36]. It depends on the sample size.

Lemma 7 *Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ to $[-M_{\mathcal{F}}, M_{\mathcal{F}}]$ with $M_{\mathcal{F}} \in \mathbb{R}_+$. Let $d(\epsilon) = \epsilon\text{-dim}(\mathcal{F})$ be finite for all $\epsilon \in (0, M_{\mathcal{F}}]$. Fix $\epsilon \in (0, 2M_{\mathcal{F}}]$. Then, provided that $n \geqslant d\left(\frac{\epsilon}{8}\right)$,*

$$\mathcal{N}_1(\epsilon, \mathcal{F}, n) \leqslant 2^{1+3\log_2\left\lceil\frac{36M_{\mathcal{F}}}{\epsilon}\right\rceil} \left(\frac{36M_{\mathcal{F}}en}{\epsilon d\left(\frac{\epsilon}{8}\right)}\right)^{3d\left(\frac{\epsilon}{8}\right)\log_2\left\lceil\frac{36M_{\mathcal{F}}}{\epsilon}\right\rceil}.$$

Proof Let $\mathcal{T}_n = \{t_i : 1 \leqslant i \leqslant n\}$ be a finite subset of $\mathcal{T}$ and $\mathbf{t}_n = (t_i)_{1\leqslant i\leqslant n}$. Let $\mathcal{F}_\epsilon$ be an ϵ-separated with respect to the pseudo-metric $d_{p,\mathbf{t}_n}$ subset of $\mathcal{F}$ of maximal cardinality. For any $\eta \in \mathbb{R}_+$, let $\tilde{\mathcal{F}}^\eta$ be a discretized set:

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

By Lemma 9 in [36],

$$\forall \eta \in (0,\epsilon),\ \mathcal{M}(\epsilon, \mathcal{F}|_{\mathcal{T}_n}, d_{1,\mathbf{t}_n}) \leqslant \mathcal{M}\left(\frac{\epsilon-\eta}{2}, \tilde{\mathcal{F}}^\eta, d_{1,\mathbf{t}_n}\right). \tag{2.8}$$

Next, we apply Lemma 11 in [36] to the right-hand side of Inequality (2.8). To this end, we set $\eta = \frac{2M_{\mathcal{F}}}{N}$ for $N \in \mathbb{N}$ satisfying $N > \frac{14M_{\mathcal{F}}}{\epsilon}$. Then, by this lemma,

$$\mathcal{M}(\epsilon, \mathcal{F}|_{\mathcal{T}_n}, d_{1,\mathbf{t}_n}) \leqslant 2^{1+3\log_2 N}\left(\frac{en(N-1)}{d_1}\right)^{3d_1\log_2 N}, \tag{2.9}$$

where $d_1 = \frac{1}{4}\left(\epsilon\text{-}\frac{14M_{\mathcal{F}}}{N}\right)\text{-dim}\left(\tilde{\mathcal{F}}^{\frac{2M_{\mathcal{F}}}{N}}\right)$. Next, we switch from the fat-shattering dimension of $\tilde{\mathcal{F}}^{\frac{2M_{\mathcal{F}}}{N}}$ to that of $\mathcal{F}$. We set $N = \left\lceil \frac{36M_{\mathcal{F}}}{\epsilon} \right\rceil$. Since the fat-shattering dimension is a non-increasing function, we have

$$d_1 \leqslant \frac{11\epsilon}{72}\text{-dim}\left(\tilde{\mathcal{F}}^{\left(\frac{2M_{\mathcal{F}}}{\left\lceil \frac{36M_{\mathcal{F}}}{\epsilon} \right\rceil}\right)}\right).$$

Applying Lemma 10 in [36], i.e.,

$$\epsilon\text{-dim}\left(\tilde{\mathcal{F}}^{\eta}\right) \leqslant \left(\epsilon - \frac{\eta}{2}\right)\text{-dim}\left(\mathcal{F}\right),$$

to the right-hand side, we get

$$d_1 \leqslant \frac{\epsilon}{8}\text{-dim}\left(\mathcal{F}|_{\mathcal{T}_n}\right).$$

On the other hand,

$$\frac{\epsilon}{8}\text{-dim}\left(\mathcal{F}|_{\mathcal{T}_n}\right) \leqslant \frac{\epsilon}{8}\text{-dim}\left(\mathcal{F}\right) = d\left(\frac{\epsilon}{8}\right),$$

and thus

$$d_1 \leqslant d\left(\frac{\epsilon}{8}\right). \tag{2.10}$$

Now, for any strictly positive u, a and b, with $u \geqslant 2$ and $b \geqslant a$,

$$u \cdot \frac{b}{a} \leqslant u^{\frac{b}{a}}.$$

Then, for any $v \geqslant b$,

$$\frac{uv}{a} = \frac{uv}{b} \cdot \frac{b}{a} \leqslant \left(\frac{uv}{b}\right)^{\frac{b}{a}} \implies \left(\frac{uv}{a}\right)^{a} \leqslant \left(\frac{uv}{b}\right)^{b}.$$

Based on this relation and assuming that $n \geqslant d\left(\frac{\epsilon}{8}\right)$, we can use Inequality (2.10) in (2.9). Combining this result with Equality (1.6) produces

$$\mathcal{M}\left(\epsilon, \mathcal{F}, d_{1,\mathbf{t}_n}\right) \leqslant 2^{3\log_2\left\lceil \frac{36M_{\mathcal{F}}}{\epsilon} \right\rceil + 1}\left(\frac{36M_{\mathcal{F}}en}{\epsilon d\left(\frac{\epsilon}{8}\right)}\right)^{3d\left(\frac{\epsilon}{8}\right)\log_2\left\lceil \frac{36M_{\mathcal{F}}}{\epsilon} \right\rceil}.$$

The proof follows by taking supremum over $\mathbf{t}_n \in \mathcal{T}^n$ from both sides and using Inequality (1.3).

■

This bound gives us the following sample complexity estimate.

Theorem 5 *Fix $\epsilon, \delta \in (0,1)$ and fix $\gamma \in (0,1]$. For $\epsilon \in (0,\gamma]$, let $d(\epsilon) = \epsilon\text{-dim}(\mathcal{F}_{\mathcal{G},\gamma})$. Then, for a sample size*

$$m \geqslant \frac{64}{\epsilon^2}\left(\frac{3d\left(\frac{\epsilon\gamma}{64}\right)}{\ln 2}\ln\left(\frac{289}{\epsilon}\right)\ln\left(\frac{565^2}{\epsilon^3}\ln\left(\frac{289}{\epsilon}\right)\right)+\ln\left(\frac{2}{\delta}\right)\right),$$

the probability (2.1) *is at most δ.*

Proof Applying Lemma 7 with some simplifications to $\mathcal{N}_1\left(\frac{\epsilon\gamma}{8}, \mathcal{F}_{\mathcal{G},\gamma}, 2m\right)$ we get:

$$\begin{aligned}\ln\mathcal{N}_1\left(\frac{\epsilon\gamma}{8}, \mathcal{F}_{\mathcal{G},\gamma}, 2m\right) &< 3d\left(\frac{\epsilon\gamma}{64}\right)\log_2\left(\frac{289\gamma}{\epsilon\gamma}\right)\ln\left(\frac{1152em\gamma}{\epsilon\gamma d\left(\frac{\epsilon\gamma}{64}\right)}\right)\\ &= 3d\left(\frac{\epsilon\gamma}{64}\right)\log_2\left(\frac{289}{\epsilon}\right)\ln\left(\frac{1152em}{\epsilon d\left(\frac{\epsilon\gamma}{64}\right)}\right).\end{aligned}$$

Next, we apply Lemma 40 in Appendix G, i.e.,

$$\ln x \leqslant Kx + \ln\frac{1}{Ke}$$

with $x, K > 0$ to the right-hand side:

$$\begin{aligned}\ln\mathcal{N}_1\left(\frac{\epsilon\gamma}{8}, \mathcal{F}_{\mathcal{G},\gamma}, 2m\right) &< \frac{3d\left(\frac{\epsilon\gamma}{64}\right)}{\ln 2}\ln\left(\frac{289}{\epsilon}\right)\left(\ln\left(\frac{1152e}{\epsilon d\left(\frac{\epsilon\gamma}{64}\right)}\right)+\ln m\right)\\ &\leqslant \frac{3d\left(\frac{\epsilon\gamma}{64}\right)}{\ln 2}\ln\left(\frac{289}{\epsilon}\right)\left(\ln\left(\frac{1152e}{\epsilon d\left(\frac{\epsilon\gamma}{64}\right)}\right)+\ln\frac{1}{Ke}+Km\right).\end{aligned}$$

Letting

$$K = \frac{\epsilon^2}{64}\left(\frac{3d\left(\frac{\epsilon\gamma}{64}\right)}{\ln 2}\ln\left(\frac{289}{\epsilon}\right)\right)^{-1}$$

we obtain

$$\ln\mathcal{N}_1\left(\frac{\epsilon\gamma}{8}, \mathcal{F}_{\mathcal{G}}, 2m\right) < \frac{3d\left(\frac{\epsilon\gamma}{64}\right)}{\ln 2}\ln\left(\frac{289}{\epsilon}\right)\ln\left(\frac{565^2}{\epsilon^3}\ln\left(\frac{289}{\epsilon}\right)\right)+\frac{m\epsilon^2}{64}.$$

We apply it to the right-hand side of (2.2):

$$P^m\left(\sup_{g\in\mathcal{G}}\left(L(g)-L_{\gamma,m}(g)\right)>\epsilon\right)\leqslant\exp\left(\ln 2+\frac{3d\left(\frac{\epsilon\gamma}{64}\right)}{\ln 2}\ln\left(\frac{289}{\epsilon}\right)\ln\left(\frac{565^2}{\epsilon^3}\ln\left(\frac{289}{\epsilon}\right)\right)-\frac{m\epsilon^2}{32}\right).$$

To conclude the proof, it suffices to upper bound the right-hand side by δ and solve for m. ■

Mendelson and Vershynin [61] provide the following dimension-free bound in the L_2-norm:

Lemma 8 (After Theorem 1 in [61]) *Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ to $[-M_{\mathcal{F}}, M_{\mathcal{F}}]$ with $M_{\mathcal{F}} \in \mathbb{R}_+$. Let $d(\epsilon) = \epsilon\text{-dim}(\mathcal{F})$. Then, for all $\epsilon \in (0, 2M_{\mathcal{F}}]$,*

$$\mathcal{N}_2(\epsilon, \mathcal{F}, n) \leqslant \left(\frac{14M_{\mathcal{F}}}{\epsilon}\right)^{20d\left(\frac{\epsilon}{96}\right)}. \tag{2.11}$$

Lemma 2 in [36] generalizes this bound to all $p \in \mathbb{N}^*$ as follows.

Lemma 9 (After Lemma 2 in [36]) *Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ into $[-M_{\mathcal{F}}, M_{\mathcal{F}}]$ with $M_{\mathcal{F}} \in \mathbb{R}_+^*$. Let $d(\epsilon) = \epsilon\text{-dim}(\mathcal{F})$ for $\epsilon \in (0, M_{\mathcal{F}}]$. Then for any $\epsilon \in (0, 2M_{\mathcal{F}}]$,*

$$\mathcal{N}_p(\epsilon, \mathcal{F}, n) \leqslant 2^{2(K_\epsilon(p)+1)} \left(\frac{6272 e K_\epsilon(p)}{3} \left(\frac{2M_{\mathcal{F}}}{\epsilon} \right)^{2p+1} \right)^{2K_\epsilon(p) d\left(\frac{\epsilon}{45}\right)}.$$

The application of this bound with $p = 1$ gives us the following sample complexity estimate.

Theorem 6 *Fix $\epsilon, \delta \in (0,1)$ and fix $\gamma \in (0,1]$. For $\epsilon \in (0, \gamma]$, let $d(\epsilon) = \epsilon\text{-dim}(\mathcal{F}_{\mathcal{G},\gamma})$. Then, for a sample size*

$$m \geqslant \frac{32}{\epsilon^2} \left(12 d\left(\frac{\epsilon\gamma}{360}\right) \ln\left(\frac{897}{\epsilon}\right) \ln\left(\left(\frac{654}{\epsilon}\right)^3 \ln\left(\frac{897}{\epsilon}\right) \right) + \ln\frac{2}{\delta} \right),$$

the probability (2.1) *is at most δ.*

Proof Applying Lemma 9 with $p = 1$ to $\mathcal{N}_1\left(\frac{\epsilon\gamma}{8}, \mathcal{F}_{\mathcal{G},\gamma}, 2m\right)$ yields:

$$\mathcal{N}_1\left(\frac{\epsilon\gamma}{8}, \mathcal{F}_{\mathcal{G},\gamma}, 2m\right) \leqslant \left(\frac{12544 \cdot 16^3 e K_{\frac{\epsilon\gamma}{8}}(1)}{3\epsilon^3} \right)^{2K_{\frac{\epsilon\gamma}{8}}(1) d\left(\frac{\epsilon\gamma}{360}\right)}, \tag{2.12}$$

where $K_{\frac{\epsilon\gamma}{8}}(1) = \left\lceil 3 \log_2 \left\lceil \frac{896}{\epsilon} \right\rceil \right\rceil$. Performing straightforward computations in the right-hand side of Inequality (2.12) and proceeding as in the proof of Lemma 5 yields the claimed bound.

■

Now, we can as well make use of Lemma 8, based on the norm ordering of covering numbers, Inequality (1.5). It produces the following result.

Theorem 7 *Fix $\epsilon, \delta \in (0,1)$ and fix $\gamma \in (0,1]$. For $\epsilon \in (0, \gamma]$, let $d(\epsilon) = \epsilon\text{-dim}(\mathcal{F}_{\mathcal{G},\gamma})$. Then, for a sample size*

$$m \geqslant \frac{32}{\epsilon^2} \left(20 d\left(\frac{\epsilon\gamma}{768}\right) \ln\left(\frac{112}{\epsilon}\right) + \ln\frac{2}{\delta} \right),$$

the probability (2.1) *is at most δ.*

The following formula summarizes the sample complexity results up to a $\ln(2/\delta)$ term that we obtained based on different combinatorial bounds:

$$m \geqslant \frac{32}{\epsilon^2} \begin{cases} 10 d\left(\frac{\epsilon\gamma}{64}\right) \ln\left(\frac{289}{\epsilon}\right) \ln\left(\frac{565^2}{\epsilon^3} \ln\left(\frac{289}{\epsilon}\right)\right), & L_1^*\text{-norm;} \\ 12 d\left(\frac{\epsilon\gamma}{360}\right) \ln\left(\frac{897}{\epsilon}\right) \ln\left(\frac{654^3}{\epsilon^3} \ln\left(\frac{897}{\epsilon}\right)\right), & L_1\text{-norm;} \\ 20 d\left(\frac{\epsilon\gamma}{768}\right) \ln\left(\frac{112}{\epsilon}\right), & L_2\text{-norm,} \end{cases}$$

where the asterisk indicates the sample-size dependent combinatorial bound. One can see that the L_1-norm dimension-free bound does not provide better sample complexity result: on the contrary, it worsens the constants. The results obtained based on the L_1-norm combinatorial bounds provide the matching dependency on ϵ with that in Inequality (5) of Bartlett and Long [12]. On the other hand, using the L_2-norm bound of [61], one can improve the dependency on ϵ by a factor of $\ln\left(\frac{1}{\epsilon^3}\ln\left(\frac{1}{\epsilon}\right)\right)$. This, however, decreases the scale of the fat-shattering dimension, thus increasing the dimension itself. In all cases, the fat-shattering dimensions that appear in our sample complexity results involve the margin parameter γ in their scales. This is the main difference with the work of Bartlett and Long.

2.3 Conclusions

This chapter focused on bounding the probability of the uniform deviation $\sup_{g\in\mathcal{G}}(L(g) - L_{\gamma,m}(g))$ in terms of an L_1-norm covering number. The technique to control the uniform deviation between the true and empirical means in the L_1-metric is due to Pollard [68]. Based on Pollard's work, Bartlett and Long [12] derived a faster rate of convergence for the uniform deviation. They obtained a sample complexity estimate for the uniform convergence based on the L_1-norm generalization of the bound of Alon et al. [1]. We generalized their uniform deviation result, Lemma 10 combined with Lemma 11 in [10], to the multi-category case. Then, we derived several sample complexity estimates using the L_1 and L_2-norm bounds [36, 61] which included dimension-free ones and the one depending on the sample-size. We observed that the dimension free L_1-norm bound does not provide smaller (thus better) sample complexity estimate. In fact, both combinatorial bounds led to a $\ln\left(\frac{1}{\epsilon}\right)\ln\left(\frac{1}{\epsilon^3}\ln\left(\frac{1}{\epsilon}\right)\right)$ dependency and this matches with that of Bartlett and Long. We demonstrated that making use of the dimension-free L_2-norm combinatorial bound, one can improve the sample complexity estimate by as much as a factor $\ln\left(\frac{1}{\epsilon^3}\ln\left(\frac{1}{\epsilon}\right)\right)$. This, however, increases the fat-shattering dimension, thus deteriorating the bound. The main difference with the work of Bartlett and Long lies in the fact that the fat-shattering dimensions in our sample complexity estimates involve the margin parameter γ. This is the consequence of the use of a margin loss function.

In the next chapter, we consider the deviation of $\sup_{g\in\mathcal{G}}(L(g) - L_{\gamma,m}(g))$ from its expectation. Handling this question via a concentration inequality and the symmetrization technique leads to an error bound involving a Rademacher complexity. The next chapter takes the form of a literature review of error bounds where the dependency on the number of categories is elaborated via the decompositions of this capacity measure.

Chapter 3

Basic Generalization Bound with a Rademacher Complexity

In the preceding chapter, we controlled the uniform deviation between the risk and the empirical margin risk by a covering number. This chapter considers an upper bound on the deviation of interest by another capacity measure, the Rademacher complexity, and constitutes the starting point of scheme (4) given in Introduction:

Rademacher complexity $\xrightarrow{\text{chaining}}$ metric entropy $\xrightarrow{\text{combinatorial bound}}$ fat-shattering dimension.

The focus of this and the following chapters is on the study of the dependencies of the Rademacher complexity on the basic parameters: the number C of categories, the sample size m and the margin parameter γ. The dependency on C is elaborated via a particular result, the decomposition result, that allows one to upper bound a capacity measure of the margin class in terms of that of component classes. In the present chapter, we give a review of the decomposition results in the literature for the Rademacher complexity. For the sample-size dependency, we discuss the Rademacher complexity bounds of several well-known classifiers.

3.1 Basic Generalization Inequality

Before the use of the Rademacher complexity as a capacity measure in learning theory, Talagrand [81] established a far-reaching concentration inequality for the suprema of empirical processes. According to his result, the supremum of an empirical process indexed by a uniformly bounded function class is concentrated around its expectation as follows.

Theorem 8 (After Theorem 4.1 in [81]) *Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ to $[-M_{\mathcal{F}}, M_{\mathcal{F}}]$ with $M_{\mathcal{F}} > 0$. Let a random variable T be an identity map on $(\mathcal{T}, \mathcal{A}_{\mathcal{F}}, P_{\mathcal{T}})$ and let $(T_i)_{1 \leqslant i \leqslant n}$ be a*

sequence of n independent copies of T. Let $F = \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} f(T_i)$ and $v = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{i=1}^{n} f^2(T_i)\right]$. Then, for any $\epsilon > 0$,

$$P_{\mathcal{T}}^{n}\left(F - \mathbb{E}\left[F\right] \geqslant \epsilon\right) \leqslant K \exp\left(-\frac{1}{K'}\frac{\epsilon}{M_{\mathcal{F}}} \ln\left(1 + \frac{\epsilon M_{\mathcal{F}}}{v}\right)\right),$$

where K and K' are constants.

In fact, Talagrand's result answered affirmatively the once long-standing question regarding the existence of the functional form of Bernstein's inequality (Inequality (B.8) in Appendix B). The quest to make the constants in his concentration inequality explicit led to modified versions of this inequality. This line of work has been started by Massart (see Theorem 3 in [56]). Below is a more refined version due to Bartlett, Bousquet and Mendelson [9] where the Rademacher complexity appears thanks to the use of the symmetrization for expectation technique [30] (see also Section 2.3 of [86]).

Theorem 9 (After Theorem 2.1 in [9]) *Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ to $[-M_{\mathcal{F}}, M_{\mathcal{F}}]$ with $M_{\mathcal{F}} > 0$. Let a random variable T be an identity map on $(\mathcal{T}, \mathcal{A}_{\mathcal{F}}, P_{\mathcal{T}})$ and let $(T_i)_{1\leqslant i\leqslant n}$ be a sequence of n independent copies of T. Assume that there is $v > 0$ such that for any $f \in \mathcal{F}$, $var(f(T)) \leqslant v$ almost surely. Fix $\delta \in (0,1)$. Then, with probability $P_{\mathcal{T}}^{n}$ at least $1-\delta$, for all $f \in \mathcal{F}$,*

$$\mathbb{E}\left[f(T)\right] \leqslant \frac{1}{n}\sum_{i=1}^{n} f(T_i) + \inf_{\alpha\geqslant 0}\left(2(1+\alpha)R_n(\mathcal{F}) + \sqrt{\frac{2v}{n}\ln\left(\frac{1}{\delta}\right)} + 2M_{\mathcal{F}}\left(\frac{1}{3} + \ln\left(\frac{1}{\delta}\right)\frac{1}{n\alpha}\right)\right).$$

Notice that this bound emphasizes the role of the variance of functions in a class, thus it is only useful when subsets of functions with small variance are considered. For the entire class, on the other hand, using a classical concentration inequality, i.e., McDiarmid's inequality (Inequality (B.9) in Appendix B), leads to a more efficient bound. In particular, for the empirical process of interest in this thesis, $\{L(g) - L_{\gamma,m}(g) : g \in \mathcal{G}\}$ with fixed $\gamma \in (0,1]$, the application of McDiarmid's inequality and the symmetrization for expectation technique gives the following basic generalization bound involving a Rademacher complexity:

Theorem 10 (After Theorem 8.1 in [63]) *Let $\mathcal{G}$ be as in Definition 3. For $\gamma \in (0,1]$, let $\mathcal{F}_{\mathcal{G},\gamma}$ be the class of functions satisfying Definition 8. Fix $\delta \in (0,1)$ and $\gamma \in (0,1]$. Then, with probability P^m at least $1-\delta$,*

$$\forall g \in \mathcal{G}, \quad L\left(g\right) \leqslant L_{\gamma,m}\left(g\right) + \frac{2}{\gamma}R_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2m}}. \tag{3.1}$$

Now, the question of interest is to elaborate the dependency of the Rademacher complexity on the number of categories and the sample size to which the upcoming sections are dedicated.

3.2 Decomposition of the Rademacher Complexity

To the best of our knowledge, the first decomposition result for the Rademacher complexity goes back to the work of Koltchinskii and Panchenko [47]. Assuming that all component classes are the same (which is usually the case), their bound admits a quadratic dependency on the number of categories.

Lemma 10 (After a partial result in the proof of Theorem 11 in [47]) *Let $\mathcal{G}$ be a class of functions satisfying Definition 3, and let $\mathcal{F}_{\mathcal{G}}$ be the class of functions deduced from $\mathcal{G}$ according to Definition 8. Then,*

$$R_m\left(\mathcal{F}_{\mathcal{G}}\right) \leqslant \frac{C}{2} \sum_{k=1}^{C} R_m\left(\mathcal{G}_k\right). \tag{3.2}$$

Proof It holds that

$$\begin{aligned} R_m\left(\mathcal{F}_{\mathcal{G}}\right) &= \frac{1}{2m} \mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g \in \mathcal{G}} \sum_{i=1}^{m} \sigma_i \left(g_{y_i}(x_i) - \max_{k \neq y_i} g_k(x_i)\right)\right] \\ &= \frac{1}{2m} \mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g \in \mathcal{G}} \sum_{i=1}^{m} \sum_{k=1}^{C} \sigma_i \left(g_k(x_i) - \max_{q \neq k} g_q(x_i)\right) \mathbb{1}_{\{k=y_i\}}\right] \\ &\leqslant \frac{1}{2m} \sum_{k=1}^{C} \mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g \in \mathcal{G}} \sum_{i=1}^{m} \sigma_i \left(g_k(x_i) - \max_{q \neq k} g_q(x_i)\right) \left(\frac{\epsilon_i}{2} + \frac{1}{2}\right)\right], \end{aligned} \tag{3.3}$$

where $\epsilon_i = 2\mathbb{1}_{\{k=y_i\}} - 1$. Denote the k-th summand in the right-hand side of (3.3) by I_k. Note that since $\epsilon_i \in \{-1, 1\}$, σ_i and $\epsilon_i \sigma_i$ follow the same distribution. Thus, for all k,

$$\begin{aligned} I_k &\leqslant \frac{1}{2} \mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g \in \mathcal{G}} \sum_{i=1}^{m} \sigma_i \epsilon_i \left(g_k(x_i) - \max_{q \neq k} g_q(x_i)\right)\right] + \frac{1}{2} \mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g \in \mathcal{G}} \sum_{i=1}^{m} \sigma_i \left(g_k(x_i) - \max_{q \neq k} g_q(x_i)\right)\right] \\ &= \mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g \in \mathcal{G}} \sum_{i=1}^{m} \sigma_i \left(g_k(x_i) - \max_{q \neq k} g_q(x_i)\right)\right]. \end{aligned}$$

Next, by sub-additivity of supremum and Lemma 29 in Appendix A,

$$\begin{aligned} I_k &\leqslant \mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g \in \mathcal{G}} \sum_{i=1}^{m} \sigma_i g_k(x_i)\right] + \mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g \in \mathcal{G}} \sum_{i=1}^{m} \sigma_i \max_{q \neq k} g_q(x_i)\right] \\ &\leqslant \mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g \in \mathcal{G}} \sum_{i=1}^{m} \sigma_i g_k(x_i)\right] + \sum_{q \neq k} \mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g \in \mathcal{G}} \sum_{i=1}^{m} \sigma_i g_q(x_i)\right] \\ &= \sum_{q=1}^{C} \mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g \in \mathcal{G}} \sum_{i=1}^{m} \sigma_i g_q(x_i)\right]. \end{aligned}$$

Substituting this bound into (3.3) concludes the proof. ■

This decomposition was refined by Kuznetsov et al. [53] by partially truncating the co-domain

of $\mathcal{F}_{\mathcal{G}}$, i.e., truncating all values above $\gamma \in (0,1]$, which improved the quadratic dependency of Lemma 10 to a linear one. This gain is based on the following line of reasoning. Since

$$\mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{m}\sigma_i\left(g_{y_i}(x_i)-\max_{k\neq y_i}g_k(x_i)\right)\right]\leqslant\mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{m}\sigma_i g_{y_i}(x_i)\right]+\mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{m}\sigma_i\max_{k\neq y_i}g_k(x_i)\right],$$

the first term can be upper bounded by $\sum_{k=1}^{C}\hat{R}_m\left(\mathcal{G}_k\right)$ as in the proof above with $g_{y_i}(x_i)-\max_{k\neq y_i}g_k(x_i)$ replaced by $g_{y_i}(x_i)$. Now, to get a similar upper bound on the second term, one needs to take the maximum over all component functions so as using Lemma 29 to obtain

$$\mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{m}\sigma_i\max_{1\leqslant k\leqslant C}g_k(x_i)\right]\leqslant\sum_{k=1}^{C}\mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{m}\sigma_i g_k(x_i)\right].$$

Now, this is precisely where the usefulness of truncating the co-domain of $\mathcal{F}_{\mathcal{G}}$ lies. This leads to the following bound involving the class $\mathcal{F}_{\mathcal{G},\gamma}$.

Lemma 11 (After a partial result in the proof of Theorem 3 in [53]) *Let $\mathcal{G}$ be a class of functions satisfying Definition 3. For $\gamma\in(0,1]$, let $\mathcal{F}_{\mathcal{G},\gamma}$ be the class of functions deduced from $\mathcal{G}$ according to Definition 8. Then*

$$R_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right)\leqslant\sum_{k=1}^{C}R_m\left(\mathcal{G}_k\right). \tag{3.4}$$

Proof For any $g\in\mathcal{G}$ and for any $\gamma\in(0,1]$, the functions $f_{g,\gamma}$ can be re-written as follows:

$$\forall(x,y)\in\mathcal{Z},\quad f_{g,\gamma}\left(z\right)=\max\left(0,\frac{1}{2}\left(g_y(x)-\max_{1\leqslant k\leqslant C}\left(g_k(x)-2\gamma\mathbb{1}_{\{k=y\}}\right)\right)\right).$$

Note that $\max(0,\cdot)$ is a 1-Lipschitz function:

$$\forall t,t'\in\mathbb{R},\quad\left|\max(0,t)-\max(0,t')\right|\leqslant\max\left(0,\left|t-t'\right|\right)\leqslant\left|t-t'\right|.$$

Then, according to Lemma 4 and by sub-additivity of the supremum function,

$$\begin{aligned}\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right)=&\ \frac{1}{2m}\mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{m}\sigma_i\max\left(0,g_{y_i}\left(x_i\right)-\max_{1\leqslant k\leqslant C}\left(g_k\left(x_i\right)-2\gamma\mathbb{1}_{\{k=y_i\}}\right)\right)\right]\\ \leqslant&\ \frac{1}{2m}\mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{m}\sigma_i\left(g_{y_i}\left(x_i\right)-\max_{1\leqslant k\leqslant C}\left(g_k\left(x_i\right)-2\gamma\mathbb{1}_{\{k=y_i\}}\right)\right)\right]\\ \leqslant&\ \frac{1}{2m}\mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{m}\sigma_i g_{y_i}\left(x_i\right)\right]+\frac{1}{2m}\mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{m}\sigma_i\max_{1\leqslant k\leqslant C}\left(g_k\left(x_i\right)-2\gamma\mathbb{1}_{\{k=y_i\}}\right)\right].\end{aligned} \tag{3.5}$$

Denote the terms in the right-hand side of (3.5) by T_1 and T_2, respectively. T_1 can be bounded as follows:

$$\begin{aligned}
T_1 &= \frac{1}{2m}\mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{m}\sum_{k=1}^{C}\sigma_i g_k\left(x_i\right)\mathbb{1}_{\{k=y_i\}}\right] \\
&\leqslant \frac{1}{2m}\sum_{k=1}^{C}\mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{m}\sigma_i g_k\left(x_i\right)\mathbb{1}_{\{k=y_i\}}\right] \\
&\leqslant \frac{1}{2m}\sum_{k=1}^{C}\mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{m}\sigma_i g_k\left(x_i\right)\left(\frac{\epsilon_i}{2}+\frac{1}{2}\right)\right] \\
&\leqslant \frac{1}{4m}\sum_{k=1}^{C}\left(\mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{m}\sigma_i\epsilon_i g_k\left(x_i\right)\right]+\mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{m}\sigma_i g_k\left(x_i\right)\right]\right) \\
&\leqslant \frac{1}{2}\sum_{k=1}^{C}\hat{R}_m\left(\mathcal{G}_k\right), \qquad (3.6)
\end{aligned}$$

where we used the fact that σ_i and $\epsilon_i\sigma_i$ follow the same distribution. Now, the term T_2 can be bounded using Lemma 29 and the fact that Rademacher variables are centered:

$$T_2 \leqslant \sum_{k=1}^{C}\frac{1}{2m}\mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{m}\sigma_i\left(g_k\left(x_i\right)-2\gamma\mathbb{1}_{\{k=y_i\}}\right)\right] = \frac{1}{2}\sum_{k=1}^{C}\hat{R}_m\left(\mathcal{G}_k\right).$$

Combining it with (3.6) in (3.5) yields the claimed bound. ■

If the decomposition of Kuznetsov and coauthors is dedicated to truncated margin functions, the result recently introduced by Maurer [57] concerns a class of Lipschitz functions with vector-valued domains. Apart from the Lipschitz property, his result makes use of the classical Khintchine's inequality:

Lemma 12 (Khintchine's inequality, after Lemma 4.1 in [54]) *Let $p \in \mathbb{R}_+$ and let $\boldsymbol{\sigma} = (\sigma_i)_{i\geqslant 1}$ be a Rademacher sequence. There exists a constant $K_p > 0$ depending only on p, such that for any $(t_i)_{i\geqslant 1} \in l_2$,*

$$\left(\sum_{i\geqslant 1} t_i^2\right)^{\frac{1}{2}} \leqslant K_p\left(\mathbb{E}_{\boldsymbol{\sigma}}\left|\sum_{i\geqslant 1}\sigma_i t_i\right|^p\right)^{\frac{1}{p}}.$$

To apply Maurer's result in the context of our work, we make use of the following result that replaces the required Lipschitz property.

Lemma 13 (After Lemma A.3 in [10]) *Fix $(x, y) \in \mathcal{X}\times[\![1, C]\!]$. Let $\mathcal{G}$ be as in Definition 3. Then, for any $g, g' \in \mathcal{G}$ and for any $p \in \mathbb{N}^*$,*

$$\left|f_g\left(x,y\right) - f_{g'}\left(x,y\right)\right| \leqslant \left\|g(x) - g'(x)\right\|_p.$$

Since the functions $\max(0, \cdot)$ and $\min(\gamma, \cdot)$ are 1-Lipschitz, Lemma 13 also holds for truncated margin functions. Thus, Maurer's result remains unchanged when one switches from the class $\mathcal{F}_{\mathcal{G}}$ to $\mathcal{F}_{\mathcal{G},\gamma}$. Particularly, the following bound holds (the proof is given in Appendix A):

Lemma 14 (After Corollary 1 in [57]) *Let $\mathcal{G}$ be a class of functions satisfying Definition 3. For $\gamma \in (0, 1]$, let $\mathcal{F}_{\mathcal{G},\gamma}$ be the class of functions deduced from $\mathcal{G}$ according to Definition 8. Let $\boldsymbol{\sigma}_{Cm} = (\sigma_i)_{1 \leqslant i \leqslant Cm}$ denote a Rademacher sequence of length $C \times m$. Then,*

$$R_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \frac{\sqrt{2}}{m} \mathbb{E}_{\boldsymbol{\sigma}_{Cm}} \sup_{g \in \mathcal{G}} \sum_{i=1}^{m} \sum_{k=1}^{C} \sigma_{C(i-1)+k} g_k(x_i).$$

Maurer's result is an improvement over that of Lei et al. [55] whose decomposition is in terms of the Gaussian complexity:

Theorem 11 (After Theorem 5 in [55]) *Let $\mathcal{G}$ be a class of functions satisfying Definition 3. For $\gamma \in (0, 1]$, let $\mathcal{F}_{\mathcal{G},\gamma}$ be the class of functions deduced from $\mathcal{G}$ according to Definition 8. Let $\boldsymbol{o}_{Cm} = (o_j)_{1 \leqslant j \leqslant Cm}$ be an orthogaussian sequence of length $C \times m$. Then*

$$R_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \frac{1}{m} \sqrt{\frac{\pi}{2}} \mathbb{E}_{\boldsymbol{o}_{Cm}} \left[\sup_{g \in \mathcal{G}} \sum_{i=1}^{m} \sum_{k=1}^{C} o_{C(i-1)+k} g_k\left(x_i\right) \right]. \tag{3.7}$$

The advantage of Lemma 14, which also holds true for Theorem 11), over Lemma 11 is that it allows the coupling between the component functions g_k to be taken into account. In particular, for linear classifiers, employing a specific coupling assumption leads to a generalization bound with a sublinear dependency on the number of categories (see Chapter 4). If no coupling is assumed, as it is the case for the main contributions of the present thesis, then one recovers Lemma 11 up to a constant factor.

To elaborate the dependency on the sample size, one could specify the component classes $\mathcal{G}_k$ and upper bound the corresponding Rademacher complexity. The following section demonstrates the technique due to Bartlett and Mendelson [13] to upper the Rademacher complexity of linear and kernel classifiers. This technique can be used to upper bound that of feedforward neural networks with multiple layers.

3.3 Dependency on the Sample Size: Standard Classifiers

For several well-known classifiers, such as support vector machines [20], feedforward neural networks [4], decision trees [19, 71], the Rademacher/Gaussian complexity bounds have been derived by Bartlett and Mendelson [13], one of the early works promoting the use of these measures in

learning theory. These results admit the optimal, $O\left(m^{-\frac{1}{2}}\right)$, dependency on the sample size. In particular, the lemma below is an upper bound on the Rademacher complexity of linear and kernel methods. It can be used as the basis to derive Rademacher complexity bounds for function classes that can be expressed as combinations of functions from simpler classes, such as feedforward neural networks.

Lemma 15 (After Lemma 2 in [13]) *Let $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a reproducing kernel and let $\left(\mathcal{H}_\kappa, \|\cdot\|_{\mathcal{H}_\kappa}\right)$ be the corresponding* reproducing kernel Hilbert space (RKHS) *[16]: for any $x \in \mathcal{X}$, $\kappa_x = \kappa(\cdot, x) \in \mathcal{H}_\kappa$, and for any $h \in \mathcal{H}_\kappa$, $h(x) = \langle h, \kappa_x \rangle$. Let $\Lambda_{\mathcal{X}}, \Lambda \in \mathbb{R}_+$. Assume that $\sup_{x \in \mathcal{X}} \|\kappa_x\|_{\mathcal{H}_\kappa} \leqslant \Lambda_{\mathcal{X}}$. Let*

$$B_\Lambda\left(\mathcal{H}_\kappa\right) = \left\{h \in \mathcal{H}_\kappa : \ \|h\|_{\mathcal{H}_\kappa} \leqslant \Lambda\right\}.$$

Then,

$$R_n\left(B_\Lambda\left(\mathcal{H}_\kappa\right)\right) \leqslant \frac{\Lambda \Lambda_{\mathcal{X}}}{\sqrt{n}}.$$

Proof For any $(x_i)_{1 \leqslant i \leqslant n} \in \mathcal{X}^n$, by Cauchy-Schwarz and Jensen's inequalities,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\sigma}_n} \sup_{h \in B_\Lambda(\mathcal{H}_\kappa)} \sum_{i=1}^n \sigma_i h(x_i) &= \mathbb{E}_{\boldsymbol{\sigma}_n} \sup_{h \in B_\Lambda(\mathcal{H}_\kappa)} \sum_{i=1}^n \sigma_i \langle h, \kappa_{x_i} \rangle \\
&\leqslant \sup_{h \in B_\Lambda(\mathcal{H}_\kappa)} \|h\|_{\mathcal{H}_\kappa} \, \mathbb{E}_{\boldsymbol{\sigma}_n} \sqrt{\sum_{i,j} \sigma_i \sigma_j \langle \kappa_{x_i}, \kappa_{x_j} \rangle} \\
&\leqslant \Lambda \sqrt{\mathbb{E}_{\boldsymbol{\sigma}_n} \sum_{i,j} \sigma_i \sigma_j \langle \kappa_{x_i}, \kappa_{x_j} \rangle} \\
&\leqslant \Lambda \sqrt{\sum_{i=1}^n \langle \kappa_{x_i}, \kappa_{x_i} \rangle} \\
&\leqslant \sqrt{n} \Lambda \Lambda_{\mathcal{X}}.
\end{aligned}$$

■

Theorem 18 in [13] provides an upper bound on the Gaussian complexity of the class of two-layer neural networks. By using the lemma above and the contraction principle, Lemma 4, one can upper bound the Rademacher complexity of the class of deep networks with the $\|\cdot\|_1$-norm constraints on weights [94]. This bound exhibits an exponential dependency on the number of layers and the optimal dependency on the sample size.

Lemma 16 (After Lemma 2 in [94]) *Let $\Lambda_{\mathcal{X}}, \Lambda \in \mathbb{R}_+$. Let $\mathcal{X} \subset \mathbb{R}^d$ with $\sup_{x \in \mathcal{X}} \|x\|_2 \leqslant \Lambda_{\mathcal{X}}$. For any $j \in [\![1, l]\!]$ with $l \geqslant 2$, let $W^{(j)}$ be a $d_j \times d_{j-1}$ weight matrix with $d_0 = d$ and $d_l = 1$. Fix*

$j \in [\![1, l]\!]$ and assume that for all $r \in [\![1, d_j]\!]$, $\left\|W_{r.}^{(j)}\right\|_1 = \sum_{q=1}^{d_{j-1}} \left|W_{rq}^{(j)}\right| \leqslant \Lambda$, where $W_{r.}^{(j)}$ denotes the r-th row of $W^{(j)}$. Let $\alpha : \mathbb{R} \to [-M, M]$ with $M > 0$ be an L-Lipschitz function. For any $\mathbf{t} = (t_i)_{1 \leqslant i \leqslant d_j} \in \mathbb{R}^{d_j}$, let $\alpha^{(j)} : \mathbf{t} \mapsto \left(\alpha(t_1), \ldots, \alpha(t_{d_j})\right)^T$. Define the class of l-layer networks as

$$H_l = \left\{ x \mapsto \alpha^{(l)}\left(W^{(l)}\alpha^{(l-1)}\left(W^{(l-1)}\ldots\alpha^{(1)}\left(W^{(1)}x\right)\ldots\right)\right) \,:\, x \in \mathbb{R}^d,\ \left\|W_{r.}^{(j)}\right\|_1 \leqslant \Lambda \right\}.$$

Then,

$$R_n\left(H_l\right) \leqslant \frac{L^l \Lambda^l \Lambda_{\mathcal{X}}}{\sqrt{n}}.$$

Proof For $l = 2$, we have $H_2 = \{x \mapsto \alpha^{(2)}\left(W^{(2)}\alpha^{(1)}\left(W^{(1)}x\right)\right)\}$. Then, by the contraction lemma (Lemma 4) and Lemma 15, we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\sigma}_n} \sup_{h \in H_2} \sum_{i=1}^n \sigma_i \alpha^{(2)}\left(W^{(2)}\alpha^{(1)}\left(W^{(1)}x_i\right)\right) &\leqslant L\mathbb{E}_{\boldsymbol{\sigma}_n} \sup_{h \in H_2} \sum_{i=1}^n \sigma_i W^{(2)}\alpha^{(1)}\left(W^{(1)}x_i\right) \\
&\leqslant L\mathbb{E}_{\boldsymbol{\sigma}_n} \sup_{h \in H_2} \sum_{i=1}^n \sigma_i \sum_{j=1}^{d_1} W_{1j}^{(2)} \alpha\left(\sum_{q=1}^d W_{jq}^{(1)} x_i^{(q)}\right) \\
&\leqslant L\mathbb{E}_{\boldsymbol{\sigma}_n} \sup_{h \in H_2} \sum_{j=1}^{d_1} W_{1j}^{(2)} \sum_{i=1}^n \sigma_i \alpha\left(\sum_{q=1}^d W_{jq}^{(1)} x_i^{(q)}\right).
\end{aligned}$$

Since $\left\|W^{(2)}\right\|_1 = \sum_{j=1}^{d_1} \left|W_{1j}^{(2)}\right| \leqslant \Lambda$, it follows that

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\sigma}_n} \sup_{h \in H_2} \sum_{i=1}^n \sigma_i \alpha^{(2)}\left(W^{(2)}\alpha^{(1)}\left(W^{(1)}x_i\right)\right) &\leqslant n\Lambda L \underbrace{\mathbb{E}_{\boldsymbol{\sigma}_n} \sup_{h \in H_1} \frac{1}{n} \sum_{i=1}^n \sigma_i \alpha\left(W^{(1)}x_i\right)}_{\hat{R}_n(H_1)} \\
&\leqslant \sqrt{n} L^2 \Lambda^2 \Lambda_{\mathcal{X}}.
\end{aligned}$$

Now, assume that the claim holds for $l-1$ layers. Then proving it for l layers following the reasoning above is straightforward:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\sigma}_n} \sup_{h \in H_l} \sum_{i=1}^n \sigma_i \alpha^{(l)}\left(W^{(l)}\alpha^{(l-1)}\left(W^{(l-1)}\ldots\alpha^{(1)}\left(W^{(1)}x_i\right)\ldots\right)\right) \\
&\quad\leqslant \Lambda L \mathbb{E}_{\boldsymbol{\sigma}_n} \sup_{h \in H_l} \sum_{i=1}^n \sigma_i \frac{W^{(l)}}{\left\|W^{(l)}\right\|_1} \alpha^{(l-1)}\left(W^{(l-1)}\ldots\alpha^{(1)}\left(W^{(1)}x_i\right)\ldots\right) \\
&\quad\leqslant n\Lambda L \underbrace{\mathbb{E}_{\boldsymbol{\sigma}_n} \sup_{h \in H_{l-1}} \frac{1}{n} \sum_{i=1}^n \sigma_i \alpha\left(W^{(l-1)}\ldots\alpha^{(1)}\left(W^{(1)}x_i\right)\ldots\right)}_{\hat{R}_n(H_{l-1})} \\
&\quad\leqslant (\Lambda L) \cdot \sqrt{n} L^{l-1} \Lambda^{l-1} \Lambda_{\mathcal{X}} \\
&\quad= \sqrt{n} L^l \Lambda^l \Lambda_{\mathcal{X}}.
\end{aligned}$$

■

Although much work is dedicated to the study of the generalization performance of deep neural networks [10, 66, 67], the analysis of the Rademacher complexities of networks with particular architectures, such as convolutional neural networks [50], is an open question.

3.4 Conclusions

This chapter considered the basic generalization bound for margin multi-category classifiers involving a Rademachar complexity as capacity measure. It constitutes the starting point for the main body of work of the subsequent chapters where the focus is on optimizing the dependencies of the Rademacher complexity on C, m and γ. In this chapter, if the dependency on γ was made explicit through the well-known contraction lemma, the one on C was based on the use of a particular result, the decomposition result, that relates a capacity measure of the margin class to the ones of the component classes. The literature provides decomposition results for the Rademacher complexity that exhibit at best $O(C)$ dependency when no coupling between the component functions is assumed, as it is the focus of this thesis. To make explicit the dependency on the sample size, after the decomposition, one could specify the classifier and bound the Rademacher complexity of the correspondig function class. For the classical classifiers, such as support vector machines and feedforward neural networks, the rate of convergence is optimal: $O\left(m^{-\frac{1}{2}}\right)$.

In the next chapter, we relate the Rademacher complexity to the metric entropy and postpone the decomposition to the level of the latter quantity. We pose the same question: what are the forms of the dependencies on C, m and γ when one manipulates the metric entropy instead of the Rademacher complexity?

Chapter 4

From Rademacher Complexity to Metric Entropy

This chapter relates the Rademacher complexity to another capacity measure, the metric entropy (the logarithm of the covering number), and studies how the decomposition at the level of the metric entropy influences the dependencies on the basic parameters in the context of the following pathway:

$$\text{Rademacher complexity} \xrightarrow{\text{chaining}} \textbf{metric entropy} \xrightarrow{\text{combinatorial bound}} \text{fat-shattering dimension.}$$

The chaining method allows one to control the supremum of a stochastic process in terms of the sum of the metric entropies of its index set. This method was pioneered by Kolmogorov, and extended to an abstract setting by Dudley in the late sixties [83]. Dudley's result applies to a (centered) stochastic process $\mathscr{G}_{\mathcal{S}} = \{\mathscr{G}_s \ : \ s \in (\mathcal{S}, \rho)\}$ indexed by a subset of a (pseudo-)metric space $(\mathcal{S}, \rho)$ which satisfies the following increment condition:

$$\forall \epsilon > 0, \quad P_{\mathcal{T}} \left(|\mathscr{G}_s - \mathscr{G}_{s'}| > \epsilon \right) \leqslant 2 \exp \left(- \frac{\epsilon^2}{2\rho^2 \left(s, s' \right)} \right). \tag{4.1}$$

This result can be formulated as follows:

$$\forall \delta > 0, \quad \mathbb{E} \sup_{\rho(s,s') \leqslant \delta} |\mathscr{G}_s - \mathscr{G}_{s'}| \leqslant K \int_0^{\delta} \sqrt{\ln \mathcal{N} \left(\epsilon, \mathcal{S}, \rho \right)} d\epsilon, \tag{4.2}$$

where K is a universal constant. In the particular case when $\mathscr{G}_{\mathcal{S}}$ is a symmetric process, such as a Gaussian process, it holds that

$$\mathbb{E} \left[\sup_{s,s'} |\mathscr{G}_s - \mathscr{G}_{s'}| \right] = 2 \mathbb{E} \left[\sup_{s} \mathscr{G}_s \right].$$

Then, Inequality (4.2) can be re-written in a handy way as

$$\mathbb{E}\sup_{s}\mathscr{G}_s \leqslant K\int_0^{\operatorname{diam}(\mathcal{S})}\sqrt{\ln\mathcal{N}(\epsilon,\mathcal{S},\rho)}d\epsilon, \tag{4.3}$$

where $\operatorname{diam}(\mathcal{S}) = \sup_{s,s'\in\mathcal{S}}\rho(s,s')$ and the constant K is different from that in Inequality (4.2). A Rademacher process is a sub-Gaussian process (see Appendix B), and thus its supremum can be controlled by Dudley's entropy integral (4.3).

As we discussed in Section 1.5, whether one can compute the integral in Inequality (4.3) is determined by the growth rate of the metric entropy as $\epsilon \to 0$. Instead, we will make use of the following bound on the empirical Rademacher complexity of $\mathcal{F} \subset L_1(\mathcal{T})$ derived in [36] (which can be bounded by the entropy integral (4.3)):

$$\forall \mathbf{t}_n \in \mathcal{T}^n, \quad \hat{R}_n(\mathcal{F}) \leqslant h(N) + 2\sum_{j=1}^{N}(h(j)+h(j-1))\sqrt{\frac{\ln\mathcal{N}(h(j),\mathcal{F},d_{2,\mathbf{t}_n})}{n}}, \tag{4.4}$$

where $N \in \mathbb{N}^*$ and $h : \mathbb{N} \to \mathbb{R}_+$ is a decreasing function with $h(0)$ greater than the diameter of $\mathcal{F}$. In other words, N is the number of steps taken to construct the chaining and $h(j)$ is the radius of (open) balls that cover $\mathcal{F}$ at step j. Note the freedom one has in the choice of the value of N and the function h. Although there is no clear justification for a particular choice of h in the literature, in the sequel we will stick to geometrically decreasing functions.

With this introduction, we can formulate the goal of the present chapter as follows: i) apply the chaining formula (4.4) to $\mathcal{F}_{\mathcal{G},\gamma}$, ii) switch from the metric entropy of $\mathcal{F}_{\mathcal{G},\gamma}$ to that of component classes $\mathcal{G}_k$, and iii) upper bound the obtained chaining formula under the choice of different combinatorial bounds and study the impact on the parameters of interest, particularly, the number C of categories.

To switch from the metric entropy of $\mathcal{F}_{\mathcal{G},\gamma}$ to that of $\mathcal{G}_k$, we use the following decomposition formula due to Guermeur [36]. This result is the generalization of Proposition 6.2 of Duan [23] to L_p-norms.

Lemma 17 (Lemma 1 in [36]) *Let $\mathcal{G}$ be a class of functions satisfying Definition 3. Let $\mathcal{F}_{\mathcal{G}}$ and $\mathcal{F}_{\mathcal{G},\gamma}$ be derived from $\mathcal{G}$ according to Definitions 4 and 8, respectively. Then, for any $p \in [1,\infty]$ and for any $\epsilon > 0$,*

$$\mathcal{N}(\epsilon,\mathcal{F}_{\mathcal{G},\gamma},d_{p,\mathbf{z}_n}) \leqslant \mathcal{N}(\epsilon,\mathcal{F}_{\mathcal{G}},d_{p,\mathbf{z}_n}) \leqslant \prod_{k=1}^{C}\mathcal{N}\left(\frac{\epsilon}{C^{\frac{1}{p}}},\mathcal{G}_k,d_{p,\mathbf{x}_n}\right), \tag{4.5}$$

where $\mathbf{z}_n \in \mathcal{Z}^n$ and $\mathbf{x}_n \in \mathcal{X}^n$.

Since the chaining is constructed in the L_2-metric, only the values of p in $[2,\infty]$ are of relevance in the decomposition formula (4.5).

The second step is based on the following considerations. First, to develop the chaining, we make the assumption that the component classes $\mathcal{G}_k$ have polynomially growing fat-shattering dimensions:

Assumption 1 *Let $\mathcal{G}$ be a class of functions satisfying Definition 3. We assume that there exists a pair $(K_{\mathcal{G}}, d_{\mathcal{G}}) \in \mathbb{R}_+^2$ such that*

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

This assumption is not restrictive, since it holds true for many well-known classifiers: support vector machines with $d_{\mathcal{G}} = 2$ (Theorem 4.6 in [14]), feedforward neural networks with $d_{\mathcal{G}} = 2l$ for l layers (Corollary 27 in [7]), and in general, Lipschitz classifiers on metric spaces, such as nearest neighbours (Corollary 4 in [32]).

For $[0, 1]$-valued function classes with polynomial fat-shattering dimensions, i.e., $\epsilon\text{-dim}\,(\mathcal{F}) \leqslant K_{\mathcal{F}} \epsilon^{-d_{\mathcal{F}}}$ with $\epsilon \in (0, 1]$, Mendelson [59, 60] derived the following result:

$$R_m(\mathcal{F}) \leqslant \frac{K}{\sqrt{m}} \begin{cases} 1, & \text{if } 0 < d_{\mathcal{F}} < 2, \\ \ln^{\frac{3}{2}} m, & \text{if } d_{\mathcal{F}} = 2, \\ m^{\frac{1}{2} - \frac{1}{d_{\mathcal{F}}}} \ln^{\frac{1}{d_{\mathcal{F}}}}(m), & \text{if } d_{\mathcal{F}} > 2, \end{cases}$$

where the constant K depends on the growth rate $d_{\mathcal{F}}$, and $K_{\mathcal{F}}$. One can notice that the chaining bound is developed into three cases based on the rate of growth of the fat-shattering dimension. In other words, the capacity of the function class dictates the computation of the chaining bound. The author rightly so calls this phenomenon the phase transition to underline the abrupt change with respect to the dependency on the sample size m when $d_{\mathcal{F}} = 2$. This result has been extended to the multi-class setting by Guermeur [36] by combining the decomposition formula (4.5) with $p = 2$ and the combinatorial bound of Mendelson and Vershynin [61]. In this extension, the phase transition is also observed with respect to the number of categories and the margin parameter γ:

Theorem 12 (After Theorem 7 in [36]) *Let $\mathcal{G}$ be a class of functions satisfying Assumption 1. For $\gamma \in (0, 1]$, let $\mathcal{F}_{\mathcal{G},\gamma}$ be class of functions deduced from $\mathcal{G}$ according to Definition 8. Then, there is a constant K that depends on $d_{\mathcal{G}}$ and $K_{\mathcal{G}}$ such that*

$$R_m(\mathcal{F}_{\mathcal{G},\gamma}) \leqslant \frac{K}{\sqrt{m}} \begin{cases} C^{\frac{d_{\mathcal{G}}+2}{4}} \gamma^{1-\frac{d_{\mathcal{G}}}{2}} \ln^{\frac{1}{2}} \left(\dfrac{\sqrt{C}}{\gamma} \right), & \text{if } 0 < d_{\mathcal{G}} < 2, \\ C \log_2 \left(\dfrac{m}{C} \right) \ln^{\frac{1}{2}} \left(\dfrac{\sqrt{m}}{\gamma C^{\frac{1}{4}}} \right), & \text{if } d_{\mathcal{G}} = 2, \\ C^{\frac{2+d_{\mathcal{G}}}{2d_{\mathcal{G}}}} \gamma^{1-\frac{d_{\mathcal{G}}}{2}} m^{\frac{1}{2}-\frac{1}{d_{\mathcal{G}}}} \ln^{\frac{1}{2}} \left(\dfrac{1}{\gamma} \left(\dfrac{m}{C} \right)^{\frac{1}{d_{\mathcal{G}}}} \right), & \text{if } d_{\mathcal{G}} > 2. \end{cases}$$

This result improves the dependency of the Rademacher complexity on C over the one obtained via the direct decomposition of this capacity measure in Chapter 3. The bounds that we obtain in the present chapter are similar to Theorem 12. Particularly, we demonstrate that one can further improve the dependency on the number of classes. In Section 4.1, we develop chaining bounds in the L_∞-norm for which the decomposition formula (17) is optimized with respect to C. This yields a radical dependency on C irrespective of the growth rate $d_{\mathcal{G}}$, but slightly worsens the one on m. Our main contribution is in Section 4.2, where we derive new combinatorial bounds (dimension-free and not) by generalizing that of Mendelson and Vershynin [61] to L_p-norms with $p \in \mathbb{N}^* \setminus \{1, 2\}$. These bounds allow us to adapt to the phase-transition phenomenon and improve the dependency on C over that in Theorem 12 without worsening the ones on m and γ.

4.1 L_∞-norm Metric Entropy

The decomposition formula, Inequality (4.5), is optimized with respect to C as p gets larger, and for $p = \infty$, C disappears altogether from the scales of the covering numbers:

$$\mathcal{N}\left(\epsilon, \mathcal{F}_{\mathcal{G},\gamma}, d_{\infty,\mathbf{z}_n}\right) \leqslant \prod_{k=1}^{C} \ln \mathcal{N}\left(\epsilon, \mathcal{G}_k, d_{\infty,\mathbf{x}_n}\right). \tag{4.6}$$

Notice that the the chaining formula (4.4) involves the covering numbers in the L_2-norm, and to apply the decomposition (4.6) to it, we will use the norm ordering of the covering numbers, Inequality (1.5). For any sample $\mathbf{z}_m = ((x_i, y_i))_{1 \leqslant i \leqslant m} \in \mathcal{Z}^m$, this gives:

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant h(N) + 2\sum_{j=1}^{N} \left(h(j) + h(j-1)\right) \sqrt{\frac{\sum_{k=1}^{C} \ln \mathcal{N}\left(\epsilon, \mathcal{G}_k, d_{\infty,\mathbf{x}_m}\right)}{n}}. \tag{4.7}$$

Below, we first derive a chaining bound for general function classes. Then, we consider classes of linear functions, for which we show that by using the corresponding metric entropy results leads to the chaining bound with a better convergence rate.

4.1.1 General Case

For classes of real-valued functions, Alon and co-authors [1] derived the following L_∞-norm combinatorial bound generalizing the classical Sauer-Shelah lemma [74, 75]:

Lemma 18 (After Lemma 3.5 in [1]) *Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ to $[-M_{\mathcal{F}}, M_{\mathcal{F}}]$ with $M_{\mathcal{F}} \in \mathbb{R}_+$. For $\epsilon \in (0, M_{\mathcal{F}}]$, let $d(\epsilon) = \epsilon\text{-}dim\left(\mathcal{F}\right)$. Then, for all $\epsilon \in (0, 2M_{\mathcal{F}}]$,*

$$\mathcal{N}_\infty\left(\epsilon, \mathcal{F}, n\right) \leqslant 2\left(\frac{16M_{\mathcal{F}}^2 n}{\epsilon^2}\right)^{d\left(\frac{\epsilon}{4}\right)\log_2\left(\frac{4M_{\mathcal{F}} en}{d\left(\frac{\epsilon}{4}\right)\epsilon}\right)}.$$

Notice that, in terms of metric entropy, this bound grows as $O(\ln^2 n)$. Theorem 4.4 of Rudelson and Vershynin [73] addresses the conjecture made in [1] concerning the question whether the exponent 2 could be reduced to 1. Their result, which is based on the comparison of the covering number of a set to the number of integer cells contained in it and its projections, reduces this exponent to some value in $(1, 2)$. Since this is achieved at the cost of deteriorating the fat-shattering dimension, in this thesis we will use Lemma 18. Applying it in the chaining formula (4.7) leads to the following bound on the Rademacher complexity:

Theorem 13 *Let $\mathcal{G}$ be as in Definition 3 and, for any $\gamma \in (0, 1]$, let $\mathcal{F}_{\mathcal{G},\gamma}$ be deduced from $\mathcal{G}$ as in Definition 8. Then, under Assumption 1, there is a constant K that depends only on $d_{\mathcal{G}}$ and $K_{\mathcal{G}}$ such that*

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

Proof We set $h(j) = \gamma 2^{-j}$ in the chaining bound (4.7):

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \gamma 2^{-N} + 6\gamma \sum_{j=1}^{N} 2^{-j} \sqrt{\frac{\sum_{k=1}^{C} \ln \mathcal{N}\left(\gamma 2^{-j}, \mathcal{G}_k, d_{\infty,\mathbf{x}_m}\right)}{m}}. \tag{4.8}$$

Let $d\left(\epsilon\right) = \max\limits_{1 \leqslant k \leqslant C} \epsilon\text{-dim}(\mathcal{G}_k)$. Then, applying Lemma 18 to $\mathcal{G}_k$, we have that for all $\epsilon \in (0, 2M_{\mathcal{G}}]$,

$$\ln \mathcal{N}\left(\epsilon, \mathcal{G}_k, d_{\infty,\mathbf{x}_m}\right) \leqslant d\left(\frac{\epsilon}{4}\right) \log_2\left(\frac{4M_{\mathcal{G}} e m}{d\left(\frac{\epsilon}{4}\right)\epsilon}\right) \ln\left(\frac{32M_{\mathcal{G}}^2 m}{\epsilon^2}\right).$$

Applying this bound in (4.8) and using the fact that $\ln(2a) \leqslant 2\ln a$ for all $a \geqslant 2$ result in

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \gamma 2^{-N} + 6\gamma \sum_{j=1}^{N} 2^{-j} \sqrt{\frac{2Cd(\frac{\gamma 2^{-j}}{4}) \log_2\left(\frac{4M_{\mathcal{G}} e m 2^j}{d(\gamma 2^{-j-2})\gamma}\right) \ln\left(\frac{16M_{\mathcal{G}}^2 m 2^{2j}}{\gamma^2}\right)}{m}}.$$

Using the fact that $d(\gamma 2^{-j-2}) \geqslant 1$ for all $j \in [\![1, N]\!]$ and that $\sqrt{m} < em$ gives

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \gamma 2^{-N} + 12\gamma \sum_{j=1}^{N} 2^{-j} \log_2\left(\frac{4M_{\mathcal{G}} e m 2^j}{\gamma}\right) \sqrt{\frac{\ln(2) Cd(\frac{\gamma 2^{-j}}{4})}{m}}.$$

Next, under the polynomial growth assumption, Assumption 1,

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \gamma 2^{-N} + 3\gamma^{1-\frac{d_{\mathcal{G}}}{2}} 2^{d_{\mathcal{G}}+2} \sqrt{\frac{CK_{\mathcal{G}}}{m}} \sum_{j=1}^{N} 2^{j\left(\frac{d_{\mathcal{G}}}{2}-1\right)} \log_2\left(\frac{4M_{\mathcal{G}} e m 2^j}{\gamma}\right). \tag{4.9}$$

Now, the way we bound the right-hand side of the above inequality is determined by the value of $d_{\mathcal{G}}$. For $d_{\mathcal{G}} \in (0,2)$, we can let $N \to \infty$ and upper bound the sum in (4.9) by the corresponding integral. For $d_{\mathcal{G}} \geqslant 2$, on the other hand, one has freedom in the choice of the value of N. It could be set in such a way so as to optimize the right-hand side of (4.9) with respect to the dependencies on C, m and γ. Thus, we have the following cases.

First case: $d_{\mathcal{G}} \in (0,2)$**.** This is the only case for which the chaining bound (4.9) can be computed in the integral form:

$$\begin{aligned}\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) &\leqslant \gamma 2^{-N} + 3\gamma^{1-\frac{d_{\mathcal{G}}}{2}} 2^{d_{\mathcal{G}}+2}\sqrt{\frac{CK_{\mathcal{G}}}{m}} \\ &\qquad \times \sum_{j=1}^{N} 2^{-j\left(\frac{2-d_{\mathcal{G}}}{2}\right)} \log_2\left(\frac{4M_{\mathcal{G}}em}{\gamma\left(2^{-j\left(\frac{2-d_{\mathcal{G}}}{2}\right)}\right)^{\frac{2}{2-d_{\mathcal{G}}}}}\right) \\ &\leqslant 3\gamma^{1-\frac{d_{\mathcal{G}}}{2}} \frac{2^{d_{\mathcal{G}}+2}}{1-2^{\frac{d_{\mathcal{G}}-2}{2}}}\sqrt{\frac{CK_{\mathcal{G}}}{m}} \int_0^{2^{\frac{d_{\mathcal{G}}-2}{2}}} \log_2\left(\frac{4M_{\mathcal{G}}em}{\gamma\epsilon^{\frac{2}{2-d_{\mathcal{G}}}}}\right) d\epsilon. \end{aligned} \tag{4.10}$$

The computation of the improper integral gives

$$\begin{aligned}\int_0^{2^{\frac{d_{\mathcal{G}}-2}{2}}} \log_2\left(\frac{4M_{\mathcal{G}}em}{\gamma\epsilon^{\frac{2}{2-d_{\mathcal{G}}}}}\right) d\epsilon &= 2^{\frac{d_{\mathcal{G}}-2}{2}} \log_2\left(\frac{8M_{\mathcal{G}}em}{\gamma}\right) + \frac{2^{\frac{d_{\mathcal{G}}}{2}}}{\ln 2\,(2-d_{\mathcal{G}})} \\ &\leqslant \frac{3\cdot 2^{\frac{d_{\mathcal{G}}}{2}} \log_2\left(\frac{8M_{\mathcal{G}}em}{\gamma}\right)}{2-d_{\mathcal{G}}}.\end{aligned}$$

Plugging this into (4.10) gives

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \frac{2^{\frac{3d_{\mathcal{G}}}{2}+6}\gamma^{1-\frac{d_{\mathcal{G}}}{2}}}{(2-d_{\mathcal{G}})\left(1-2^{\frac{d_{\mathcal{G}}-2}{2}}\right)} \log_2\left(\frac{8M_{\mathcal{G}}em}{\gamma}\right)\sqrt{\frac{CK_{\mathcal{G}}}{m}}.$$

Second case: $d_{\mathcal{G}} \geqslant 2$**.** In this case, we can bound the right-hand side of (4.9) as:

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \gamma 2^{-N} + 3\gamma^{1-\frac{d_{\mathcal{G}}}{2}} 2^{d_{\mathcal{G}}+2}\sqrt{\frac{CK_{\mathcal{G}}}{m}} \log_2\left(\frac{4M_{\mathcal{G}}em2^N}{\gamma}\right) \sum_{j\in\mathcal{J}} 2^{j\left(\frac{d_{\mathcal{G}}}{2}-1\right)}$$

We set $N = \left\lceil \log_2 m^{\frac{1}{d_{\mathcal{G}}}} \right\rceil$. Then,

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \frac{\gamma}{m^{\frac{1}{d_{\mathcal{G}}}}} + 3\cdot 2^{2+d_{\mathcal{G}}}\gamma^{1-\frac{d_{\mathcal{G}}}{2}} \log_2\left(\frac{8M_{\mathcal{G}}em^{1+\frac{1}{d_{\mathcal{G}}}}}{\gamma}\right)\sqrt{\frac{CK_{\mathcal{G}}}{m}} \sum_{j=1}^{N} 2^{j\left(\frac{d_{\mathcal{G}}}{2}-1\right)}. \tag{4.11}$$

If $d_{\mathcal{G}} = 2$, then:

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \frac{\gamma}{\sqrt{m}} + 48\log_2^2\left(\frac{8M_{\mathcal{G}}em^{\frac{3}{2}}}{\gamma}\right)\sqrt{\frac{CK_{\mathcal{G}}}{m}}.$$

Otherwise, we can bound the geometric sum in (4.11) by

$$\sum_{j=1}^{N} 2^{j\left(\frac{d_{\mathcal{G}}}{2}-1\right)} = \frac{2^{\left(\frac{d_{\mathcal{G}}}{2}-1\right)(N+1)} - 2^{\left(\frac{d_{\mathcal{G}}}{2}-1\right)}}{2^{\left(\frac{d_{\mathcal{G}}}{2}-1\right)}-1} < \frac{2^{\left(\frac{d_{\mathcal{G}}}{2}-1\right)(N+1)}}{2^{\left(\frac{d_{\mathcal{G}}}{2}-1\right)}-1},$$

which finally gives

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \frac{\gamma}{m^{\frac{1}{d_{\mathcal{G}}}}} + \frac{3\cdot 2^{2d_{\mathcal{G}}}\gamma^{1-\frac{d_{\mathcal{G}}}{2}}}{2^{\left(\frac{d_{\mathcal{G}}}{2}-1\right)}-1}\log_2\left(\frac{8M_{\mathcal{G}}em^{1+\frac{1}{d_{\mathcal{G}}}}}{\gamma}\right)\frac{\sqrt{CK_{\mathcal{G}}}}{m^{\frac{1}{d_{\mathcal{G}}}}}.$$

To complete the proof, in all cases we take expectation with respect to the sample. ■

While having the matching dependency on the margin parameter γ with that of Theorem 12, Theorem 13 improves the dependency on C to a radical one irrespective of the capacity of the function class. This, however, is achieved at the expense of the extra logarithmic factor involving the sample size. Below we demonstrate that specifying a classifier improves upon this sample-size dependency.

4.1.2 Linear Classifiers

We switch from the general case to a specific one involving linear classifiers. To the best of our knowledge, the only metric entropy bounds for linear classifiers have been derived by Bartlett [7], Williamson et al. [91] and Zhang [92]. At the basis of the L_2-norm bound of Bartlett, Lemma 22 in [7], lies Maurey's lemma (Lemma 33 in Appendix A), and Theorem 3 of Zhang [92] generalizes his result. The L_∞-norm bound of Zhang, Theorem 4 in [92], on the other hand, relies on the mistake bound due to Grove et al. [33], and it is comparable to that of Williamson and coauthors obtained based on the operatory theory methods. We first demonstrate the extension of this bound to balls of a RKHS. To this end, we appeal to a special case of the mistake bound of Grove et al.: the well known perceptron's convergence theorem [72, 62] extended to a Hilbert space. Then, we consider a particular RKHS: the Gaussian RKHS. Finally, the chaining bounds are developed based on these metric entropy bounds and compared.

4.1.2.1 Unspecified Kernel

We extend Theorem 4 of Zhang [92] to an RKHS. This relies on the convergence result of the perceptron algorithm applied to the data living in a Hilbert space which is the direct sum of the RKHS $\mathcal{H}_\kappa$ and $\mathbb{R}$ (Proposition 4 in Appendix D).

Lemma 19 *Let $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a reproducing kernel and let $(\mathcal{H}_\kappa, \|\cdot\|_{\mathcal{H}_\kappa})$ be the corresponding RKHS. Let $\Lambda_{\mathcal{X}}, \Lambda \in \mathbb{R}_+$. Suppose that $\sup_{x\in\mathcal{X}}\|\kappa_x\|_{\mathcal{H}_\kappa} \leqslant \Lambda_{\mathcal{X}}$. Let*

$$B_\Lambda(\mathcal{H}_\kappa) = \left\{h \in \mathcal{H}_\kappa : \ \|h\|_{\mathcal{H}_\kappa} \leqslant \Lambda\right\}.$$

Then, for any $\epsilon > 0$,

$$\log_2 \mathcal{N}_\infty^{ext}(\epsilon, B_\Lambda(\mathcal{H}_\kappa), n) \leqslant \frac{26\Lambda^2\Lambda_{\mathcal{X}}^2}{\epsilon^2} \log_2\left(2n\left\lceil\frac{4\Lambda_{\mathcal{X}}\Lambda}{\epsilon} + 2\right\rceil + 1\right).$$

Proof Let $\mathbb{H} = \mathcal{H}_\kappa \oplus \mathbb{R}$ be the *Hilbert space direct sum* of $\mathcal{H}_\kappa$ and $\mathbb{R}$ with an inner product $\langle\cdot,\cdot\rangle_{\mathbb{H}}$ defined as

$$\forall (f_1, f_2) \in \mathcal{H}_\kappa^2, \ \forall (r_1, r_2) \in \mathbb{R}^2, \quad \langle (f_1, r_1), (f_2, r_2)\rangle_{\mathbb{H}} = \langle f_1, f_2\rangle_{\mathcal{H}_\kappa} + r_1 r_2.$$

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

$$B_\Lambda(\mathcal{H}_\kappa)|_{\mathbf{x}_n} = \{(f(x_i))_{1\leqslant i\leqslant n} = (\langle f, \kappa_{x_i}\rangle)_{1\leqslant i\leqslant n} : f \in B_\Lambda(\mathcal{H}_\kappa)\} \subseteq [-\Lambda_{\mathcal{X}}\Lambda, \Lambda_{\mathcal{X}}\Lambda]^n.$$

Divide the interval $\left[-\Lambda_{\mathcal{X}}\Lambda - \frac{\epsilon}{2}, \Lambda_{\mathcal{X}}\Lambda + \frac{\epsilon}{2}\right]$ into $l = \left\lceil\frac{2}{\epsilon}(2\Lambda_{\mathcal{X}}\Lambda + \epsilon)\right\rceil$ sub-intervals, each of length no greater than $\epsilon/2$. Let $(\theta_j)_{0\leqslant j\leqslant l}$ be the boundaries of these sub-intervals with $\theta_{j+1} - \theta_j \leqslant \epsilon/2$, $0 \leqslant j \leqslant l-1$. Fix $f \in B_\Lambda(\mathcal{H}_\kappa)$. For any $i \in [\![1, n]\!]$, choose maximum and minimum indices, $j1(i, f) \in [\![0, l-1]\!]$ and $j2(i, f) \in [\![1, l]\!]$, respectively, so that

$$f(x_i) - \theta_{j1(i,f)} \geqslant \epsilon/2 \ \text{ and } \ -f(x_i) + \theta_{j2(i,f)} \geqslant \epsilon/2. \tag{4.12}$$

Now we want to find an ϵ-approximation of f in the $d_{\infty,\mathbf{x}_n}$ metric. Based on (4.12), a function $\bar{f} \in B_\Lambda(\mathcal{H}_\kappa)$ with

$$\forall i \in [\![1, n]\!], \quad \bar{f}(x_i) - \theta_{j1(i,f)} > 0 \ \text{ and } \ -\bar{f}(x_i) + \theta_{j2(i,f)} > 0, \tag{4.13}$$

satisfies

$$\max_{1\leqslant i\leqslant n} \left|f(x_i) - \bar{f}(x_i)\right| < \epsilon.$$

To find such a function we will use Proposition 4. First, we need to design a new dataset for the perceptron. Rewrite (4.12) as

$$\begin{gathered}\langle (f, \Lambda), (\kappa_{x_i}, \ -\theta_{j1(i,f)}/\Lambda)\rangle_{\mathbb{H}} \geqslant \epsilon/2 \\ \text{and} \\ \langle (f, \Lambda), (-\kappa_{x_i}, \ \theta_{j2(i,f)}/\Lambda)\rangle_{\mathbb{H}} \geqslant \epsilon/2.\end{gathered} \tag{4.14}$$

Let $h_i = (\kappa_{x_i}, -\theta_{j1(i,f)}/\Lambda)$ and $h_{n+i} = (-\kappa_{x_i}, \theta_{j2(i,f)}/\Lambda)$ for $1 \leqslant i \leqslant n$. Let $D = ((h_i, y_i))_{1\leqslant i\leqslant 2n} \in (\mathbb{H} \times \{-1, +1\})^{2n}$ be an input for the perceptron with all $y_i = +1$. According to (4.14), the

function $\langle w, \cdot \rangle_{\mathbb{H}}$ with $w = (f, \Lambda) \in \mathbb{H}$, yields a margin $\gamma \geqslant \frac{\epsilon}{2}$ on D. Thus, using Inequality (D.1), the number n' of updates is upper bounded as:

$$\begin{aligned} n' \leqslant \frac{\|w\|_{\mathbb{H}}^2 \max\limits_{1 \leqslant k \leqslant m} \|h_{i_k}\|_{\mathbb{H}}^2}{\gamma^2} &\leqslant \frac{4}{\epsilon^2} \left(\|f\|_{\mathcal{H}_\kappa}^2 + \Lambda^2 \right) \left(\Lambda_{\mathcal{X}}^2 + \left(\Lambda_{\mathcal{X}} + \frac{\epsilon}{2\Lambda} \right)^2 \right) \\ &\leqslant \frac{4}{\epsilon^2} \cdot 2\Lambda^2 \cdot \left(\Lambda_{\mathcal{X}}^2 + \Lambda_{\mathcal{X}}^2 + \frac{\Lambda_{\mathcal{X}} \epsilon}{\Lambda} + \frac{\epsilon^2}{4\Lambda^2} \right) \\ &\leqslant \frac{8\Lambda^2}{\epsilon^2} \cdot \left(2\Lambda_{\mathcal{X}}^2 + \frac{\Lambda_{\mathcal{X}}^2 \Lambda}{\Lambda} + \frac{\Lambda_{\mathcal{X}}^2 \Lambda^2}{4\Lambda^2} \right) \\ &\leqslant 26 \left(\frac{\Lambda \Lambda_{\mathcal{X}}}{\epsilon} \right)^2. \end{aligned} \tag{4.15}$$

The function the algorithm converges to is of the form:

$$\langle (f^*, z\Lambda), \cdot \rangle_{\mathbb{H}} = \left\langle \sum_{k=1}^{m} h_{i_k}, \cdot \right\rangle_{\mathbb{H}} = \left\langle \sum_{i=1}^{n} \alpha_i \left(\kappa_{x_i}, -\theta_{j1(i,f)}/\Lambda \right) + \sum_{i=1}^{n} \beta_i \left(-\kappa_{x_i}, \theta_{j2(i,f)}/\Lambda \right), \cdot \right\rangle_{\mathbb{H}}, \tag{4.16}$$

where $\alpha_i, \beta_i \in \mathbb{N}$ indicate the number of times h_i and h_{n+i}, respectively, appear in the updates of the perceptron with $\sum_{i=1}^{n} (\alpha_i + \beta_i) = n'$. According to Proposition 4 (in Appendix D), for all $i \in [\![1, n]\!]$,

$$\langle (f^*, z\Lambda), (\kappa_{x_i},\ -\theta_{j1(i,f)}/\Lambda) \rangle_{\mathbb{H}} > 0 \ \text{ and } \ \langle (f^*, z\Lambda), (-\kappa_{x_i},\ \theta_{j2(i,f)}/\Lambda) \rangle_{\mathbb{H}} > 0. \tag{4.17}$$

Note that this ensures that $z > 0$. Now, rewriting (4.13) as

$$\langle (\bar{f}, \Lambda), (\kappa_{x_i},\ -\theta_{j1(i,f)}/\Lambda) \rangle_{\mathbb{H}} > 0 \ \text{ and } \ \langle (\bar{f}, \Lambda), (-\kappa_{x_i},\ \theta_{j2(i,f)}/\Lambda) \rangle_{\mathbb{H}} > 0,$$

one can see that the function $\bar{f}$ can be constructed as $\bar{f} = f^*/z$. Thus, the covering number of $B_\Lambda(\mathcal{H}_\kappa)$ is no greater than the number of functions that can be expressed as (4.16) and that satisfy (4.17). Estimating this number calls for the introduction of the following notation: let

$$S_1 = \left((\kappa_{x_i}, -\theta_j/\Lambda) \right)_{1 \leqslant i \leqslant n, 0 \leqslant j \leqslant l-1} \in \mathbb{H}^{nn'},$$

and

$$S_2 = \left((-\kappa_{x_i}, \theta_j/\Lambda) \right)_{1 \leqslant i \leqslant n, 1 \leqslant j \leqslant l} \in \mathbb{H}^{nn'}.$$

Let $n_{i,j}$, $m_{i,j}$ be non-negative integers that indicate the number of times the elements of S_1 and S_2, respectively, appear in the updates of the perceptron. Denote the right-hand side of Inequality (4.15) by $M = 26 \left(\frac{\Lambda \Lambda_{\mathcal{X}}}{\epsilon} \right)^2$. Then, the covering number of $B_\Lambda(\mathcal{H}_\kappa)$ is no greater than the number s of non-negative integer solutions of

$$\sum_{i=1}^{n} \left(\sum_{j=0}^{l-1} n_{i,j} + \sum_{j=1}^{l} m_{i,j} \right) \leqslant M.$$

The number of terms on the left hand side is $|S_1| + |S_2| = 2nl$. Let $k = 2nl$. Using Lemma 42 in Appendix G, we get

$$s \leqslant 1 + \sum_{l=1}^{M} k^l \leqslant (k+1)^M.$$

Consequently, we obtain

$$\ln \mathcal{N}_\infty^{ext}\left(\epsilon, B_\Lambda\left(\mathcal{H}_\kappa\right), n\right) \leqslant M \ln\left(k+1\right).$$

Substituting the values of M and k in the right-hand side gives the desired bound. ■

Clearly, this result is an improvement over that of Alon et al., Lemma 18 applied to linear classifiers, by a factor $\ln\left(\frac{1}{\epsilon}\right)$, and this gain is propagated through the chaining bound:

Theorem 14 (After Lemma 19 in [93]) *Let $\mathcal{G} = B_\Lambda\left(H_\kappa\right)^C$. For $\gamma \in (0,1]$, let $\mathcal{F}_{\mathcal{G},\gamma}$ be the class of functions deduced from $\mathcal{G}$ according to Definition 8. Suppose $\sup_{x\in\mathcal{X}} \|\kappa_x\|_{H_\kappa} \leqslant \Lambda_{\mathcal{X}}$. Then*

$$R_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \frac{\gamma}{\sqrt{m}} + 62\Lambda\Lambda_{\mathcal{X}}\sqrt{\frac{C}{m}}\log_2\left(2\sqrt{m}\right)\sqrt{\ln\left(2m\left\lceil\frac{16\Lambda\Lambda_{\mathcal{X}}}{\gamma}\sqrt{m}+2\right\rceil+1\right)}.$$

Proof Set $h(j) = \dfrac{\gamma}{2^j}$ in the chaining formula (4.4) applied to $\mathcal{F}_{\mathcal{G},\gamma}$:

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \gamma 2^{-N} + \frac{6\gamma}{\sqrt{m}}\sum_{j=1}^{N} 2^{-j}\sqrt{\ln\left(\mathcal{N}\left(\gamma 2^{-j}, \mathcal{F}_{\mathcal{G},\gamma}, d_{2,\mathbf{z}_m}\right)\right)}.$$

Apply in sequence the norm ordering of covering numbers (1.5), the decomposition formula (4.5) and Inequality (1.2) to the metric entropy:

$$\begin{aligned}
\ln\mathcal{N}\left(\gamma 2^{-j}, \mathcal{F}_{\mathcal{G},\gamma}, d_{2,\mathbf{z}_m}\right) &\leqslant \ln\mathcal{N}\left(\gamma 2^{-j}, \mathcal{F}_{\mathcal{G},\gamma}, d_{\infty,\mathbf{z}_m}\right) \\
&\leqslant C\ln\mathcal{N}\left(\gamma 2^{-j}, \mathcal{B}_\Lambda\left(H_\kappa\right), d_{\infty,\mathbf{x}_m}\right) \\
&\leqslant C\ln\mathcal{N}^{ext}\left(\gamma 2^{-(j+1)}, \mathcal{B}_\Lambda\left(H_\kappa\right), d_{\infty,\mathbf{x}_m}\right).
\end{aligned}$$

Thus,

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \gamma 2^{-N} + 6\gamma\sqrt{\frac{C}{m}}\sum_{j=1}^{N} 2^{-j}\sqrt{\ln\mathcal{N}^{ext}\left(\gamma 2^{-(j+1)}, \mathcal{B}_\Lambda\left(H_\kappa\right), d_{\infty,\mathbf{x}_m}\right)}.$$

Applying Lemma 19 yields

$$R_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \gamma 2^{-N} + 62\Lambda\Lambda_{\mathcal{X}}\sqrt{\frac{C}{m}}\sum_{j=1}^{N}\sqrt{\ln\left(2m\left\lceil\frac{8\Lambda\Lambda_{\mathcal{X}}}{\gamma}2^j+2\right\rceil+1\right)}.$$

Finally, setting $N = \left\lceil \frac{1}{2} \log_2 (m) \right\rceil$ and upper bounding the sum in a straightforward way give the desired result. ■

Since according to Theorem 4.6 in [14], the growth rate $d_{\mathcal{G}}$ of the class $\mathcal{G}$ of Theorem 14 is 2, then we can compare this result with the corresponding case of Theorem 13, our contribution. Notice that both results provide the same dependency on C, but Theorem 14 yields a sharper dependency on the sample size: $O\left(\frac{\ln^{\frac{3}{2}}(m)}{\sqrt{m}}\right)$ instead of $O\left(\frac{\ln^2(m)}{\sqrt{m}}\right)$.

4.1.2.2 Gaussian Kernel

We now specify the kernel and focus on the most common choice: the Gaussian kernel defined as

$$\forall \sigma > 0,\ \forall x, x' \in \mathbb{R}^d, \quad \kappa_\sigma\left(x, x'\right) = \exp\left(-\sigma^{-2} \left\|x - x'\right\|_2^2\right),$$

where σ is the width of the kernel κ_σ and $\|\cdot\|_2$ denotes the Euclidean norm. Note that Lemma 19 is insensitive to the specifics of the RKHS induced by a particular kernel. The specification of the kernel calls for a dedicated approach: be it of interest in the probability or the learning theory literature, much work has been dedicated to the entropy estimates of balls in Gaussian RKHSs [31, 51, 91, 95, 38, 52, 78, 77]. Particularly, Farooq and Steinwart [28] derived the following result based on a result by van der Vaart and van Zanten [85].

Theorem 15 (After Theorem 5 in [28]) *Let H_{κ_σ} be a Gaussian RKHS over $\mathcal{X} \subset \mathbb{R}^d$. Let $B_\Lambda(H_{\kappa_\sigma})$ denote a ball of radius $\Lambda \in \mathbb{R}_+$ in H_{κ_σ}. Then, there exists a constant $K > 0$ which depends on $\mathcal{X}$, such that for all $\epsilon \in (0, \Lambda/2)$ and all $p \in (0, 1]$,*

$$\ln \mathcal{N}^{ext}\left(\epsilon, B_\Lambda\left(H_{\kappa_\sigma}\right), d_\infty\right) \leqslant \frac{K\Lambda^p}{\sigma^d \epsilon^p} \left(\frac{d+1}{ep}\right)^{d+1}, \tag{4.18}$$

where d_∞ denotes the metric induced from the supremum norm.

One can notice how the width parameter σ comes into play in the entropy bound. From this result it follows that the balls of the Gaussian RKHSs satisfy Pollard's entropy condition:

$$\begin{aligned}
\int_0^\infty \sqrt{\ln \mathcal{N}^{ext}\left(\epsilon, B_\Lambda\left(H_{\kappa_\sigma}\right), d_\infty\right)} d\epsilon &\leqslant \sqrt{\frac{K\Lambda^p}{\sigma^d}\left(\frac{d+1}{ep}\right)^{d+1}} \int_0^{\Lambda/2} \epsilon^{-\frac{p}{2}} d\epsilon \\
&= \frac{2}{2-p}\left(\frac{\Lambda}{2}\right)^{\frac{2-p}{2}} \sqrt{\frac{K\Lambda^p}{\sigma^d}\left(\frac{d+1}{ep}\right)^{d+1}} \\
&< \infty.
\end{aligned} \tag{4.19}$$

Thus, according to Theorem 2, they are universal Donsker classes, and the corresponding chaining bound is computed in a straightforward manner as demonstrated below.

Theorem 16 *Let $\mathcal{G} = B_\Lambda (H_{\kappa_\sigma})^C$. For $\gamma \in (0,1]$, let $\mathcal{F}_{\mathcal{G},\gamma}$ be the class of functions deduced from $\mathcal{G}$ according to Definition 8. There exists a constant K which depends on $\mathcal{X}$ such that for all $p \in (0,1]$,*

$$R_m (\mathcal{F}_{\mathcal{G},\gamma}) \leqslant \frac{24}{2-p} \sqrt{\frac{(2\Lambda)^p K}{\sigma^d} \left(\frac{d+1}{ep}\right)^{d+1} \left(\frac{\gamma}{2}\right)^{2-p}} \cdot \sqrt{\frac{C}{m}}.$$

Proof For the proof we use the integral form of the chaining formula (4.3) with $K = 8\sqrt{2}$ (see Bartlett's lecture notes):

$$\hat{R}_m (\mathcal{F}_{\mathcal{G},\gamma}) \leqslant 8\sqrt{2} \int_0^{\gamma/2} \sqrt{\frac{\ln \mathcal{N} (\epsilon, \mathcal{F}_{\mathcal{G},\gamma}, d_{2,\mathbf{z}_m})}{m}} d\epsilon. \tag{4.20}$$

Using the same chain of inequalities as in the proof of Theorem 14, as well as the fact that for all $f, f' \in \mathcal{F}$, $d_{\infty,\mathbf{t}_n}(f,f') \leqslant d_\infty(f,f')$, we obtain:

$$\ln \mathcal{N} (\epsilon, \mathcal{F}_{\mathcal{G},\gamma}, d_{2,\mathbf{z}_m}) \leqslant C \ln \mathcal{N}^{ext} \left(\frac{\epsilon}{2}, \mathcal{B}_\Lambda (H_{\kappa_\sigma}), d_\infty\right).$$

To conclude the proof, we apply the last inequality to the right-hand side of (4.20), then use (4.19) with $\frac{\Lambda}{2}$ replaced by $\frac{\gamma}{4}$:

$$\begin{aligned}
\hat{R}_m (\mathcal{F}_{\mathcal{G},\gamma}) &\leqslant 8\sqrt{2} \int_0^{\gamma/2} \sqrt{\frac{C \ln \left(\mathcal{N}^{ext} (\epsilon/2, \mathcal{B}_\Lambda (H_{\kappa_\sigma}), d_\infty)\right)}{m}} d\epsilon \\
&\leqslant 8\sqrt{2} \sqrt{\frac{(2\Lambda)^p C K}{m\sigma^d} \left(\frac{d+1}{ep}\right)^{d+1}} \int_0^{\gamma/2} \left(\frac{\epsilon}{2}\right)^{-p/2} d\epsilon \\
&= \frac{16\sqrt{2}}{2-p} \sqrt{\frac{(2\Lambda)^p C K}{m\sigma^d} \left(\frac{d+1}{ep}\right)^{d+1} \left(\frac{\gamma}{4}\right)^{2-p}}.
\end{aligned}$$

■

The radical dependency of Theorem 15 on the number of categories matches with that of Theorem 14, but it provides the optimal, $O\left(\frac{1}{\sqrt{m}}\right)$, convergence rate. However, the specification of the kernel brought new parameters into play: in view of the exponential dependency $\left(\frac{d}{\sigma}\right)^d$ on the dimension of the description space, the absolute amelioration provided by Theorem 16 over Theorem 14 is not clear.

4.2 L_p-norm Combinatorial Bound

In the preceding section, we saw that for general function classes using the decomposition result with $p = \infty$ and the metric entropy bounds in the L_∞-norm, one can obtain a radical dependency on C. But it leads to a slightly worse convergence rate (which is especially true for

Donsker classes) than the one in Theorem 12 which is based on the L_2-norm. In this section, we consider the values of p between these two "extreme" ones: 2 and ∞. We extend the L_2-norm combinatorial bound of Mendelson and Vershynin [61] (Lemma 8 in Chapter 2) to L_p-norms with $p \in \mathbb{N}^* \setminus \{1, 2\}$. Their bound does not depend on the sample size thanks to the use of the probabilistic extraction principle. We extend this bound in two ways: in one of them we keep the dependency on the sample size, and in the other, we remove it using the L_p-norm generalization of the aforementioned principle. This allows us to optimize the dependency on C, while not degrading the ones on m and γ, by applying one or the other combinatorial bound in the chaining based on the value of $d_{\mathcal{G}}$.

Theorem 17 *Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ to $[-M_{\mathcal{F}}, M_{\mathcal{F}}]$ with $M_{\mathcal{F}} \in [1, +\infty)$. For $\epsilon \in (0, M_{\mathcal{F}}]$, let $d(\epsilon) = \epsilon\text{-dim}(\mathcal{F})$. For all values of $p \in \mathbb{N}^* \setminus \{1, 2\}$ and $\epsilon \in (0, M_{\mathcal{F}}]$,*

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

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

(b)

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

Proof Let $\mathcal{T}_n = \{t_i : 1 \leqslant i \leqslant n\} \subset \mathcal{T}$ and $\boldsymbol{t}_n = (t_i)_{1 \leqslant i \leqslant n}$. Let $\mathcal{F}_\epsilon$ be a subset of $\mathcal{F}$ of maximal cardinality ϵ-separated with respect to the pseudo-metric $d_{p,\boldsymbol{t}_n}$. We distinguish three major steps in the proof: i) discretize the functions in the set $\mathcal{F}_\epsilon|_{\mathcal{T}_n}$, ii) demonstrate that the set of discretized functions is separated, and iii) upper bound the cardinality of the discretized set. The purpose of discretization is to reduce the original problem to the one that can be addressed by combinatorial means: we upper bound the packing number of the discretized set which is then related to that of the original set via the step (ii).

(a) Let $\epsilon' = 4(4K_p)^{1/p}$ with $K_p = \sum_{k=1}^{\infty} k^p/2^k$ (this quantity arises in Lemma 37 in Appendix E), $\eta = \dfrac{\epsilon}{\epsilon' + 2}$ and $N = \lfloor 2M_{\mathcal{F}}/\eta \rfloor$. Define the class $\tilde{\mathcal{F}}^\eta$ of functions from $\mathcal{T}_n$ to $[\![0, N]\!]$ obtained by discretizing the ones in $\mathcal{F}_\epsilon|_{\mathcal{T}_n}$ as follows:

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

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

$\lfloor b \rfloor|^p \geqslant (\max(0, |a-b|-1))^p$ for all $a, b \in \mathbb{R}_+$,

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

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

$$d_{p,\boldsymbol{t}_n}\left(\tilde{f}_1, \tilde{f}_2\right) \geqslant \frac{1}{\eta}\left(\frac{1}{n}\sum_{i\in I}|f_1(t_i)-f_2(t_i)|^p\right)^{\frac{1}{p}} - \left(\frac{|I|}{n}\right)^{\frac{1}{p}}$$
$$\geqslant \frac{1}{\eta}\left(\frac{1}{n}\sum_{i\in I}|f_1(t_i)-f_2(t_i)|^p\right)^{\frac{1}{p}} - 1. \tag{4.21}$$

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

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

It follows that

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

Applying the last inequality to (4.21) and using $((a-b)+b) \leqslant ((a-b)^{1/p}+b^{1/p})^p$ with $a, b \in \mathbb{R}_+$ and $a \geqslant b$ (where we set $a = (\epsilon'+2)^p$ and $b = 1$), we get

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

This proves our claim. Then, it follows that

$$\mathcal{M}(\epsilon, \mathcal{F}_\epsilon, d_{p,\boldsymbol{t}_n}) \leqslant \mathcal{M}(\epsilon', \tilde{\mathcal{F}}^\eta, d_{p,\boldsymbol{t}_n}) = |\tilde{\mathcal{F}}^\eta|. \tag{4.22}$$

The major step that remains to perform to obtain the claimed bound is to upper bound the right-hand side of (4.22). To this end, we will appeal to Proposition 7 in Appendix E. Let d_s be the strong dimension of $\tilde{\mathcal{F}}^\eta$. By part (1) of Lemma 3.2 in [1],

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

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

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

We can substitute this result in the upper bound on d_s based on the fact that the fat-shattering dimension is a non-increasing function of the scale:

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

Now, according to Proposition 7,

$$\begin{aligned} |\tilde{\mathcal{F}}^\eta| &\leqslant \left(\frac{eNn}{d\left(\frac{\epsilon}{15p}\right)}\right)^{2d\left(\frac{\epsilon}{15p}\right)} \\ &\leqslant \left(\frac{en}{d\left(\frac{\epsilon}{15p}\right)}\left\lfloor\frac{2M_{\mathcal{F}}}{\eta}\right\rfloor\right)^{2d\left(\frac{\epsilon}{15p}\right)} \\ &\leqslant \left(\frac{en}{d\left(\frac{\epsilon}{15p}\right)}\left(\frac{8M_{\mathcal{F}}(4K_p)^{1/p} + 4M_{\mathcal{F}}}{\epsilon}\right)\right)^{2d\left(\frac{\epsilon}{15p}\right)}. \end{aligned} \tag{4.23}$$

Applying Lemma 1 to the right-hand side of (4.23) and simplifying we get

$$|\tilde{\mathcal{F}}^\eta| \leqslant \left(\frac{15enM_{\mathcal{F}}p}{\epsilon d\left(\frac{\epsilon}{15p}\right)}\right)^{2d\left(\frac{\epsilon}{15p}\right)}. \tag{4.24}$$

We apply the relation (4.22) and Lemma 2 in sequence to the left-hand side of (4.24), and take the supremum over $\boldsymbol{t}_n \in \mathcal{T}^n$ of both sides to obtain the claimed result.

(b) To derive a dimension-free combinatorial bound we use the L_p-norm generalization of the probabilistic extraction principle: Lemma 8 of [36]. According to this lemma, there exists a subset $\mathcal{T}_q = \{t_{i_k} : 1 \leqslant k \leqslant q\}$ of $\mathcal{T}_n$ of cardinality

$$q \leqslant \frac{112\,(2M_{\mathcal{F}})^{2p}\ln\left(|\mathcal{F}_\epsilon|\right)}{3\epsilon^{2p}}, \tag{4.25}$$

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

$$|\mathcal{F}_\epsilon| = \mathcal{M}\left(\epsilon_1, \mathcal{F}_\epsilon, d_{p,\boldsymbol{t}_q}\right) = \mathcal{M}\left(\epsilon_1, \mathcal{F}_\epsilon|_{\mathcal{T}_q}, d_{p,\boldsymbol{t}_q}\right) = |\,\mathcal{F}_\epsilon|_{\mathcal{T}_q}\,|. \tag{4.26}$$

Let $\eta = \frac{\epsilon_1}{\epsilon' + 2}$. We discretize $\mathcal{F}_\epsilon|_{\mathcal{T}_q}$ in a similar way as in part (a):

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

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

$$\mathcal{M}\left(\epsilon_1, \mathcal{F}_\epsilon, d_{p,\boldsymbol{t}_q}\right) \leqslant \mathcal{M}(\epsilon', \tilde{\mathcal{F}}^\eta, d_{p,\boldsymbol{t}_q}) = |\tilde{\mathcal{F}}^\eta|. \tag{4.27}$$

By Proposition 7 in Appendix E,

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

where d_s is the strong dimension of $\tilde{\mathcal{F}}^\eta$. Plugging in the value of N and performing similar computations as in Inequalties (4.23)-(4.24) of part (a), we get

$$|\tilde{\mathcal{F}}^\eta| \leqslant \left(\frac{23eqM_\mathcal{F}(4K_p)^{1/p}}{\epsilon d_s}\right)^{2d_s}. \tag{4.28}$$

Now, we go back from the discretized set $\tilde{\mathcal{F}}^\eta$ to $\mathcal{F}_\epsilon$ using the relations (4.26) and (4.27) which yield: $|\mathcal{F}_\epsilon| \leqslant |\tilde{\mathcal{F}}^\eta|$. Using this relation and (4.25) in Inequality (4.28) gives:

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

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

$$\ln\left(|\mathcal{F}_\epsilon|\right) \leqslant 4d_s \ln\left(\frac{2576 \cdot 2^{2p+1} e M_\mathcal{F}^{2p+1} (4K_p)^{1/p}}{3\epsilon^{2p+1}}\right). \tag{4.29}$$

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

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

Substituting the last inequality in (4.29) and applying Lemma 36 to K_p, we obtain

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

The claim follows from the fact that $|\mathcal{F}_\epsilon| = \mathcal{M}(\epsilon, \mathcal{F}, d_{p,\boldsymbol{t}_n})$, the application of Lemma 1.3 and taking the supremum over $\boldsymbol{t}_n \in \mathcal{T}^n$ of both sides of the bound. ■

From the decomposition formula (4.5) one can see that, based on $C^{\frac{1}{p}} = 2^{\frac{1}{p}\log_2(C)}$, the dependency on C in the scale of covering numbers can be eliminated for all $p \geqslant \log_2(C)$. Thus, we combine Inequality (4.5) with Theorem 17 using $p = \lceil \log_2(C) \rceil$ for $C > 4$. It yields the following corollary:

Corollary 1 *Let $\mathcal{G}$ be a class of functions as in Definition 3. For $\gamma \in (0,1]$, let $\mathcal{F}_{\mathcal{G},\gamma}$ be the class of functions deduced from $\mathcal{G}$ according to Definition 8. For $\epsilon \in (0, M_{\mathcal{G}}]$, let $d(\epsilon) = \max_{1 \leqslant k \leqslant C} \epsilon\text{-}dim(\mathcal{G}_k)$. Then, for $\epsilon \in (0, \gamma]$ and $C > 4$,*

$$\ln \mathcal{N}_p(\epsilon, \mathcal{F}_{\mathcal{G},\gamma}, m) \leqslant 2Cd\left(\frac{\epsilon}{30\log_2(2C)}\right) \ln\left(\frac{30em\log_2(2C)\, M_{\mathcal{G}}}{\epsilon}\right), \tag{4.30}$$

and

$$\ln \mathcal{N}_p(\epsilon, \mathcal{F}_{\mathcal{G},\gamma}, m) \leqslant 10C\log_2(2C)\, d\left(\frac{\epsilon}{72\log_2(2C)}\right) \ln\left(\frac{14\log_2^{\frac{1}{7}}(2C)\, M_{\mathcal{G}}}{\epsilon}\right). \tag{4.31}$$

Proof Inequality (4.30) follows from the application of Inequality (4.5) and part (a) of Theorem 17 (where we drop $d(\epsilon)$ from the denominator inside the logarithm as it is greater than one), along with the fact that $C^{1/\lceil \log_2(C) \rceil} < 2$ and $\lceil \log_2(C) \rceil < \log_2(2C)$. We obtain Inequality (4.31) in a similar way using part (b) of Theorem 17. ■

4.2.1 Chaining Bound

The availability of two kinds of combinatorial bounds allows us to adapt to the phase transition in the chaining in the following manner. For $d_{\mathcal{G}} \in (0,2)$, the formula (4.4) can be upper bounded by the corresponding integral, and the use of the dimension-free bound (4.31) leads to the optimized result with respect to the number of classes without losing in m and γ. For $d_{\mathcal{G}} \geqslant 2$, such a result is obtained from the application of the bound (4.30) in (4.4). As it was explained before, the second case can also be characterized by the fact that there is a freedom in the choice of the number N of steps to construct the chaining. To optimize this construction when $d_{\mathcal{G}} > 2$, we make the non restrictive assumption that m is greater than a small power of C.

Theorem 18 (Theorem 3 in [64]) *Let $\mathcal{G}$ be a class of functions as in Definition 3. For $\gamma \in (0,1]$, let $\mathcal{F}_{\mathcal{G},\gamma}$ be the class of functions deduced from $\mathcal{G}$ according to Definition 8. Then, under*

Assumption 1, there is a constant K that depends only on $d_{\mathcal{G}}$ and $K_{\mathcal{G}}$ such that for all $C > 4$,

$$R_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant K\sqrt{\frac{C}{m}} \begin{cases} \gamma^{1-\frac{d_{\mathcal{G}}}{2}} \left(\ln(C)\right)^{\frac{d_{\mathcal{G}}}{2}+\frac{1}{2}} \ln^{\frac{1}{2}}\left(\dfrac{\ln^{\frac{1}{7}} C}{\gamma}\right), & \text{if } 0 < d_{\mathcal{G}} < 2, \\ \ln(C)\ln\left(\dfrac{m}{C}\right) \ln^{\frac{1}{2}}\left(\dfrac{m^{\frac{3}{2}}\ln C}{\gamma\sqrt{C}}\right), & \text{if } d_{\mathcal{G}} = 2, \\ \dfrac{\gamma^{1-\frac{d_{\mathcal{G}}}{2}} m^{\frac{1}{2}-\frac{1}{d_{\mathcal{G}}}}}{\left(\ln(C)\right)^{\frac{d_{\mathcal{G}}}{2}-2}} \ln^{\frac{1}{2}}\left(\dfrac{m^{1+\frac{1}{d_{\mathcal{G}}}}}{\gamma\ln(C)}\right), & \text{if } d_{\mathcal{G}} > 2 \text{ and } m \geqslant C^{1.2}. \end{cases}$$

Compared to Theorem 12, one can see that in all three cases, we have the matching dependencies on m and γ, but the dependency on C is improved: the powers of C are replaced by powers of $\ln(C)$. It is interesting to note that, in the third case, when $d_{\mathcal{G}} \geqslant 4$, which is true for instance for feedforward neural networks (see Corollary 27 in [7]), the dependency on C is slightly better than radical. This is, however, at the cost of the constant factor $d_{\mathcal{G}}^{d_{\mathcal{G}}}$.

Proof [Proof of Theorem 18] For all $j \in \mathbb{N}$, we set $h(j) = \gamma 2^{-\alpha(d_{\mathcal{G}})j}$ where $\alpha(d_{\mathcal{G}}) > 0$ for all $d_{\mathcal{G}} \in \mathcal{R}_+^*$ in (4.4) applied to $\mathcal{F}_{\mathcal{G},\gamma}$:

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \gamma 2^{-\alpha(d_{\mathcal{G}})N} + 2\sum_{j=1}^{N} \left(\gamma 2^{-\alpha(d_{\mathcal{G}})j} + \gamma 2^{-\alpha(d_{\mathcal{G}})(j-1)}\right) \sqrt{\frac{\ln \mathcal{N}\left(\gamma 2^{-\alpha(d_{\mathcal{G}})j}, \mathcal{F}_{\mathcal{G},\gamma}, d_{2,\mathbf{z}_m}\right)}{m}}, \tag{4.32}$$

First case: $d_{\mathcal{G}} \in (0,2)$. Apply Inequalities (1.5) and (4.31) in sequence to the right-hand side of (4.32) and use Assumption 1 to get

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

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

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

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

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

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

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

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

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

Applying the integration by parts formula, we obtain

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

Consequently,

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

Second case: $d_{\mathcal{G}} \geqslant 2$. In this case, we apply Inequalities (1.5) and (4.30) to (4.32) and use Assumption 1 to get

$$\begin{aligned}\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) &\leqslant \gamma 2^{-\alpha(d_{\mathcal{G}})N} + 2\sqrt{\frac{2C}{m}} \sum_{j=1}^{N} \left(\gamma 2^{-\alpha(d_{\mathcal{G}})j} + \gamma 2^{-\alpha(d_{\mathcal{G}})(j-1)}\right) \\ &\qquad \times \left[d\left(\frac{\gamma 2^{-\alpha(d_{\mathcal{G}})j}}{30\log_2(2C)}\right) \ln\left(\frac{30emM_{\mathcal{G}}\log_2\left(2C\right)}{\gamma 2^{-\alpha(d_{\mathcal{G}})j}}\right)\right]^{1/2} \\ &\leqslant \gamma 2^{-\alpha(d_{\mathcal{G}})N} + 2\sqrt{\frac{2CK_{\mathcal{G}}}{m}} \left(30\log_2(2C)\right)^{d_{\mathcal{G}}/2} \gamma^{1-\frac{d_{\mathcal{G}}}{2}} \left(1 + 2^{\alpha(d_{\mathcal{G}})}\right) \\ &\qquad \times \sum_{j=1}^{N} 2^{\alpha(d_{\mathcal{G}})\left(\frac{d_{\mathcal{G}}-2}{2}\right)j} \ln^{\frac{1}{2}}\left(\frac{30emM_{\mathcal{G}}\log_2\left(2C\right)\cdot 2^{\alpha(d_{\mathcal{G}})j}}{\gamma}\right). \end{aligned} \tag{4.33}$$

Unlike the first case, we now control the number of steps N in (4.33) through C and m. The aim is to optimize the dependencies on them while making sure that (i) N is a strictly positive integer, and (ii) as $m \to \infty$, $N \to \infty$.

Now, if $d_{\mathcal{G}} = 2$, set $\alpha(d_{\mathcal{G}}) = 1$. Thus, from (4.33), we have

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

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

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

For the final case, $d_{\mathcal{G}} > 2$, we set $\alpha(d_{\mathcal{G}}) = \dfrac{2}{d_{\mathcal{G}}-2}$ in (4.33) and bound the geometric series:

$$\begin{aligned}\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) &\leqslant \gamma 2^{-\frac{2}{d_{\mathcal{G}}-2}N} + 2\sqrt{\frac{2CK_{\mathcal{G}}}{m}} \left(30\log_2(2C)\right)^{d_{\mathcal{G}}/2} \gamma^{1-\frac{d_{\mathcal{G}}}{2}} \left(1 + 2^{\frac{2}{d_{\mathcal{G}}-2}}\right) \\ &\qquad \times \sum_{j=1}^{N} 2^{j} \ln^{\frac{1}{2}}\left(\frac{30emM_{\mathcal{G}}\log_2\left(2C\right)\cdot 2^{\frac{2}{d_{\mathcal{G}}-2}j}}{\gamma}\right) \\ &\leqslant \gamma 2^{-\frac{2}{d_{\mathcal{G}}-2}N} + 4\cdot 2^{N}\sqrt{\frac{2CK_{\mathcal{G}}}{m}} \left(30\log_2(2C)\right)^{d_{\mathcal{G}}/2} \gamma^{1-\frac{d_{\mathcal{G}}}{2}} \left(1 + 2^{\frac{2}{d_{\mathcal{G}}-2}}\right) \\ &\qquad \times \ln^{\frac{1}{2}}\left(\frac{30emM_{\mathcal{G}}\log_2\left(2C\right)\cdot 2^{\frac{2}{d_{\mathcal{G}}-2}N}}{\gamma}\right). \end{aligned} \tag{4.34}$$

Now, let $N = \left\lceil \frac{d_{\mathcal{G}} - 2}{2d_{\mathcal{G}}} \log_2 \left(\frac{m}{\log_2^{2d_{\mathcal{G}}}(2C)^{\frac{1}{d_{\mathcal{G}}}}} \right) \right\rceil$. Note that, with the assumption $m \geqslant C^{1.2}$, $m > \log_2^{2d_{\mathcal{G}}}(2C)^{\frac{1}{d_{\mathcal{G}}}}$ for all $d_{\mathcal{G}} > 2$ and thus, N is a strictly positive integer. Applying it to (4.34), we get

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

■

4.3 Conclusions

In this chapter, we related the Rademacher complexity to the metric entropy via the chaining method, and performed the decomposition at the level of the latter measure. We studied what impact different combinatorial bounds, the results that relate the metric entropy to the fat-shattering dimension, have on the dependencies of the Rademacher complexity on C, m and γ. As in [59, 36], we assumed that the fat-shattering dimensions of the component classes grow no faster than polynomially with the inverse of their scales. The combinatorial bound considered in [36] is the L_2-norm one of Mendelson and Vershynin [61]. When applied in the chaining, this gives a sublinear (but still close to a linear) dependency on C, and the dependency on the sample size matches with that in [60].

The fact that the decomposition result is optimized with respect to C in the L_∞-norm motivated us to use the combinatorial bound of Alon et al. [1] in the chaining. It led to a better (radical) dependency on C, irrespective of the growth rate of the fat-shattering dimension. However, the $O(\ln^2(m))$ dependence of this combinatorial bound propagated through the chaining and deteriorated the convergence rate in m. Yet, we demonstrated that by specifying a classifier, and thus making use of the corresponding metric entropy bound in the chaining, the dependency on the sample size could be improved. In particular, we extended the metric entropy bound of Zhang [92] for linear classes to RKHSs. Its application led to a bound in $O\left(\frac{\ln^{\frac{3}{2}}(m)}{\sqrt{m}}\right)$ compared to the $O\left(\frac{\ln^2(m)}{\sqrt{m}}\right)$ rate obtained with the bound of Alon and coauthors. Focusing on Gaussian RKHSs, we could improve it further to $O\left(\frac{1}{\sqrt{m}}\right)$ at the cost of introducing new parameters: the

dimensionality of the description space and the bandwith parameter.

For general classes, to find a good trade-off between the dependencies on C and m, we extended the L_2-norm metric entropy bound of Mendelson and Vershynin [61] to L_p-norms with $p \in \mathbb{N}^* \setminus \{1, 2\}$ in two ways: in one we kept the dependency on the sample size, and in the other, we removed it by means of the probabilistic extraction principle. The application of these bounds in the chaining gave us a radical dependency on C up to logarithmic factors without worsening those on m and γ: a uniform improvement upon the result of [36].

So far, we considered decomposing at the level of the Rademacher complexity (in Chapter 3), and the metric entropy (in this chapter), and the way these decompositions affect the dependencies on the basic parameters. The case that remains to be studied is the decomposition at the last level of scheme (4): that of the fat-shattering dimension. This is the subject of the next chapter.

Chapter 5

Decomposition of the Fat-shattering Dimension

This chapter focuses on the decomposition of the last capacity measure appearing in scheme (4), the fat-shattering dimension:

$$\text{Rademacher complexity} \xrightarrow{\text{chaining}} \text{metric entropy} \xrightarrow{\text{combinatorial bound}} \textbf{fat-shattering dimension}.$$

So far, we dealt with upper bounds on the metric entropy (via combinatorial bounds), and the Rademacher complexity (via the chaining method) in terms of the fat-shattering dimension. But the converse relationship also holds: the fat-shattering dimension can be controlled in terms of these capacity measures.

An argument made by Mendelson [59] leads to an upper bound on the fat-shattering dimension in terms of the Rademacher complexity. Thus, a tight upper bound on the Rademacher complexity implies one on the fat-shattering dimension. As discussed in Section 5.1, these results are dedicated to linear function classes.

Decompositions of the fat-shattering dimension of a composite class in terms of that of component classes via the metric entropy are due to Bartlett [7] and Duan [23]. Bartlett's result concerns a specific classifier: feedforward neural networks. For such a network with two layers, the decomposition can be viewed as that of the fat-shattering dimension of the convex hull of a bounded function class. On the other hand, Duan decomposed the fat-shattering dimension of a composite class built based on a unformly continuous function with a vector-valued domain. At the basis of this result lie the decomposition of the L_2-norm metric entropy and Talagrand's bound on the fat-shattering dimension [82]. The extension of Duan's result to the multi-category classification setting is straightforward [37]. In Section 5.2, we present a new decomposition result

for the fat-shattering dimension based on the L_∞-norm metric entropy. This gives an improved dependency on the number C of categories compared to that based on Duan's result. We also consider the matrix covering bound of Bartlett and coauthors [10]. The basic idea of this method is to collect the images of vector-valued functions into matrices and estimate the metric entropy of a set of matrices following the method of Zhang [92]. The advantage over the standard metric entropy bounds (which are based on vectors) is that it allows one to exploit the interactions between the component functions, consequently, eliminating a linear factor C from the bounds (a linear dependency on C is usually a consequence of dealing with each component function independently). Using this bound gives a decomposition result with a logarithmic dependency on C. This, however, applies only to linear classifiers.

In Section 5.3, we decompose the fat-shattering dimension of yet another family of classifiers, Lipschitz classifiers on doubling spaces studied in [32]. The derivation of this result with a linear dependency on C is rather straightforward.

Section 5.4 derives a new combinatorial bound for the margin class using the decomposition of the fat-shattering dimension for general function classes. This result is in terms of the fat-shattering dimensions of the component classes, and it can be compared to Corollary 1 of Chapter 4. The advantage of the new bound over Corollary 1 lies in the fact that it is dimension-free and that there is no dependency on C in the scales of the component fat-shattering dimensions. In Section 5.4.2, we use this result to derive the sample complexity estimate for the deviation probability (2.2) of Chapter 2 with an explicit dependency on C. We compare this result to the one obtained via the decomposition at the level of the L_∞-norm metric entropy. Both yield a linear dependency on C up to a logarithmic factor. Section 5.4.3 applies the new combinatorial bound in the chaining formula (4.4) of Chapter 3, for which we obtain a better than radical dependency on C, the result comparable to Theorem 13 of Chapter 4. Lastly, we dedicate the chaining bound to Lipschitz classifiers considered in Section 5.3. Using the decomposition result of the fat-shattering dimension of these classifiers leads to a multi-class generalization of Lemma 7 of Gottlieb and coauthors [32].

5.1 Decomposition via the Rademacher Complexity

Based on the argument made by Mendelson [59], one can deduce the following relationship between the fat-shattering dimension and the Rademacher complexity of a class $\mathcal{F}$ of real-valued functions on $\mathcal{T}$: if the worst-case empirical Rademacher complexity of $\mathcal{F}$ over $n \in \mathbb{N}^*$ points is less than some $\epsilon > 0$, then its fat-shattering dimension at scale ϵ is less than n. Thus, an efficient

bound on the Rademacher complexity implies one on the fat-shattering dimension.

Lemma 20 *Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ to $[-M_{\mathcal{F}}, M_{\mathcal{F}}]$ with $M_{\mathcal{F}} \in \mathbb{R}_+$. For $\epsilon \in (0, M_{\mathcal{F}}]$, let $d(\epsilon) = \epsilon\text{-dim}(\mathcal{F})$. For all $\epsilon \in (0, M_{\mathcal{F}}]$, if $\sup_{\mathbf{t}_n \in \mathcal{T}^n} \hat{R}_n(\mathcal{F}) \leqslant \epsilon$ for some $n \in \mathbb{N}^*$, then $d(\epsilon) \leqslant n$.*

Proof Let $S = \{t_i : 1 \leqslant i \leqslant d\} \in \mathcal{T}$ be the set of maximal cardinality d ϵ-shattered by $\mathcal{F}$. By definition of ϵ-shattering, for any $\mathbf{s}_d = (s_i)_{1 \leqslant i \leqslant d} \in \{-1, 1\}^d$, there exists $f_{\mathbf{s}_d}$ in $\mathcal{F}$ such that

$$\sum_{i=1}^{d} s_i \left(f_{\mathbf{s}_d}(t_i) - u(t_i)\right) \geqslant d\epsilon.$$

It implies

$$\forall \mathbf{s}_d \in \{-1, 1\}^d, \quad \sup_{f \in \mathcal{F}} \sum_{i=1}^{d} s_i \left(f(t_i) - u(t_i)\right) \geqslant d\epsilon.$$

Then,

$$\frac{1}{2^d} \sum_{\mathbf{s}_n \in \{-1,1\}^d} \sup_{f \in \mathcal{F}} \sum_{i=1}^{d} s_i \left(f(t_i) - u(t_i)\right) \geqslant d\epsilon,$$

which is equivalent to

$$\mathbb{E}_{\boldsymbol{\sigma}_d} \sup_{f \in \mathcal{F}} \sum_{i=1}^{d} \sigma_i \left(f(t_i) - u(t_i)\right) \geqslant d\epsilon.$$

Since Rademacher variables are centered, the above bound reduces to

$$\frac{1}{d} \mathbb{E}_{\boldsymbol{\sigma}_d} \sup_{f \in \mathcal{F}} \sum_{i=1}^{d} \sigma_i f(t_i) \geqslant \epsilon.$$

It follows that if for some $n \in \mathbb{N}^*$,

$$\frac{1}{n} \sup_{\mathbf{t}_n \in \mathcal{T}^n} \mathbb{E}_{\boldsymbol{\sigma}_n} \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i f(t_i) \leqslant \epsilon,$$

then $d \leqslant n$. ■

Remark 2 *Srebro et al. [76] upper bounded the fat-shattering dimension directly by the Rademacher complexity as follows:*

$$d(\epsilon) \leqslant \frac{4n\hat{R}_n^2(\mathcal{F})}{\epsilon^2}.$$

Given that this bound involves a factor 4 and that this Rademacher complexity is defined based on the absolute value of the sum $\sum_{i=1}^{n} \sigma_i f(t_i)$, using Lemma 20 one can get a slightly tighter bound on the fat-shattering dimension.

Our goal now is to make use of the efficient bounds on the Rademacher complexity of the margin class. As we pointed out in Chapter 3, the decomposition results of Lei et al. [55], and Maurer [57] allow one to employ interactions between the component functions, and as a consequence, for linear function classes these give bounds on the Rademacher complexity with as tight as a logarithmic dependency on C. Below, we compare these two approaches, and demonstrate the bounds they produce on the fat-shattering dimension. We start with the result of Lei et al. [55] (the proof is given in Appendix F).

Theorem 19 (After Corollary 8 [55]) *Let $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a reproducing kernel and let $(\mathcal{H}_\kappa, \|\cdot\|_{\mathcal{H}_\kappa})$ be the corresponding RKHS. Let $\Lambda_{\mathcal{X}}, \Lambda \in \mathbb{R}_+$. Suppose that $\sup_{x\in\mathcal{X}}\|\kappa_x\|_{\mathcal{H}_\kappa} \leqslant \Lambda_{\mathcal{X}}$. Let $p \leqslant 2$. For any $h = (h_k)_{1\leqslant k\leqslant C} \in \mathcal{H}_\kappa^C$, let $\|h\|_{\mathcal{H}_\kappa,p} = \left(\sum_{k=1}^C \|h_k\|_{\mathcal{H}_\kappa}^p\right)^{\frac{1}{p}}$ and let*

$$\mathcal{B}_{p,\Lambda} = \left\{h \in \mathcal{H}_\kappa^C : \|h\|_{\mathcal{H}_\kappa,p} \leqslant \Lambda\right\}.$$

For any $\gamma \in (0,1]$, let $\mathcal{F}_{\mathcal{B}_{p,\Lambda},\gamma}$ be the margin class built from $\mathcal{B}_{p,\Lambda}$ according to Definition 8. Then,

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{B}_{p,\Lambda},\gamma}\right) < \begin{cases} 4\sqrt{e\pi}\Lambda\Lambda_{\mathcal{X}}\dfrac{\ln C}{\sqrt{m}}, & \text{if } p = 1; \\ 4\sqrt{\pi}\Lambda\Lambda_{\mathcal{X}}\sqrt{\dfrac{C}{m}}, & \text{if } p = 2. \end{cases} \tag{5.1}$$

Thanks to Lemma 20, we can take benefit from this result to bound the fat-shattering dimension of $\mathcal{F}_{\mathcal{B}_{p,\Lambda},\gamma}$ as follows.

Corollary 2 *Let $\mathcal{F}_{\mathcal{B}_{p,\Lambda},\gamma}$ be as in Theorem 19. Fix $\epsilon \in (0,\gamma]$ and let $d(\epsilon) = \epsilon\text{-dim}\left(\mathcal{F}_{\mathcal{B}_{p,\Lambda},\gamma}\right)$. Then, for $n \in \mathbb{N}^*$ satisfying $\hat{R}_n\left(\mathcal{F}_{\mathcal{B}_{p,\Lambda},\gamma}\right) \leqslant \epsilon$,*

$$d(\epsilon) \leqslant \begin{cases} \dfrac{16e\pi\Lambda^2\Lambda_{\mathcal{X}}^2\ln^2 C}{\epsilon^2}, & \text{if } p = 1; \\ \dfrac{16\pi\Lambda^2\Lambda_{\mathcal{X}}^2 C}{\epsilon^2}, & \text{if } p = 2. \end{cases}$$

Proof It suffices to set the right-hand side of (5.1) less than ϵ and solve for n. ■

With Maurer's approach, Lemma 14 in Chapter 3, the following bound holds.

Theorem 20 *Let $\mathcal{F}_{\mathcal{B}_{p,\Lambda},\gamma}$ be as in Theorem 19. Then,*

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{B}_{p,\Lambda},\gamma}\right) < \begin{cases} 2\sqrt{e}\Lambda\Lambda_{\mathcal{X}}\sqrt{\dfrac{\ln C}{m}}, & \text{if } p = 1; \\ 2\Lambda\Lambda_{\mathcal{X}}\sqrt{\dfrac{C}{m}}, & \text{if } p = 2. \end{cases}$$

The proof (given in Appendix F) follows that of Theorem 19, with all orthogaussian sequences replaced by Rademacher ones. Switching from the Gaussian complexity to the Rademacher one results in the following improvement. Compared to Theorem 19, for $p = 1$, the gain is by a factor $2\sqrt{\pi \ln C}$ and for $p = 2$, the gain is only in terms of a constant: $2\sqrt{\pi}$. This has the following consequence on the growth of the fat-shattering dimension:

Corollary 3 *Let $\mathcal{F}_{\mathcal{B}_{p,\Lambda},\gamma}$ be as in Theorem 19. Fix $\epsilon \in (0, \gamma]$ and let $d(\epsilon) = \epsilon\text{-}dim\left(\mathcal{F}_{\mathcal{B}_{p,\Lambda},\gamma}\right)$. Then, for $n \in \mathbb{N}^*$ satisfying $\hat{R}_n\left(\mathcal{F}_{\mathcal{B}_{p,\Lambda},\gamma}\right) \leqslant \epsilon$,*

$$d(\epsilon) \leqslant \begin{cases} \dfrac{4e\Lambda^2\Lambda_{\mathcal{X}}^2 \ln C}{\epsilon^2}, & \text{if } p = 1; \\ \dfrac{4\Lambda^2\Lambda_{\mathcal{X}}^2 C}{\epsilon^2}, & \text{if } p = 2. \end{cases}$$

Remark 3 *The bound of Corollary 3 can be compared to that on the margin Natarajan dimension of the multi-class support vector machines, i.e., the one established in Lemma 10 in [37]:*

$$d_N(\epsilon) \leqslant \frac{\Lambda^2\Lambda_{\mathcal{X}}^2 C}{\epsilon^2}.$$

Given that the fat-shattering dimension of the margin class dominates the margin Natarajan dimension (Inequality (1.7)), when $p = 2$, there is a very small gap between these two bounds: a factor of 4. Clearly, the gain provided in Corollary 3 with respect to the dependency on C corresponds to the case $p = 1$ for which it is $O(\ln C)$.

In the following, we show another approach to decompose the fat-shattering dimension which is based on the metric entropy.

5.2 Decomposition via the Metric Entropy

The decomposition of the fat-shattering dimension of a composite class in terms of that of component classes via the metric entropy are due to Bartlett [7] and Duan [23]. Bartlett's result concerns a specific classifier: feedforward neural networks. For such a network with two layers, his decomposition can be viewed as that of the fat-shattering dimension of the absolutely convex hull of a bounded function class:

Theorem 21 (After Theorem 17 in [7]) *Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ to $[-M_{\mathcal{F}}, M_{\mathcal{F}}]$ with $M_{\mathcal{F}} \in \mathbb{R}_+$. Let*

$$absconv(\mathcal{F}) = \left\{ \sum_{i=1}^{N} \lambda_i f_i : N \in \mathbb{N}^*, f_i \in \mathcal{F}, \lambda_i \in \mathbb{R}, \sum_{i=1}^{N} |\lambda_i| \leqslant 1 \right\}.$$

For $\epsilon \in (0, M_{\mathcal{F}}]$, let $d(\epsilon) = \epsilon\text{-}dim(\mathcal{F})$. Then,

$$\epsilon\text{-}dim(absconv(\mathcal{F})) \leqslant \frac{K M_{\mathcal{F}} d\left(\frac{\epsilon}{32}\right)}{\epsilon^2} \ln^2 \left(\frac{M_{\mathcal{F}} d\left(\frac{\epsilon}{32}\right)}{\epsilon} \right),$$

where K is a universal constant.

On the other hand, Duan's result, Theorem 6.2 in [23], concerns the fat-shattering dimension of a function class obtained on the basis of a uniformly continuous function with a vector-valued domain. In the margin multi-category classification setting, his result takes the following form:

Lemma 21 (Lemma 6 in [37]) *Let $\mathcal{G}$ be as in Definition 3 and $\mathcal{F}_{\mathcal{G}}$ be as in Definition 4. Let $d(\epsilon) = \epsilon\text{-}dim(\mathcal{F}_{\mathcal{G}})$ and let $d_{\mathcal{G}_*}(\epsilon) = \max_{1 \leqslant k \leqslant C} \epsilon\text{-}dim(\mathcal{G}_k)$. Then, for all $\epsilon \in (0, M_{\mathcal{G}}]$,*

$$d(\epsilon) \leqslant 462 C d_{\mathcal{G}_*} \left(\frac{\epsilon}{96\sqrt{C}} \right) \ln \left(\frac{24 M_{\mathcal{G}} \sqrt{C}}{\epsilon} \right). \tag{5.2}$$

At the basis of this result lies the following chain of bounds: i) an upper bound on the fat-shattering dimension by the L_2-norm metric entropy due to Talagrand [82] (Proposition 1 below), ii) the decomposition of the L_2-norm metric entropy, Lemma 17, and finally iii) the L_2-norm metric entropy bound of Mendelson and Vershynin [61]. Notice how the number of categories appears in the scale of the fat-shattering dimension—this is due to the instantiation of Lemma 17 in the L_2-norm. Now, for classes with polynomially growing fat-shattering dimensions, i.e., for the ones satisfying Assumption 1, the bound (5.2) exhibits a superlinear dependency on C:

$$d(\epsilon) \leqslant 462 C K_{\mathcal{G}} \left(\frac{96\sqrt{C}}{\epsilon} \right)^{d_{\mathcal{G}}} \ln \left(\frac{24 M_{\mathcal{G}} \sqrt{C}}{\epsilon} \right). \tag{5.3}$$

For instance, for support vector machines, for which $d_{\mathcal{G}} = 2$, this result scales as a $O\left(C^2 \ln\left(\sqrt{C}\right)\right)$.

In the following subsections, we demonstrate two approaches with at most a linear (up to a logarithmic factor) dependency on C. The first one makes use of the fact that Lemma 17 is optimized with respect to C when instantiated in the L_∞-norm, i.e., C disappears from the scales of covering numbers of the component classes $\mathcal{G}_k$. We also take benefit from the fact that there is a straightforward relationship between the fat-shattering dimension and the L_∞-norm metric entropy, as Lemma 22 below demonstrates. The second approach uses the matrix covering bound developed by Bartlett and coauthors [10] which employs a coupling assumption allowing one to get rid of a linear factor C. This approach, however, is dedicated to linear function classes. We start with the first one.

5.2.1 L_∞-norm Metric Entropy

Bartlett et al. [11] upper bounded the fat-shattering dimension in terms of the L_1-norm metric entropy. Theorem 28 in Appendix A provides an optimized version of this bound under the assumption that the fat-shattering dimension at scale 2ϵ is strictly greater than 4. Talagrand's result [82], Proposition 1 in Section 5.2.2, concerns the L_2-norm metric entropy. Rudelson and Vershynin [73], on the other hand, established a deeper connection, namely, the equivalence of the fat-shattering dimension and the L_2-norm metric entropy under minimal regularity assumptions. If these results follow from non-trivial arguments, below we demonstrate that there is a rather straightforward connection between the fat-shattering dimension and the L_∞-norm metric entropy.

Lemma 22 *Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ to $[-M_\mathcal{F}, M_\mathcal{F}]$ with $M_\mathcal{F} \in \mathbb{R}_+$ and let $d(\epsilon) = \epsilon\text{-}dim\,(\mathcal{F})$. Then, for all $\epsilon \in (0, M_\mathcal{F}]$,*

$$d(\epsilon) \leqslant \log_2 \mathcal{N}_\infty\left(\epsilon, \mathcal{F}, d\right).$$

Proof Let $\mathcal{T}_d \subset \mathcal{T}$ be the set of maximal cardinality d ϵ-shattered by $\mathcal{F}$. Then, there is a subset $\mathcal{F}' \subseteq \mathcal{F}$ of cardinality 2^d such that

$$\forall f, f' \in \mathcal{F}',\ \exists t \in \mathcal{T}_d, \quad \left|f(t) - f'(t)\right| \geqslant 2\epsilon.$$

This implies that $\max_{t\in\mathcal{T}_d} |f(t) - f'(t)| \geqslant 2\epsilon$. Consequently,

$$2^d \leqslant \mathcal{M}\left(2\epsilon, \mathcal{F}, d_{\infty,\mathbf{t}_d}\right) \leqslant \mathcal{M}_\infty\left(2\epsilon, \mathcal{F}, d\right),$$

and the claimed bound follows from

$$\mathcal{M}_\infty\left(2\epsilon, \mathcal{F}, d\right) \leqslant \mathcal{N}_\infty\left(\epsilon, \mathcal{F}, d\right).$$

■

Combining this result, the decomposition of the L_∞-norm metric entropy and the L_∞-norm combinatorial bound, Lemma 18 in Chapter 4, give us the following bound.

Theorem 22 *Let $\mathcal{G}$ and $\mathcal{F}_\mathcal{G}$ be as in Definitions 3 and 8, respectively. For all $\epsilon \in (0, M_\mathcal{G}]$, let $d(\epsilon) = \epsilon\text{-}dim\,(\mathcal{F}_\mathcal{G})$ and let $d_{\mathcal{G}_*}(\epsilon) = \max_{1\leqslant k\leqslant C} \epsilon\text{-}dim(\mathcal{G}_k)$. Then, for all $\epsilon \in (0, M_\mathcal{G}]$,*

$$d(\epsilon) \leqslant 32C d_{\mathcal{G}_*}\left(\frac{\epsilon}{4}\right) \ln^2\left(\frac{320CM_\mathcal{G}^2}{\epsilon^2} d_{\mathcal{G}_*}\left(\frac{\epsilon}{4}\right)\right). \tag{5.4}$$

Proof Let $d = d(\epsilon)$. Based on Lemmas 22 and 17, we have

$$d \leqslant \sum_{k=1}^{C} \log_2 \mathcal{N}_\infty \left(\epsilon, \mathcal{G}_k, d\right). \tag{5.5}$$

Next, we bound the right-hand side of Inequality (5.5) using Lemma 18:

$$\begin{aligned} d &\leqslant C d_{\mathcal{G}_*}\left(\frac{\epsilon}{4}\right) \log_2\left(\frac{4M_{\mathcal{G}} e d}{d_{\mathcal{G}_*}\left(\frac{\epsilon}{4}\right)\epsilon}\right) \log_2\left(\frac{20M_{\mathcal{G}}^2 d}{\epsilon^2}\right) \\ &\leqslant C d_{\mathcal{G}_*}\left(\frac{\epsilon}{4}\right) \log_2^2\left(\frac{20M_{\mathcal{G}}^2 d}{\epsilon^2}\right), \end{aligned}$$

where the last inequality is due to the fact that $d_{\mathcal{G}_*}\left(\frac{\epsilon}{4}\right) \geqslant 1$. Applying Lemma 41 in Appendix G to the right-hand side of the last inequality yields

$$d \leqslant \frac{d}{2} + 16C d_{\mathcal{G}_*}\left(\frac{\epsilon}{4}\right) \ln^2\left(\frac{320CM_{\mathcal{G}}^2}{\epsilon^2} d_{\mathcal{G}_*}\left(\frac{\epsilon}{4}\right)\right)$$

and the result follows. ■

Notice that compared to Inequality (5.3), the dependency on C in Theorem 22 remains unaffected under Assumption 1. On the other hand, the maximum component fat-shattering dimension appears inside a squared logarithmic term, similarly to Theorem 21.

5.2.1.1 Application to Linear Classes

The following lemma demonstrates that for linear function classes, using Lemma 19 in Inequality (5.5), instead of instantiating the fat-shattering dimension in the general formula (5.4), allows one to obtain a bound with a $O\left(C \ln C\right)$ dependence. It is an improvement over Theorem 5.4 by a factor $\ln C$.

Lemma 23 *Let $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a reproducing kernel and let $(\mathcal{H}_\kappa, \|\cdot\|_{\mathcal{H}_\kappa})$ be the corresponding RKHS. Let $\Lambda_{\mathcal{X}}, \Lambda \in \mathbb{R}_+$. Suppose that $\sup_{x\in\mathcal{X}} \|\kappa_x\| \leqslant \Lambda_{\mathcal{X}}$. Let*

$$\mathcal{B}_\Lambda = \left\{h \in \mathcal{H}_\kappa : \ \|h\|_{\mathcal{H}_\kappa} \leqslant \Lambda\right\},$$

and let $\mathcal{G} = \mathcal{B}_\Lambda^C$. Let $\mathcal{F}_{\mathcal{G},\gamma}$ be the class built from $\mathcal{G}$ as in Definition 8. Denote $d(\epsilon) = \epsilon\text{-}dim\left(\mathcal{F}_{\mathcal{G},\gamma}\right)$. Then, for any $\epsilon \in (0, \gamma]$,

$$d(\epsilon) \leqslant \frac{104C\Lambda^2\Lambda_{\mathcal{X}}^2}{\epsilon^2} \log_2\left(\frac{38C\Lambda^2\Lambda_{\mathcal{X}}^2}{\epsilon^2}\right). \tag{5.6}$$

Proof Let $d = d(\epsilon)$. Applying Lemma 19 in Chapter 4 to (5.5) we get:

$$\begin{aligned} d &\leqslant \frac{26C\Lambda^2\Lambda_{\mathcal{X}}^2}{\epsilon^2} \log_2 \left(2d \left\lceil \frac{4\Lambda_{\mathcal{X}}\Lambda}{\epsilon} + 2 \right\rceil + 1 \right) \\ &\leqslant \frac{26C\Lambda^2\Lambda_{\mathcal{X}}^2}{\epsilon^2} \left(\log_2 d + \log_2 \left(3 \left\lceil \frac{4\Lambda_{\mathcal{X}}\Lambda}{\epsilon} + 2 \right\rceil \right) \right). \end{aligned}$$

Next, we apply Lemma 40 in Appendix G to the first term on the right-hand side. To this end, we set $\alpha = \dfrac{\epsilon^2}{52C\Lambda^2\Lambda_{\mathcal{X}}^2}$. Then

$$\frac{26C\Lambda^2\Lambda_{\mathcal{X}}^2 \log_2 d}{\epsilon^2} \leqslant \frac{d}{2} + \frac{26C\Lambda^2\Lambda_{\mathcal{X}}^2}{\epsilon^2} \log_2 \left(\frac{52C\Lambda^2\Lambda_{\mathcal{X}}^2}{2 \ln 2 \epsilon^2} \right),$$

and thus,

$$d \leqslant \frac{d}{2} + \frac{26C\Lambda^2\Lambda_{\mathcal{X}}^2}{\epsilon^2} \left(\log_2 \left(\frac{38C\Lambda^2\Lambda_{\mathcal{X}}^2}{\epsilon^2} \right) + \log_2 \left(3 \left\lceil \frac{4\Lambda_{\mathcal{X}}\Lambda}{\epsilon} + 2 \right\rceil \right) \right).$$

Solving for d and simplifying the bound yield the claimed result. ■

Since no coupling between the component functions is employed, Lemma 23 is inferioir to the one obtained based on Maurer's decomposition in Section 5.1. In the following, we consider yet another approach that allows one to exploit the interactions between the component functions and admits a $O(\ln C)$ dependence.

5.2.2 Matrix Covering Bound for Linear Classes

The matrix covering bound of Bartlett et al., Lemma 3.2 in [10], is based on Theorem 3 of Zhang [92] dealing with the L_2-norm metric entropy of linear classifers. Dealing with matrices instead of vectors allows one to take benefit from the coupling assumption between the component functions. As a consequence, applying this approach to the class $\mathcal{G}$ from Definition 3 leads to the metric entropy bound with a $O(\ln C)$ dependence: an improvement over Lemma 17 of Zhang [93] with a linear dependency on C. This implies the same dependency for the metric entropy of $\mathcal{F}_{\mathcal{G}}$, since it can be controlled by that of $\mathcal{G}$ as demonstrated below.

Lemma 24 *Let $\mathcal{G}$ be as in Definition 3 and let $\mathcal{F}_{\mathcal{G}}$ be as in Definition 4. For any $g, g' \in \mathcal{G}$, we define*

$$\forall \mathbf{x}_n = (x_i)_{1 \leqslant i \leqslant n} \in \mathcal{X}^n, \quad d_{p,p,\mathbf{x}_n}(g, g') = \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{C} \left| g_k(x_i) - g'_k(x_i) \right|^p \right)^{\frac{1}{p}}.$$

Then, for all $p \geqslant 1$,

$$\mathcal{N}(\epsilon, \mathcal{F}_{\mathcal{G}}, d_{p,\mathbf{z}_n}) \leqslant \mathcal{N}(\epsilon, \mathcal{G}, d_{p,p,\mathbf{x}_n}). \tag{5.7}$$

Proof According to Lemma 13,

$$\left|f_g(z) - f_{g'}(z)\right|^p \leqslant \sum_{k=1}^{C} \left|g_k(x) - g'_k(x)\right|^p .$$

This implies

$$\left(\frac{1}{n}\sum_{i=1}^{n} \left|f_g(z_i) - f_{g'}(z_i)\right|^p\right)^{\frac{1}{p}} \leqslant \left(\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{C} \left|g_k(x_i) - g'_k(x_i)\right|^p\right)^{\frac{1}{p}} .$$

This means that an ϵ-cover of $\mathcal{G}$ with respect to the $d_{p,p,\mathbf{x}_n}$ metric is also an ϵ-cover for $\mathcal{F}_{\mathcal{G}}$ with respect to the $d_{p,\mathbf{z}_n}$ metric. ■

For a matrix $\mathfrak{G} \in \mathbb{R}^{n\times d}$, we define the matrix norm $\|\mathfrak{G}\|_{p,q}$ as

$$\|\mathfrak{G}\|_{p,q} = \left(\sum_{i=1}^{n} \|\mathfrak{G}_{i\cdot}\|_q^p\right)^{\frac{1}{p}} = \left(\sum_{i=1}^{n}\left(\sum_{j=1}^{d} |\mathfrak{G}_{ij}|^q\right)^{\frac{p}{q}}\right)^{\frac{1}{p}} ,$$

where $\mathfrak{G}_{i\cdot}$ denotes the i-th row of $\mathfrak{G}$. The following is the matrix covering bound that we use in the decomposition of the fat-shattering dimension. It corresponds to a special case of Lemma 3.2 in [10] with the $\|\cdot\|_{1,2}$-norm on the weight matrix.

Theorem 23 *Let $X = [x_1, \ldots, x_n]^T \in \mathbb{R}^{n\times d}$ with*

$$\|X\|_{2,2} = \sqrt{\sum_{i=1}^{n} \|x_i\|_2^2} \leqslant \Lambda_{\mathcal{X}}.$$

Let

$$\mathcal{G} = \left\{ g(x) = Wx \ : \ W = [w_1, \ldots, w_C]^T \in \mathbb{R}^{d\times C},\ \|W\|_{1,2} = \sum_{k=1}^{C} \|w_k\|_2 \leqslant \Lambda \right\}$$

and

$$\mathfrak{G} = \left\{ XW = [g(x_i), \ldots, g(x_n)]^T \in \mathbb{R}^{n\times C} \ : \ g \in \mathcal{G} \right\}.$$

Let $\mathcal{F}_{\mathcal{G}}$ be a class built from $\mathcal{G}$ as in Definition 4. Then, for all $\epsilon \in (0, 2\Lambda\Lambda_{\mathcal{X}}]$,

$$\log_2 \mathcal{N}\left(\epsilon, \mathcal{F}_{\mathcal{G}}, d_{2,\mathbf{z}_n}\right) \leqslant \left\lceil \frac{\Lambda^2\Lambda_{\mathcal{X}}^2}{\epsilon^2} \right\rceil \log_2 (Cd + 1) . \tag{5.8}$$

We need the following result in the proof of Theorem 23.

Lemma 25 *Let $(x_i)_{1\leqslant i\leqslant n} \in \mathcal{X}^n$. Let the class $\mathcal{G}$ be as in Definition 3. Let*

$$\mathfrak{G} = \left\{ G \in \mathbb{R}^{n\times C} \ : \ G_{ik} = g_k(x_i), g \in \mathcal{G} \right\}.$$

Then, for all $p \geqslant 1$,

$$\mathcal{N}\left(\epsilon, \mathcal{G}, d_{p,p,\mathbf{x}_n}\right) = \mathcal{N}\left(n^{\frac{1}{p}}\epsilon, \mathfrak{G}, \|\cdot\|_{p,p}\right).$$

Proof For any $G, G' \in \mathfrak{G}$,

$$\|G - G'\|_{p,p} = \left(\sum_{i=1}^{n} \sum_{k=1}^{C} \left| G_{ik} - G'_{ik} \right|^p \right)^{\frac{1}{p}} = \left(\sum_{i=1}^{n} \sum_{k=1}^{C} \left| g_k(x_i) - g'_k(x_i) \right|^p \right)^{\frac{1}{p}}.$$

Then,

$$\|G - G'\|_{p,p} = \left(\sum_{i=1}^{n} \sum_{k=1}^{C} \left| g_k(x_i) - g'_k(x_i) \right|^p \right)^{\frac{1}{p}} \leqslant n^{\frac{1}{p}} \epsilon$$

is equivalent to

$$\left(\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{C} |g_k(x_i) - g\prime_k(x_i)|^p \right)^{\frac{1}{p}} \leqslant \epsilon.$$

■

Proof [proof of Theorem 23] We denote by $A_{j.}$ and $A_{.j}$ the j-th row and the j-th column, respectively, of the matrix A. Let X' and W' be derived from X and W, respectively, based on the following normalization:

$$X'_{.j} = \frac{\sqrt{n}\Lambda\Lambda_{\mathcal{X}}}{\|X_{.j}\|_2} X_{.j} \quad \text{and} \quad W'_{j.} = \frac{\|X_{.j}\|_2}{\sqrt{n}\Lambda\Lambda_{\mathcal{X}}} W_{j.}.$$

Then, $XW = X'W'$. Now, we represent W' as a column vector

$$\bar{w} = [W_{.1}^T, \ldots, W_{.C}^T]^T \in \mathbb{R}^{dC}.$$

Then,

$$\sum_{j=1}^{dC} |\bar{w}_j| = \frac{1}{\sqrt{n}\Lambda\Lambda_{\mathcal{X}}} \sum_{k=1}^{C} \sum_{j=1}^{d} \|X_{.j}\|_2 |W_{jk}|.$$

By Cauchy-Schwarz inequality it holds that

$$\begin{aligned} \sum_{j=1}^{dC} |\bar{w}_j| &\leqslant \frac{1}{\sqrt{n}\Lambda\Lambda_{\mathcal{X}}} \sum_{k=1}^{C} \left(\sqrt{\sum_{j=1}^{d} \|X_{.j}\|_2^2} \right) \left(\sqrt{\sum_{j=1}^{d} |W_{jk}|^2} \right) \\ &\leqslant \frac{\sqrt{n}\Lambda\Lambda_{\mathcal{X}}}{\sqrt{n}\Lambda\Lambda_{\mathcal{X}}} \leqslant 1. \end{aligned}$$

Let

$$\bar{X} = \begin{bmatrix} X' & 0 & \ldots & 0 \\ 0 & X' & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & X' \end{bmatrix} \in \mathbb{R}^{nC \times dC}.$$

We let $K = \left\lceil \frac{\Lambda^2 \Lambda_{\mathcal{X}}^2}{\epsilon^2} \right\rceil$. Then, by Maurey's lemma, Lemma 33 in Appendix G, for any

$$\bar{X}\bar{w} = \sum_{j=1}^{dC} |w_j| \left(\bar{X}_{\cdot j} \cdot \operatorname{sign}(w_j) \right),$$

one can find integers $(k_j)_{1 \leqslant j \leqslant dC}$ such that $\sum_{j=1}^{dC} |k_j| \leqslant K$ and

$$\begin{aligned} \left\| \bar{X}\bar{w} - \frac{1}{K} \sum_{j=1}^{dC} k_j \bar{X}_{\cdot j} \right\|_2^2 &\leqslant \frac{\max_{1 \leqslant j \leqslant dC} \left\| \bar{X}_{\cdot j} \right\|_2^2 \sum_{j=1}^{dC} |\bar{w}_j|}{K} \\ &\leqslant \frac{\left(\max_{1 \leqslant j \leqslant d} n\Lambda^2 \Lambda_{\mathcal{X}}^2 / \|X_{\cdot j}\|_2^2 \cdot \|X_{\cdot j}\|_2^2 \right) \cdot 1}{K} \\ &\leqslant \frac{n\Lambda^2 \Lambda_{\mathcal{X}}^2}{K} \leqslant n\epsilon^2. \end{aligned} \tag{5.9}$$

Note that

$$\left\| \bar{X}\bar{w} \right\|_2 = \|XW\|_{2,2}.$$

Thus, an ϵ-covering number of the set $A = \{\bar{X}\bar{w} \ : \ \bar{w} \in \mathbb{R}^{dC}\}$ with respect to the $\|\cdot\|_2$ norm is an ϵ-covering number of $\mathfrak{G}$ with respect to the $\|\cdot\|_{2,2}$ norm. From Inequality (5.9) it follows that the $\epsilon\sqrt{n}$-covering number of A is no greater than the number of integer solutions of $\sum_{j=1}^{dC} |k_j| \leqslant K$. According to Lemma 42 in Appendix G, this number is at most $(Cd+1)^K$. Then, we have

$$\log_2 \mathcal{N}\left(\sqrt{n}\epsilon, \mathfrak{G}, \|\cdot\|_{2,2}\right) \leqslant \left\lceil \frac{\Lambda^2 \Lambda_{\mathcal{X}}^2}{\epsilon^2} \right\rceil \log_2 (Cd+1).$$

Applying Lemma 25 and Lemma 24 in sequence to the left-hand side concludes the proof. ■

Now, the dimensionality d in Inequality (5.8) can be replaced by the size n based on Corollary 3 in [92]. This is useful when $n \leqslant d$. The idea is to consider the subspace S spanned by $\{x_i : 1 \leqslant i \leqslant n\}$ whose dimensionality is at most n. Then, we can express each w_k in Theorem 23 as a sum of projected and orthogonal to S components:

$$\forall k \in [\![1, C]\!], \quad w_k = w_k^{\parallel} + w_k^{\perp},$$

with $w_k^{\parallel} \in S$ and $w_k^{\perp} \in S^{\perp}$, and $S^{\perp}$ being the orthogonal complement of S. Clearly,

$$\forall i \in [\![1, n]\!], \forall k \in [\![1, C]\!], \quad \langle w_k, x_i \rangle = \langle w_k^{\parallel}, x_i \rangle$$

and

$$\sum_{k=1}^{C} \|w_k^{\parallel}\|_2 \leqslant \sum_{k=1}^{C} \|w_k\|_2 \leqslant \Lambda.$$

We can collect all $w_k^{\|}$ into a matrix $W^{\|} = X^T\mathcal{B}$ with $\mathcal{B} \in \mathbb{R}^{n\times C}$. Then, $XW^{\|} = XX^T\mathcal{B}$. Applying the reasoning in the proof above to X replaced now by XX^T and W replaced by $\mathcal{B}$ gives:

$$\log_2 \mathcal{N}\left(\epsilon, \mathcal{F}_{\mathcal{G}}, d_{2,\mathbf{z}_n}\right) \leqslant \left\lceil \frac{\Lambda^2\Lambda_{\mathcal{X}}^2}{\epsilon^2} \right\rceil \log_2\left(Cn+1\right). \tag{5.10}$$

We will use the following result that estimates the fat-shattering dimension in terms of the L_2-norm metric entropy. This result provides an explicit value for the constant in Proposition 1.4 of Talagrand [82].

Proposition 1 (After Proposition 5 in [37]) *Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ to $[-M_{\mathcal{F}}, M_{\mathcal{F}}]$ with $M_{\mathcal{F}} \in \mathbb{R}_+$. For any $\epsilon \in (0, M_{\mathcal{F}}]$, let $d = \epsilon$-dim$(\mathcal{F})$. Then,*

$$d \leqslant 16 \log_2 \mathcal{N}_2\left(\frac{\epsilon}{2}, \mathcal{F}, d\right).$$

Combining it with Inequality (5.10) gives:

Corollary 4 *Let $\mathcal{F}_{\mathcal{G}}$ be the class given in Theorem 23. Let $\mathcal{F}_{\mathcal{G},\gamma}$ be deduced from $\mathcal{F}_{\mathcal{G}}$ according to Definition 8. Let $d(\epsilon) = \epsilon$-dim$(\mathcal{F}_{\mathcal{G},\gamma})$. For all $\epsilon \in (0, \gamma]$,*

$$d(\epsilon) \leqslant 32 \left\lceil \frac{4\Lambda^2\Lambda_{\mathcal{X}}^2}{\epsilon^2} \right\rceil \log_2\left(3C \left\lceil \frac{4\Lambda^2\Lambda_{\mathcal{X}}^2}{\epsilon^2} \right\rceil\right).$$

Proof By the left hand-side of Inequality (4.5), Inequality (5.10), and Proposition 1,

$$d(\epsilon) \leqslant 16 \left\lceil \frac{4\Lambda^2\Lambda_{\mathcal{X}}^2}{\epsilon^2} \right\rceil \log_2\left(Cd(\epsilon)+1\right).$$

Let $K = 16 \left\lceil \frac{4\Lambda^2\Lambda_{\mathcal{X}}^2}{\epsilon^2} \right\rceil$ and apply Lemma 40 in Appendix G to the right-hand side:

$$\begin{aligned} d(\epsilon) &\leqslant K\left(\log_2(2C) + \log_2 d(\epsilon)\right) \\ &\leqslant K \log_2\left(\frac{2CK}{\ln 2}\right) + \frac{d(\epsilon)}{2}. \end{aligned}$$

Thus, $d(\epsilon) \leqslant 2K \log_2\left(\frac{2CK}{\ln 2}\right)$ and the result follows. ■

In the following section, we decompose the fat-shattering dimension of yet another family of classifiers: Lipschitz classifiers.

5.3 Multi-category Lipschitz Classifiers

This section considers Lipschitz classifiers, an example of which is the nearest neighbor [49]. We also assume that the description space is a doubling space. A metric space $(\mathcal{T}, \rho)$ is said to be

a doubling space if there exists a constant $\lambda < \infty$ such that every ball of radius r in $\mathcal{T}$ can be covered by at most λ balls of half the radius $r/2$. Then the smallest such value λ is said to be the doubling constant of $\mathcal{T}$, and the doubling dimension D of $\mathcal{T}$ is $D = \log_2 \lambda$. An example of a doubling space is a d-dimensional Euclidean space whose doubling dimension is roughly d [39].

It is a well-known fact that the fat-shattering dimension of a class of Lipschitz functions can be controlled via the packing number of their domain (see, for instance, Theorem 13 in [7]). If the domain is a doubling space, then according to the following result, the doubling dimension has a direct impact on the fat-shattering dimension.

Lemma 26 (Lemma 1 in [32]) *Let $(\mathcal{T}, \rho)$ be a metric space with doubling dimension D. Suppose $S \subset \mathcal{T}$ is finite and $\inf_{s,s' \in S} \rho(s, s') = \alpha > 0$. Let $diam(S) = \sup_{s,s' \in S} \rho(s, s')$. Then the size of S is*

$$|S| \leqslant \left(\frac{2diam(S)}{\alpha} \right)^D .$$

The upper bound on the fat-shattering dimension of Lipschitz classifiers is provided by Theorem 3 of Gottlieb et al. [32]. Our generalization of this result to the multi-category setting is straightforward. Indeed, it is obtained by using the Lipschitz property of component functions and an idea inspired from the proof of Lemma 4.2 of Bartlett and Shawe-Taylor [14].

Lemma 27 *Let $(\mathcal{X}, \rho)$ be a metric space with doubling dimension D. Suppose that $\sup_{x,x' \in \mathcal{X}} \rho(x, x') \leqslant \Lambda_{\mathcal{X}}$. Let $\mathcal{G}$ be a class of functions given in Definition 3 and $\mathcal{F}_{\mathcal{G}}$ in Definition 4. We assume that for all $k \in [\![1, C]\!]$, $\mathcal{G}_k$ is a class of L-Lipschitz functions. Let $d(\epsilon) = \epsilon\text{-}dim(\mathcal{F}_{\mathcal{G}})$. Then, for all $\epsilon \in (0, \mathcal{M}_{\mathcal{G}}]$,*

$$d(\epsilon) \leqslant C \left(\frac{L\Lambda_{\mathcal{X}}}{\epsilon} \right)^D . \tag{5.11}$$

Proof Let $S = \{(x_i, y_i) : 1 \leqslant i \leqslant n\}$ be a set of n ordered pairs ϵ-shattered by $\mathcal{F}_{\mathcal{G}}$ and let $u : \mathcal{X} \times \mathcal{Y} \to \mathcal{R}$ be a witness to this shattering. Now, partition S into subsets S_k with $k \in [\![1, C]\!]$ according to the value of y_i in each pair. Let $I_k \subseteq [\![1, n]\!]$ be a set of indices of pairs in S_k. First, assume that $|S_k| > 1$. Then, consider a partitioning of I_k into two subsets I' and I'' so that

$$\big(I' \neq \emptyset \text{ and } I'' \neq \emptyset\big) \text{ and } \forall (i, j) \in I' \times I'',\ u(x_i, k) \geqslant u(x_j, k). \tag{5.12}$$

For any such partition there exists $g \in \mathcal{G}$ satisfying

$$\begin{cases} \forall i \in I', & \frac{1}{2} \left(g_k(x_i) - \max_{q \neq k} g_q(x_i) \right) - u(x_i, k) \geqslant \gamma, \\ \forall j \in I'', & \frac{1}{2} \left(-g_k(x_j) + \max_{q \neq k} g_q(x_j) \right) + u(x_j, k) \geqslant \gamma. \end{cases}$$

Then, adding up both lines yields

$$\forall (i,j) \in I' \times I'',\ \frac{1}{2}\left(g_k(x_i) - g_k(x_j)\right) + \frac{1}{2}\left(g_l(x_j) - g_l(x_i)\right) - \left(u(x_i,k) - u(x_j,k)\right) \geqslant 2\gamma,$$

where $l = \max_{q \neq k} g_q(x_j)$. By the assumption (5.12), it follows that

$$\frac{1}{2}\left(g_k(x_i) - g_k(x_j)\right) + \frac{1}{2}\left(g_l(x_j) - g_l(x_i)\right) \geqslant 2\gamma.$$

By the Lipschitz property on the other hand,

$$L\rho\left(x_i, x_j\right) \geqslant 2\gamma.$$

By considering all partitions satisfying the assumption (5.12), one can establish that any two elements in S_k are $\dfrac{2\gamma}{L}$-separated. Therefore,

$$|I_k| = \mathcal{M}\left(\frac{2\gamma}{L}, \mathcal{S}_k, \rho\right) \leqslant \mathcal{M}\left(\frac{2\gamma}{L}, \mathcal{S}, \rho\right).$$

Now, the above relation holds trivially when S_k is empty or contains one element. Then, repeating this procedure for all $k \in [\![1, C]\!]$, it follows that

$$n = \sum_{k=1}^{C} |I_k| \leqslant C\mathcal{M}\left(\frac{2\gamma}{L}, S, \rho\right).$$

Applying Lemma 26 to the packing number yields the required result. ■

5.4 Application of the Decomposition Results

In this section, we focus on the application of the decomposition results, Theorem 22 and Lemma 27. The use of the former result leads to a new combinatorial bound for the class $\mathcal{F}_{\mathcal{G},\gamma}$ in terms of the fat-shattering dimensions of the component classes $\mathcal{G}_k$. Then, we apply the new combinatorial bound in two contexts: in Section 5.4.2, we present a new sample complexity estimate for the deviation probability, Theorem 4, with an explicit dependency on C, and in Section 5.4.3, we apply the combinatorial bound in the chaining bound. Finally, we dedicate the chaining bound to Lipschitz classifiers considered in Section 5.3, where we make use of the corresponding decomposition result of the fat-shattering dimension, Lemma 27.

5.4.1 New Combinatorial Bound

When it comes to the margin class, one can derive a sharper bound than that of Mendelson and Vershynin, Lemma 8 in Chapter 2, taking benefit from the following result involving the

margin graph dimension, and the fact that the this dimension is dominated by the fat-shattering dimension, Inequality (1.7) in Chapter 1.

Lemma 28 (After Lemma 7 in [37]) *For $\epsilon \in (0, M_{\mathcal{G}}]$, let $d_G(\epsilon) = \epsilon\text{-}G\text{-}dim(\mathcal{F}_{\mathcal{G}})$. Then,*

$$\ln \mathcal{N}_2(\epsilon, \mathcal{F}_{\mathcal{G},\gamma}, n) \leqslant 20 d_G\left(\frac{\epsilon}{48}\right) \ln\left(\frac{6\gamma}{\epsilon}\right). \tag{5.13}$$

Applying the right-hand side of Inequality (1.7) to the above bound gives:

Corollary 5 *For $\epsilon \in (0, M_{\mathcal{G}}]$, let $d(\epsilon) = \epsilon\text{-}dim(\mathcal{F}_{\mathcal{G}})$. Then, for any $\epsilon \in (0, \gamma]$,*

$$\ln \mathcal{N}_2(\epsilon, \mathcal{F}_{\mathcal{G},\gamma}, n) \leqslant 20 d\left(\frac{\epsilon}{48}\right) \ln\left(\frac{6\gamma}{\epsilon}\right). \tag{5.14}$$

Notice that, compared to Lemma 8 applied to $\mathcal{F}_{\mathcal{G},\gamma}$ the scale of the fat-shattering dimension (of $\mathcal{F}_{\mathcal{G}}$) is increased which decreases the dimension tightening the bound. Applying the decomposition of the fat-shattering dimension, Theorem 22, to Inequality (5.14) yields the following result.

Corollary 6 *For $\epsilon \in (0, M_{\mathcal{G}}]$, let $d(\epsilon) = \max_{1 \leqslant k \leqslant C} \epsilon\text{-}dim(\mathcal{G}_k)$. Then, for any $\gamma \in (0, 1]$ and for any $\epsilon \in (0, \gamma]$,*

$$\ln \mathcal{N}_2(\epsilon, \mathcal{F}_{\mathcal{G},\gamma}, n) \leqslant 640 C d\left(\frac{\epsilon}{192}\right) \ln^2\left(\frac{2^{14} C M_{\mathcal{G}}^2}{\epsilon^2} d\left(\frac{\epsilon}{192}\right)\right) \ln\left(\frac{6\gamma}{\epsilon}\right). \tag{5.15}$$

The particularity of this result lies in the fact that it is dimension-free and that there is no dependency on C in the scale of the fat-shattering dimension. In this sense, Corollary 6 represents the best of the two kinds of combinatorial bounds: the L_2-norm bound, Lemma 8, and the L_∞-norm bound, Lemma 18. On the downside, the fat-shattering dimension now appears also inside the logarithm, and we are dealing with $O\left(\ln^3\left(\frac{1}{\epsilon}\right)\right)$ as $\epsilon \to 0$.

In the following sections, we focus on several applications of this combinatorial bound, as well as Corollary 5.

5.4.2 Sample Complexity: Explicit Dependency on C

In Chapter 2, we derived several sample complexity estimates in terms of the fat-shattering dimension of the margin class $\mathcal{F}_{\mathcal{G},\gamma}$. Below we derive sample complexity estimates with explicit dependencies on the number of categories scaling as a $O(C \ln^2 C)$. Using the new combinatorial bound we get:

Theorem 24 *Fix $\epsilon, \delta \in (0,1)$ and fix $\gamma \in (0,1]$. For $\epsilon \in (0, M_{\mathcal{G}}]$, let $d(\epsilon) = \max_{1 \leqslant k \leqslant C} \epsilon\text{-}dim(\mathcal{G}_k)$. Then, for sample size*

$$m \geqslant \frac{32}{\epsilon^2}\left(640Cd\left(\frac{\epsilon\gamma}{1536}\right)\ln^2\left(\frac{2^{20}CM_{\mathcal{G}}^2}{\epsilon^2\gamma^2}d\left(\frac{\epsilon\gamma}{1536}\right)\right)\ln\left(\frac{6}{\epsilon}\right)+\ln\frac{2}{\delta}\right),$$

the probability (2.1) *is at most δ.*

Proof Apply Corollary 6 to the right hand-side of Inequality (2.2) based on the norm ordering of covering numbers, Inequality (1.5) in Chapter 1. This gives

$$\begin{aligned}P^m\left(\sup_{g\in\mathcal{G}}\left(L(g)-L_{\gamma,m}(g)\right)>\epsilon\right) &\leqslant 2\mathcal{N}_2\left(\frac{\epsilon\gamma}{8},\mathcal{F}_{\mathcal{G},\gamma},2m\right)\exp\left(-\frac{m\epsilon^2}{32}\right)\\ &\leqslant \exp\left(\ln 2+640Cd\left(\frac{\epsilon\gamma}{1536}\right)\ln^2\left(\frac{2^{20}CM_{\mathcal{G}}^2}{\epsilon^2\gamma^2}d\left(\frac{\epsilon\gamma}{1536}\right)\right)\ln\left(\frac{6}{\epsilon}\right)\right)\\ &\quad\times\exp\left(-\frac{m\epsilon^2}{32}\right).\end{aligned}$$

Fix $\delta \in (0,1)$ and let the right-hand side of the last inequality be at most δ. Solving for m yields the claimed bound. ■

This result can be compared to the one obtained via the decomposition of the L_∞-norm metric entropy, and the use of the L_∞-norm combinatorial bound, Lemma 18. Recall that the motivation to instantiate the decomposition in this norm is that the scales of the component covering numbers do not depend on C.

Theorem 25 *Fix $\epsilon, \delta \in (0,1)$ and fix $\gamma \in (0,1]$. For $\epsilon \in (0, M_{\mathcal{G}}]$, let $d(\epsilon) = \max_{1 \leqslant k \leqslant C} \epsilon\text{-}dim(\mathcal{G}_k)$. Then, for sample size*

$$m \geqslant \frac{32}{\epsilon^2}\left(9Cd\left(\frac{\epsilon\gamma}{32}\right)\ln^2\left(\frac{942CM_{\mathcal{G}}^2}{\epsilon^2\gamma^2}d\left(\frac{\epsilon\gamma}{32}\right)\right)+\ln\frac{2}{\delta}\right),$$

the probability (2.1) *is at most δ.*

Proof The right hand-side of Lemma 18 can be upper bounded as follows:

$$\ln\mathcal{N}\left(\epsilon,\mathcal{G}_k,d_{\infty,\mathbf{x}_m}\right)\leqslant 1.5d\left(\frac{\epsilon}{4}\right)\ln^2\left(\frac{20M_{\mathcal{G}}^2m}{\epsilon^2}\right).$$

Then by decomposition of covering numbers, Lemma 17, and the norm ordering of covering numbers, the right hand-side of (2.2) can be upper bounded as

$$\ln\mathcal{N}_1\left(\frac{\epsilon\gamma}{8},\mathcal{F}_{\mathcal{G}},2m\right)\leqslant 1.5Cd\left(\frac{\epsilon\gamma}{32}\right)\ln^2\left(\frac{2560M_{\mathcal{G}}^2m}{\epsilon^2\gamma^2}\right).$$

Based on the fact that $a^2 + b^2 \geqslant 2ab$ for any a and b, we have

$$\ln \mathcal{N}_1\left(\frac{\epsilon\gamma}{8}, \mathcal{F}_\mathcal{G}, 2m\right) \leqslant 3Cd\left(\frac{\epsilon\gamma}{32}\right)\left(\ln^2\left(\frac{2560M_\mathcal{G}^2}{\epsilon^2\gamma^2}\right) + \ln^2 m\right). \tag{5.16}$$

We upper bound $\ln^2 m$ as follows. Consider

$$f(m) = \ln m - \sqrt{Km},$$

with $K > 0$. By a standard computation, we find that f takes its maximum value at $m = \frac{4}{K}$. Therefore,

$$\ln m \leqslant \sqrt{Km} + \ln\left(\frac{4}{Ke^2}\right).$$

Substituting it in the right-hand side of Inequality (5.16) yields

$$\begin{aligned}\ln \mathcal{N}_1\left(\frac{\epsilon\gamma}{8}, \mathcal{F}_\mathcal{G}, 2m\right) &\leqslant 3Cd\left(\frac{\epsilon\gamma}{32}\right)\left(\ln^2\left(\frac{2560M_\mathcal{G}^2}{\epsilon^2\gamma^2}\right) + \left(\sqrt{Km} + \ln\left(\frac{4}{Ke^2}\right)\right)^2\right) \\ &\leqslant 3Cd\left(\frac{\epsilon\gamma}{32}\right)\left(\ln^2\left(\frac{2560M_\mathcal{G}^2}{\epsilon^2\gamma^2}\right) + 2Km + 2\ln^2\left(\frac{4}{Ke^2}\right)\right).\end{aligned}$$

Now, let $K = \epsilon^2 / \left(384Cd\left(\frac{\epsilon\gamma}{32}\right)\right)$. Then,

$$\begin{aligned}\ln \mathcal{N}_1\left(\frac{\epsilon\gamma}{8}, \mathcal{F}_\mathcal{G}, 2m\right) &\leqslant 3Cd\left(\frac{\epsilon\gamma}{32}\right)\ln^2\left(\frac{2560M_\mathcal{G}^2}{\epsilon^2\gamma^2}\right) + \frac{m\epsilon^2}{64} + 6Cd\left(\frac{\epsilon\gamma}{32}\right)\ln^2\left(\frac{1536Cd\left(\frac{\epsilon\gamma}{32}\right)}{\epsilon^2 e^2}\right) \\ &\leqslant \frac{m\epsilon^2}{64} + 9Cd\left(\frac{\epsilon\gamma}{32}\right)\ln^2\left(\frac{2560Cd\left(\frac{\epsilon\gamma}{32}\right)M_\mathcal{G}^2}{\epsilon^2\gamma^2 e^2}\right).\end{aligned}$$

Substituting the last inequality in the right-hand side of Inequality (2.2) gives

$$P^m\left(\sup_{g\in\mathcal{G}}\left(L\left(g\right) - L_{\gamma,m}\left(g\right)\right) > \epsilon\right) \leqslant \exp\left(\ln 2 + 9Cd\left(\frac{\epsilon\gamma}{32}\right)\ln^2\left(\frac{2560Cd\left(\frac{\epsilon\gamma}{32}\right)M_\mathcal{G}^2}{\epsilon^2\gamma^2 e^2}\right) - \frac{m\epsilon^2}{32}\right).$$

Finally, letting the right-hand side to be at most δ and solving for m conclude the proof. ■

Although Theorem 25 is obtained using the sample-size dependent combinatorial bound, compared to Theorem 24, it provides a more efficient result in terms of the constants, the scale of the component fat-shattering dimension and the dependency on ϵ.

5.4.3 Chaining Bound

This section deals with the application of the combinatorial bounds, Corollary 6, and Corollary 5 combined with Lemma 27 (the decomposition result for the fat-shattering dimension of Lipschitz classifiers) in the chaining formula. We start with the first bound concerning general function classes.

5.4.3.1 General Function Classes

For general classes of functions (excluding Donsker classes for which the sample-size dependence is worse than that of Theorem 13), using Corollary 6 in the chaining gives the following bound on the Rademacher complexity. This bounds admits sharper than a radical dependency on C by slightly deteriorating that on m, which is comparable to Theorem 13 in Chapter 4.

Theorem 26 *Let $\mathcal{G}$ be a class of functions as in Definition 3. For $\gamma \in (0,1]$, let $\mathcal{F}_{\mathcal{G},\gamma}$ be the class of functions deduced from $\mathcal{G}$ according to Definition 8. Then, under Assumption 1, there is a constant K that depends only on $d_{\mathcal{G}}$ and $K_{\mathcal{G}}$ such that,*

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

Proof To derive the bound we follow similar computations as in the proof of Theorem 18. Apply Assumption 1 in Corollary 6 and take into account the obtained bound in the chaining formula (4.32):

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant h(N) + 2\sqrt{\frac{640 \cdot 192^{d_{\mathcal{G}}} C K_{\mathcal{G}}}{m}} \times \sum_{j=1}^{N} \frac{h(j)+h(j-1)}{(h(j))^{d_{\mathcal{G}}/2}} \ln\left(\frac{2^{14} \cdot 192^{d_{\mathcal{G}}} K_{\mathcal{G}} M_{\mathcal{G}}^2 C}{(h(j))^{2+d_{\mathcal{G}}}}\right) \sqrt{\ln\left(\frac{6\gamma}{h(j)}\right)}.$$

For all $j \in \mathbb{N}$, we set $h(j) = \gamma 2^{-\alpha(d_{\mathcal{G}})j}$ with $\alpha(d_{\mathcal{G}}) > 0$ for all $d_{\mathcal{G}} \in \mathbb{R}_+$. Then,

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \gamma 2^{-\alpha(d_{\mathcal{G}})N} + 2\sqrt{\frac{640 \cdot 192^{d_{\mathcal{G}}} C K_{\mathcal{G}}}{m}} \gamma^{1-\frac{d_{\mathcal{G}}}{2}} \left(1 + 2^{\alpha(d_{\mathcal{G}})}\right) \times \sum_{j=1}^{N} 2^{\alpha(d_{\mathcal{G}})\left(\frac{d_{\mathcal{G}}-2}{2}\right)j} \ln\left(\frac{2^{14} \cdot 192^{d_{\mathcal{G}}} K_{\mathcal{G}} M_{\mathcal{G}}^2 C}{(\gamma 2^{-\alpha(d_{\mathcal{G}})j})^{2+d_{\mathcal{G}}}}\right) \sqrt{\ln\left(6 \cdot 2^{\alpha(d_{\mathcal{G}})j}\right)}.$$

Now, if $d_{\mathcal{G}} = 2$, set $\alpha(d_{\mathcal{G}}) = 1$, and thus

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \gamma 2^{-N} + 6\sqrt{\frac{640 \cdot 192^2 C K_{\mathcal{G}}}{m}} \sum_{j=1}^{N} \ln\left(\frac{2^{14} \cdot 192^2 2^{4j} K_{\mathcal{G}} M_{\mathcal{G}}^2 C}{\gamma^4}\right) \sqrt{\ln\left(6 \cdot 2^j\right)}.$$

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

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \gamma\sqrt{\frac{C}{m}} + 6\sqrt{\frac{640 \cdot 192^2 C K_{\mathcal{G}}}{m}} \left\lceil \log_2 \sqrt{\frac{m}{C}} \right\rceil \sqrt{\ln\left(12\sqrt{\frac{m}{C}}\right)} \ln\left(\frac{2^{18} \cdot 192^2 K_{\mathcal{G}} M_{\mathcal{G}}^2 m^2}{\gamma^4 C}\right).$$

For $d_{\mathcal{G}} > 2$, we set $\alpha(d_{\mathcal{G}}) = \dfrac{2}{d_{\mathcal{G}} - 2}$ and $N = \left\lceil \dfrac{d_{\mathcal{G}} - 2}{2d_{\mathcal{G}}} \log_2 \left(\dfrac{m}{\log_2^{2d_{\mathcal{G}}}(2C)^{\frac{1}{d_{\mathcal{G}}}}} \right) \right\rceil$. Consequently,

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \frac{\gamma \log_2^2(2C)^{\frac{1}{d_{\mathcal{G}}}}}{m^{\frac{1}{d_{\mathcal{G}}}}} + 4\sqrt{\frac{640 \cdot 192^{d_{\mathcal{G}}} C K_{\mathcal{G}}}{m}} \gamma^{1-\frac{d_{\mathcal{G}}}{2}} (1 + 2^{\frac{2}{d_{\mathcal{G}}-2}}) \frac{m^{\frac{d_{\mathcal{G}}-2}{2d_{\mathcal{G}}}}}{\log_2^{d_{\mathcal{G}}-2}(2C)^{\frac{1}{d_{\mathcal{G}}}}}$$
$$\times \ln^{\frac{1}{2}} \left(\frac{6 \cdot 2^{\frac{2}{d_{\mathcal{G}}-2}} d_{\mathcal{G}}^2 m^{\frac{1}{d_{\mathcal{G}}}}}{\log_2^2(2C)} \right) \ln \left(\frac{2^{14} \cdot 192^{d_{\mathcal{G}}} 2^{\frac{2(2+d_{\mathcal{G}})}{d_{\mathcal{G}}-2}} K_{\mathcal{G}} M_{\mathcal{G}}^2}{\gamma^{2+d_{\mathcal{G}}}} \cdot \frac{m^{1+\frac{2}{d_{\mathcal{G}}}}}{\log_2^{2d_{\mathcal{G}}+4}(2C)^{\frac{1}{d_{\mathcal{G}}}}} \right).$$

■

5.4.3.2 Lipschitz Classifiers

We apply the combination of the decomposition result for Lipschitz classifiers, Lemma 27, and the combinatorial bound, Corollary 5, in the chaining bound. This result corresponds to a multi-category extension of Lemma 7 of Gottlieb and co-authors [32]. In particular, we obtain a matching dependency on the sample size up to a logarithmic factor, and a radical dependency on the number of categories. On the downside, this result is useful only when the sample size is much greater than the doubling dimension of the description space.

Theorem 27 *Let $(\mathcal{X}, \rho)$ be a metric space with doubling dimension D and suppose that $\sup_{x,x' \in \mathcal{X}} \rho(x, x') \leqslant \Lambda_{\mathcal{X}}$. Let $\mathcal{G}$ be as in Definition 3, where for all $k \in [\![1, C]\!]$, $\mathcal{G}_k$ is a class of L-Lipschitz functions. Let $\mathcal{F}_{\mathcal{G},\gamma}$ be derived from $\mathcal{G}$ as in Definition 4. Then, for any $\ln m \geqslant D$,*

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \frac{\gamma}{m^{\frac{1}{D}}} + \frac{27\sqrt{C}}{\gamma^{\frac{D}{2}-1} m^{\frac{1}{D}}} \left(48 L \Lambda_{\mathcal{X}}\right)^D \ln^{\frac{3}{2}}\left(12 m^{\frac{1}{D}}\right).$$

Proof Substitute Inequality (5.11) in the right-hand side of Inequality (5.14) to obtain

$$\ln \mathcal{N}_2\left(\epsilon, \mathcal{F}_{\mathcal{G},\gamma}, m\right) \leqslant 20C \left(\frac{48 L \Lambda_{\mathcal{X}}}{\epsilon} \right)^D \ln\left(\frac{6\gamma}{\epsilon}\right).$$

Applying this bound in the chaining formula (4.4) yields:

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant h(N) + \frac{2\sqrt{20C}}{\sqrt{m}} \left(48 L \Lambda_{\mathcal{X}}\right)^D \sum_{j=1}^{N} \frac{h(j) + h(j-1)}{h(j)^{\frac{D}{2}}} \ln^{\frac{1}{2}}\left(\frac{6\gamma}{h(j)}\right).$$

Let $h(j) = \gamma 2^{-j}$ and $N = \left\lceil \log_2(m^{\frac{1}{D}}) \right\rceil$. Then, by a straightforwad computation it follows that

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \frac{\gamma}{m^{\frac{1}{D}}} + \frac{6\sqrt{20C}}{\gamma^{\frac{D}{2}-1} m^{\frac{1}{D}}} \left(48 L \Lambda_{\mathcal{X}}\right)^D \left\lceil \log_2(m^{\frac{1}{D}}) \right\rceil \ln^{\frac{1}{2}}\left(12 m^{\frac{1}{D}}\right).$$

■

5.5 Conclusions

In this chapter, we employed several approaches to decompose the fat-shattering dimension of the margin class in terms of that of the component classes. The first approach was based on an argument made by Mendelson [59] that allows one to upper bound the fat-shattering dimension of a class via its Rademacher complexity. The literature provides efficient upper bounds on the Rademacher complexity of linear classifiers: the bound of Lei et al. [55], and its improvement due to Maurer [57]. Particularly, using Maurer's approach we could upper bound the fat-shattering dimension with as sharp as a $O(\ln C)$ dependence.

The second approach was based on the fact that there exists a converse relation between the fat-shattering dimension and the metric entropy, the line of work due to Bartlett et al. [11] in the L_1-norm, and Talagrand [82] in the L_2-norm. In this chapter, we demonstrated that there is rather a straightforward relation between the fat-shattering dimension and the L_∞-norm metric entropy. Then, appealing to the combinatorial bound of Alon et al. [1] and using the decomposition result for the covering numbers in the L_∞-norm, produced a decomposition growing as a $O(C \ln^2(C))$. For classes of linear functions, this dependence could be improved to a $O(C \ln(C))$ if one appeals to the metric entropy bound of Zhang [92]. The extension of this entropy bound to matrices by Bartlett and co-authors [10], allows one to make use of a coupling assumption between the component functions, and eliminate the linear factor C. Consequently, we obtained a decomposition result scaling as a $O(\ln C)$, similar to the case with Maurer's approach.

We also considered the decomposition of the fat-shattering dimension of yet another family of classifiers: Lipschitz classifiers on doubling spaces [32]. We generalized the bound on the fat-shattering dimension, Corollary 4 in [32], to the multi-category case in a rather straightforwad fashion with a linear dependency on C.

Combining the decomposition of the fat-shattering dimension of general function classes and the combinatorial bound of Guermeur [37], Lemma 7, led to a new combinatorial bound for the margin class involving the fat-shattering dimensions of the component classes. Comparing to a similar result, Corollary 1 of Chapter 4, the advantage of the new bound is that it is dimension-free, and there is no dependency on C in the scale of the fat-shattering dimension involved. The latter, however, now also appears inside a squared logarithmic term, which is the price to pay for this gain. This result was applied in two contexts.

First, we were interested in elaborating the dependency on the number of categories of the sample complexity for the deviation probability (2.2) of Chapter 1. To this end, we used the new combinatorial bound, and for a comparison, we considered decomposing directly at the level

of the covering number. For the latter, we considered the L_∞-norm covering number, since under this norm, the decomposition for the covering numbers is optimized with respect to C. Although both sample complexity estimates scale as a $O(C \ln^2 C)$, the result based on the new combinatorial bound degrades the dependency on ϵ.

In the context of the chaining, on the other hand, the application of the new combinatorial bound improved the dependency on C to a sharper than radical one. As expected, the appearance of the fat-shattering dimension inside a squared logarithmic term slightly deteriorated the convergence rate in m compared to Theorem 18 of Chapter 4. Overall, this bound is comparable to Theorem 13 of Chapter 4 displaying a $O\left(\sqrt{C}\right)$ dependence.

Finally, we dedicated the chaining bound to Lipschitz classifiers and applied the corresponding decomposition of the-shattering dimension to generalize Lemma 7 of Gottlieb and coauthors [32] to the multi-category setting. This generalization yielded a radical dependency on C, and the matching with Lemma 7 dependency (up to a logarithmic factor) on m.

Chapter 6

Conclusions and Future Work

This thesis deals with the generalization performance of margin multi-category classifiers. We studied the uniform deviation of the empirical margin risk from the risk in terms of the basic parameters: the number C of classes, the sample size m and the margin parameter γ. The margin risk of a classifier is defined based on the margin loss function. Both in the bi-category and multi-category settings, the generalization performance is well studied for the indicator margin loss function. However, there still remain open questions when the margin loss function used is Lipschitz continuous.

One of the open questions is to control the uniform deviation between the margin risk and the empirical one by a covering number. Chapter 2 addressed this problem: under the assumption that the fat-shattering dimension of the margin class is finite, we upper bounded the deviation probability in terms of a L_1-norm covering number. This result was then translated into a sample complexity estimate. For fixed $\epsilon, \delta \in (0, 1)$, the sample complexity is the minimal sample size sufficient for the empirical risk to be ϵ close to the risk with probability at least $1 - \delta$. We derived several estimates appealing to different combinatorial bounds, the results that relate the covering number to a combinatorial dimension, which included a bound that depends on the sample-size and those that do not, i.e., the dimension-free bounds. This line of work corresponds to a multi-category extension of that of Bartlett and Long [12]. The combinatorial bound they appealed to was their extension of the L_∞-norm bound of Alon and coauthors [1] to the L_1-norm. We used the L_1 and L_2-norm combinatorial bounds [36, 61]. Apart from ϵ, δ, the sample complexity is also characterized by a combinatorial dimension of the function class, in this thesis, as in [12], the fat-shattering dimension. The sample complexity estimates that we obtained based on the L_1-norm combinatorial bounds match with that of Bartlett and Long [12] in terms of the dependency on ϵ and δ. On the other hand, we showed that the dimension-free L_2-norm bound

improves the dependency on ϵ, at the cost of degrading the scale of the fat-shattering dimension. A particularity of our extension is that, due to the use of a margin loss function, our sample complexity estimates also involve the margin parameter γ. In these results, this parameter appears in the scale of the fat-shattering dimension, implying that the sample complexity increases more rapidly as $\gamma \to 0$.

To make explicit the dependency of the sample complexity on the number of categories, we considered a particular kind of result, a decomposition result, that allows one to estimate a capacity of the margin class in terms of that of the component classes. We derived sample complexity estimates with explicit dependencies on C in two ways via the decomposition at the level of the L_∞-norm covering number or of the fat-shattering dimension. Both results exhibit the same $O(C \ln^2 C)$ dependence.

Another open question concerns the generalization bound for margin multi-category classifiers involving a Rademacher complexity. More precisely, we were interested in optimizing the dependencies on C and m of this capacity measure under the following very general assumptions: i) no interactions are assumed between the component functions, and ii) the fat-shattering dimensions of the component classes grow no faster than polynomially with the inverse of their scales. In the bi-class setting, under the aforementioned polynomial growth assumption, Mendelson [60] elaborated the convergence rate by relating the Rademacher complexity to another capacity measure, the metric entropy, and the latter to the fat-shattering dimension. Implementing this pathway in the multi-class setting, one has a choice to decompose at the level of any of these capacity measures to elaborate the dependencies on the basic parameters, especially C. We addressed this question as follows.

In Chapter 3, we gave a literature review of the decomposition results for the Rademacher complexity and we discussed the way they impact the dependency on C. For independent component classes, the decompositions [53, 57] scale at best linearly with the number of categories.

Chapter 4 related the Rademacher complexity to the metric entropy via the chaining method, and performed the decomposition at this level. Guermeur [36] showed that postponing the decomposition to the level of the metric entropy opens up a possibility to improve the dependency on C to a sublinear one. This work was based on the use of the L_2-norm metric entropy. We improved upon this result in two ways. First, since the decomposition result for the covering numbers is optimized with respect to C in the L_∞-norm, one can obtain a chaining bound with a radical dependency on C using the combinatorial bound of Alon et al. [1], and this is irrespective of the growth rate of the fat-shattering dimensions. The dependency on the sample size, however, is worsened. Second, by generalizing the combinatorial bound of Mendelson and Vershynin [61]

to L_p-norms for p between 2 and ∞, we could improve the dependency on C upon that in [36]. Although inferior to that in the L_∞-norm, the use of combinatorial bound did not degrade the convergence rate in m (nor in the margin parameter γ).

Finally, Chapter 5 considered the decomposition at the last level: that of the fat-shattering dimension. The decomposition at this level was based on the converse relationships between the fat-shattering dimension and the other two capacity measures. Our decomposition results based on the Rademacher complexity concerned only classes of linear functions. For such classes, Lei et al. [55], and Maurer [57] obtained a sublinear dependency on C of the Rademacher complexity by employing a rather natural coupling assumption between the component functions. Particularly, the decomposition of the fat-shattering dimension based on Maurer's result yields a $O(\ln C)$ dependency. For the decomposition via the metric entropy, on the other hand, we considered both specific and general function classes. For linear classes, using the matrix covering of Bartlett et al. [10], which also exploits a coupling assumption, led to a logarithmic dependency on C, similar to the one based on Maurer's approach. For general classes of functions, under no coupling assumption, which was the main focus of this thesis, this dependency worsens. Yet, appealing to the L_∞-norm metric entropy, we could obtain a decomposition result with a growth rate no more than $O(C \ln^2 C)$, independently of the capacities of the component classes. Using this decomposition, we derived a new combinatorial bound for the margin class in terms of the fat-shattering dimensions of the component classes. This result is dimension-free and there is no dependency on C in the scale of the fat-shattering dimension, an improvement over our decomposition of L_p-norm metric entropies in Chapter 4. This gain propagates in the chaining yielding a better than radical dependency on C, which is comparable to the L_∞-norm result of Chapter 4 (obtained through the decomposition of the metric entropy).

Our main results lead to the conclusion that there is a trade-off when optimizing the dependencies on the parameters of interest, particularly C and m: improving the dependency on C deteriorates the one on m. The basic question is how much milder these dependencies could be made.

We saw that for specific families of classifiers, exploiting a coupling assumption leads to an improved dependency of the metric entropy, and the Rademacher complexity on the number of categories. It is based on the fact that this kind of assumption allows one to eliminate the linear factor C from the bounds, the consequence of treating the component functions independently. Whether one can take benefit from a coupling assumption for general function classes, rather than just for linear ones, is an essential open question.

Improving the dependency of the Rademacher complexity on the basic parameters, C, m and

γ, could also be addressed at the level of the construction of the chaining. Since the choice of the function h, which gives the radius of balls of a minimal cover of a set, involves these parameters, optimizing it could lead to a tighter chaining bound. This concerns especially the optimization of the chaining bound with respect to γ which was not in the focus of this work. One possible approach would be based on the fact that γ constitutes the diameter of the truncated margin class for which the chaining bound is actually derived, and it is upper bounded by 1. Then, using the maximum value of γ when setting the function h would improve the growth rate of the chaining bound with respect to this parameter.

Regarding the dependency on the sample size, an important source for improvement lies in the combinatorial bound of Alon and coauthors [1]. As mentioned above, this bound concerns the covering numbers in the L_∞-norm, for which the decomposition formula takes the optimal form with respect to the dependency on C, and this is precisely the reason why we obtained a tighter than radical dependency on C in Chapter 5. But because the mentioned bound grows as $O(\ln^2(m))$, the dependency on m is worsened. In fact, it has been questioned in [1] whether the exponent 2 could be reduced to 1. This question has been addressed by Rudelson and Vershynin [73] by reducing the exponent 2 to any number larger than 1, at the cost of deteriorating the scale of the fat-shattering dimension. It appears, then, that to answer affirmatively the question posed by Alon and coauthors, without incurring any trade-off, is indeed a non-trivial endeavour. As a matter of fact, if it were possible, then it would allow one to improve the dependency on the number of categories without degrading that on the sample size.

In the context of extreme classification, one interesting question is whether it is possible to exploit the unbalanced class distribution phenomenon (when some classes are overrepresented in data) [5] to derive dedicated generalization bounds with a better dependency on C. One possible approach, for instance, would be to handle the capacities of the component classes in a non-uniform way based on coupling constraints inspired from the approach of Maurer, and Lei and coauthors. The knowledge of the distribution of the classes would particularly help in constraining the component classes, thus leading to tighter bounds.

Appendix A

Capacity Measures

The present Appendix gathers some results related to the Rademacher/Gaussian complexity, metric entropy and the fat-shattering dimension.

Proof of Lemma 4 Let $\mathbf{t}_n = (t_i)_{1 \leqslant i \leqslant n} \in \mathcal{T}^n$. We have

$$\hat{R}_n(\psi \circ \mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}_n}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \psi(f(t_i))\right] = \mathbb{E}_{\boldsymbol{\sigma}_{n-1}} \mathbb{E}_{\sigma_n}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \psi(f(t_i)) \Big| \boldsymbol{\sigma}_{n-1}\right],$$

where $\boldsymbol{\sigma}_{n-1} = (\sigma_i)_{1 \leqslant i \leqslant n-1}$. Then,

$$\begin{aligned}
&\hat{R}_n(\psi \circ \mathcal{F}) \\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}_{n-1}}\left[\sup_{f \in \mathcal{F}} \left(\sum_{i=1}^{n-1} \sigma_i \psi(f(t_i)) + \psi(f(t_n))\right)\right] + \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}_{n-1}}\left[\sup_{f \in \mathcal{F}} \left(\sum_{i=1}^{n-1} \sigma_i \psi(f(t_i)) - \psi(f(t_n))\right)\right] \\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}_{n-1}}\left[\sup_{f, f' \in \mathcal{F}} \left(\sum_{i=1}^{n-1} \left(\sigma_i \psi(f(t_i)) + \sigma_i \psi(f'(t_i))\right) + \psi(f(t_n)) - \psi(f'(t_n))\right)\right].
\end{aligned}$$

Next, we make use of L-Lipschitz property of ψ:

$$\begin{aligned}
\hat{R}_n(\psi \circ \mathcal{F}) &\leqslant \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}_{n-1}}\left[\sup_{f, f' \in \mathcal{F}} \left(\sum_{i=1}^{n-1} \left(\sigma_i \psi(f(t_i)) + \sigma_i \psi(f'(t_i))\right) + L\left|f(t_n) - f'(t_n)\right|\right)\right] \\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}_{n-1}}\left[\sup_{f, f' \in \mathcal{F}} \left(\sum_{i=1}^{n-1} \left(\sigma_i \psi(f(t_i)) + \sigma_i \psi(f'(t_i))\right) + L\left(f(t_n) - f'(t_n)\right)\right)\right] \\
&= \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}_{n-1}}\left[\sup_{f \in \mathcal{F}} \left(\sum_{i=1}^{n-1} \sigma_i \psi(f(t_i)) + Lf(t_n)\right)\right] \\
&\quad + \frac{1}{2}\mathbb{E}_{\boldsymbol{\sigma}_{n-1}}\left[\sup_{f \in \mathcal{F}} \left(\sum_{i=1}^{n-1} \sigma_i \psi(f(t_i)) - Lf(t_n)\right)\right],
\end{aligned}$$

where the first equality is due to the fact that the supremum is taken with respect to f, f' from the same set and the last one is based on

$$\sup_{f,f' \in \mathcal{F}} \big(f(t) + f'(t)\big) = \sup_{f \in \mathcal{F}} f(t) + \sup_{f' \in \mathcal{F}} f'(t). \tag{A.1}$$

Then,

$$\hat{R}_n(\psi \circ \mathcal{F}) \leqslant \mathbb{E}_{\sigma_n} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{n-1} \left(\sigma_i \psi(f\left(t_i\right))\right) + \sigma_n L f\left(t_n\right) \right].$$

Repeating it for the remaining terms involving $\sigma_1, ..., \sigma_{n-1}$ yields the required result.

Proof of Lemma 5 The proof is based on the following reasoning. By the assumption we have that

$$\forall f_1, \ldots, f_N \in \mathcal{F},\ \forall t \in \mathcal{T}, \quad \sum_{j=1}^{N} \alpha_j f_j(t) \leqslant \sum_{j=1}^{N} \alpha_j f_l(t) \leqslant f_l(t),$$

where $l = \operatorname{argmax}_j f_j(t)$. Consequently, $\sup_{\alpha_1,\ldots,\alpha_j} \sum_{j=1}^{N} \alpha_j f_j(t) = f_l(t)$ is achieved for the configuration where $\alpha_l = 1$ and $\alpha_j = 0$ for all $j \neq l$. Then,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\sigma}_n} \left[\sup_{f \in \operatorname{conv}(\mathcal{F})} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f\left(t_i\right) \right] &= \mathbb{E}_{\boldsymbol{\sigma}_n} \left[\sup_{\substack{f_1,\ldots,f_N \in \mathcal{F} \\ \alpha_1,\ldots,\alpha_N}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \sum_{j=1}^{N} \alpha_j f_j\left(t_i\right) \right] \\
&= \mathbb{E}_{\boldsymbol{\sigma}_n} \left[\sup_{\substack{f_1,\ldots,f_N \in \mathcal{F} \\ \alpha_1,\ldots,\alpha_N}} \sum_{j=1}^{N} \alpha_j \frac{1}{n} \sum_{i=1}^{n} \sigma_i f_j\left(t_i\right) \right] \\
&= \mathbb{E}_{\boldsymbol{\sigma}_n} \left[\sup_{\alpha_1,\ldots,\alpha_N} \sum_{j=1}^{N} \alpha_j \sup_{f_j \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f_j\left(t_i\right) \right] \\
&= \mathbb{E}_{\boldsymbol{\sigma}_n} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f\left(t_i\right) \right],
\end{aligned}$$

where the third equality holds thanks to the positivity of all α_j.

Proof of Lemma 6 Based on the fact that a standard Gaussian random variable is symmetric and using Jensen's inequality, it holds that

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

Proof of Theorem 11 It holds that

$$\begin{aligned}\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) =& \mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g\in\mathcal{G}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f_{g,\gamma}\left(z_i\right)\right] \\ =& \mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g\in\mathcal{G}}\frac{1}{2m}\sum_{i=1}^{m}\sigma_i \max\left(0, g_{y_i}\left(x\right) - \max_{1\leqslant k\leqslant C}\left(g_k\left(x_i\right) - 2\gamma\mathbb{1}_{\{k=y_i\}}\right)\right)\right] \\ \leqslant& \mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g\in\mathcal{G}}\frac{1}{2m}\sum_{i=1}^{m}\sigma_i \left(g_{y_i}\left(x\right) - \max_{1\leqslant k\leqslant C}\left(g_k\left(x_i\right) - 2\gamma\mathbb{1}_{\{k=y_i\}}\right)\right)\right].\end{aligned}$$

The last inequality is derived using the contraction principle, Lemma 4, since the $\max\left(0,\cdot\right)$ function is 1-Lipschitz. Next, based on the sub-additivity of the supremum and the symmetry of a Rademacher variable, we have

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g\in\mathcal{G}}\frac{1}{2m}\sum_{i=1}^{m}\sigma_i g_{y_i}\left(x_i\right)\right] + \mathbb{E}_{\boldsymbol{\sigma}_m}\left[\sup_{g\in\mathcal{G}}\frac{1}{2m}\sum_{i=1}^{m}\sigma_i \max_{1\leqslant k\leqslant C}\left(g_k\left(x_i\right) - 2\gamma\mathbb{1}_{\{k=y_i\}}\right)\right].$$

According to Lemma 6, we can upper bound both terms by the Gaussian complexity of the corresponding class:

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant \frac{1}{2m}\sqrt{\frac{\pi}{2}}\mathbb{E}_{\hat{\boldsymbol{o}}_m}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{m}\hat{o}_i g_{y_i}\left(x_i\right)\right] + \frac{1}{2m}\sqrt{\frac{\pi}{2}}\mathbb{E}_{\tilde{\boldsymbol{o}}_m}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{m}\tilde{o}_i \max_{1\leqslant k\leqslant C}\left(g_k(x_i) - 2\gamma\mathbb{1}_{\{k=y_i\}}\right)\right],$$

where $\hat{\boldsymbol{o}}_m = \left(\hat{o}_i\right)_{1\leqslant i\leqslant m}$ and $\tilde{\boldsymbol{o}}_m = \left(o_i\right)_{1\leqslant i\leqslant m}$ are orthogaussian sequences. Finally, applying a comparison result for Gaussian processes, Lemma 31, to both terms and based on the fact that a standard Gaussian random variable is centered (which eliminates the term $2\gamma\mathbb{1}_{\{k=y_i\}}$ for all i) conclude the proof.

Proof of Lemma 14 Clearly, the claim holds for $n = 1$. Let $n = 2$. Then, according to Lemma 30 it holds that

$$\begin{aligned}\mathbb{E}_{\sigma_2}\mathbb{E}_{\sigma_1}\left[\sup_{g\in\mathcal{G}}\sigma_1 f_g\left(z_1\right) + \sigma_2 f_g\left(z_2\right)\Big|\sigma_2\right] &\leqslant \mathbb{E}_{\sigma_2}\mathbb{E}_{\boldsymbol{\sigma}_C}\left[\sup_{g\in\mathcal{G}}\sum_{k=1}^{C}\sqrt{2}\sigma_k g_k(x_1) + \sigma_2 f_g\left(z_2\right)\Big|\sigma_2\right] \\ &= \mathbb{E}_{\boldsymbol{\sigma}_C}\mathbb{E}_{\sigma_2}\left[\sup_{g\in\mathcal{G}}\left(\sigma_2 f_g\left(z_2\right) + \sum_{k=1}^{C}\sqrt{2}\sigma_k g_k(x_1)\right)\Big|\boldsymbol{\sigma}_C\right] \\ &\leqslant \sqrt{2}\mathbb{E}_{\boldsymbol{\sigma}_{2C}}\sup_{g\in\mathcal{G}}\sum_{i=1}^{2}\sum_{k=1}^{C}\sigma_{C(i-1)+k} g_k(x_i),\end{aligned}$$

where $\boldsymbol{\sigma}_C = (\sigma_k)_{1\leqslant k\leqslant C}$. Assume now that $n \geqslant 3$ and that the claim holds for $n-1$. Then,

$$\begin{aligned}
\mathbb{E}_{\sigma_n}\mathbb{E}_{\boldsymbol{\sigma}_{n-1}} &\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{n-1}\sigma_i f_g(z_i) + \sigma_n f_g(z_n)\,\middle|\sigma_n\right] \\
&\leqslant \mathbb{E}_{\boldsymbol{\sigma}_{C(n-1)}}\mathbb{E}_{\sigma_n}\left[\sup_{g\in\mathcal{G}}\left(\sigma_n f_g(z_n) + \sum_{i=1}^{n-1}\sum_{k=1}^{C}\sqrt{2}\sigma_{C(i-1)+k}g_k(x_i)\right)\middle|\boldsymbol{\sigma}_{C(n-1)}\right] \\
&\leqslant \mathbb{E}_{\boldsymbol{\sigma}_{C(n-1)}}\mathbb{E}_{\tilde{\boldsymbol{\sigma}}_C}\left[\sup_{g\in\mathcal{G}}\left(\sum_{k=1}^{C}\sqrt{2}\tilde{\sigma}_k g_k(x_n) + \sum_{i=1}^{n-1}\sum_{k=1}^{C}\sqrt{2}\sigma_{C(i-1)+k}g_k(x_i)\right)\middle|\boldsymbol{\sigma}_{C(n-1)}\right] \\
&= \sqrt{2}\mathbb{E}_{\boldsymbol{\sigma}_{Cn}}\sup_{g\in\mathcal{G}}\sum_{i=1}^{n-1}\sum_{k=1}^{C}\sigma_{C(i-1)+k}g_k(x_i),
\end{aligned}$$

where $\tilde{\boldsymbol{\sigma}}_C = (\tilde{\sigma}_i)_{1\leqslant i\leqslant C}$.

Lemma 29 (After Lemma 2 in [47]) *Let $J \in \mathbb{N}^*$. For any $j \in [\![1, J]\!]$, let $\mathcal{F}_j$ be a class of functions from $\mathcal{T}$ to $[-M_{\mathcal{F}}, M_{\mathcal{F}}]$ with $M_{\mathcal{F}} \in \mathbb{R}_+$. Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ to $[-M_{\mathcal{F}}, M_{\mathcal{F}}]$ defined as $\mathcal{F} = \left\{\max_{1\leqslant j\leqslant J} f_j : (f_j)_{1\leqslant j\leqslant J} \in \prod_{j=1}^{J}\mathcal{F}_j\right\}$. Then,*

$$\hat{R}_n(\mathcal{F}) \leqslant \sum_{j=1}^{J}\hat{R}_n(\mathcal{F}_j).$$

Lemma 30 (After Lemma 7 in [57]) *Let $\mathcal{G}$ be a class of functions satisfying Definition 3 and for any $\gamma \in (0,1]$, let $\mathcal{F}_{\mathcal{G},\gamma}$ be a class of functions satisfying Definition 8. Let $\boldsymbol{\sigma}_C = (\sigma_i)_{1\leqslant i\leqslant C}$ be a Rademacher sequence and $u : \mathcal{G} \to \mathbb{R}$. Then, for any fixed $(x, y) \in \mathcal{X} \times [\![1, C]\!]$,*

$$\mathbb{E}\sup_{g\in\mathcal{G}}\left(\sigma f_{g,\gamma}(x,k) + u(g)\right) \leqslant \sqrt{2}\mathbb{E}_{\boldsymbol{\sigma}_C}\sup_{g\in\mathcal{G}}\left(\sum_{k=1}^{C}\sigma_k g_k(x) + u(g)\right). \tag{A.2}$$

Proof As the truncation of functions f_g play no role to obtain the claimed bound (because of the Lipschitz continuity of $f_{g,\gamma}$), to keep the notation simple, we prove it for functions f_g. By definition of a Rademacher variable and Lemma 13, it holds that

$$\begin{aligned}
2\mathbb{E}\sup_{g\in\mathcal{G}}\left(\sigma f_g(x,k) + u(g)\right) &\leqslant \sup_{g\in\mathcal{G}}\left(f_g(x,k) + u(g)\right) + \sup_{g\in\mathcal{G}}\left(-f_g(x,k) + u(g)\right) \\
&\leqslant \sup_{g,g'\in\mathcal{G}}\left(f_g(x,k) - f_{g'}(x,k) + u(g) + u(g')\right) \\
&\leqslant \sup_{g,g'\in\mathcal{G}}\left(\left\|g(x) - g'(x)\right\|_2 + u(g) + u(g')\right).
\end{aligned} \tag{A.3}$$

Now, one can apply Khintchine's inequality, Lemma 12, to $\|g(x) - g'(x)\|_2$. For $p = 1$, the smallest value for the constant in this inequality is due to Szarek [79]: $K_1 = \sqrt{2}$. Applying it in

Inequality (A.3), then using Jensen's inequality, the fact that the supremum is with respect to two elements of the same set and the definition of a Rademacher variable, we derive

$$
\begin{aligned}
2\mathbb{E}\sup_{g\in\mathcal{G}}\left(\sigma f_g\left(x,k\right)+u(g)\right) &\leqslant \sqrt{2}\sup_{g,g'\in\mathcal{G}}\mathbb{E}_{\boldsymbol{\sigma}_C}\left(\left|\sum_{k=1}^{C}\sigma_k\left(g_k(x)-g'_k(x)\right)\right|+u(g)+u(g')\right)\\
&\leqslant \sqrt{2}\mathbb{E}_{\boldsymbol{\sigma}_C}\sup_{g,g'\in\mathcal{G}}\left(\sum_{k=1}^{C}\sigma_k\left(g_k(x)-g'_k(x)\right)+u(g)+u(g')\right)\\
&= \sqrt{2}\mathbb{E}_{\boldsymbol{\sigma}_C}\sup_{g\in\mathcal{G}}\left(\sum_{k=1}^{C}\sigma_k g_k(x)+u(g)\right)+\sqrt{2}\mathbb{E}_{\boldsymbol{\sigma}_C}\sup_{g\in\mathcal{G}}\left(\sum_{k=1}^{C}-\sigma_k g_k(x)+u(g)\right)\\
&= 2\sqrt{2}\mathbb{E}_{\boldsymbol{\sigma}_C}\sup_{g\in\mathcal{G}}\sum_{k=1}^{C}\left(\sigma_k g_k(x)+u(g)\right).
\end{aligned}
$$

■

Lemma 31 (After Lemma 4 in [55]) *Let $\mathcal{G}$ be a class of functions given in Definition 3 and let $\hat{\boldsymbol{o}}_n=(\hat{o}_i)_{1\leqslant i\leqslant n}$, $\tilde{\boldsymbol{o}}_n=(\tilde{o}_i)_{1\leqslant i\leqslant n}$ and $\boldsymbol{o}_{Cn}=(o_j)_{1\leqslant j\leqslant Cn}$ be orthogaussian sequences. Then*

$$
\mathbb{E}_{\hat{\boldsymbol{o}}_n}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{n}\hat{o}_i g_{y_i}(x_i)\right]\leqslant\mathbb{E}_{\boldsymbol{o}_{Cn}}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{n}\sum_{k=1}^{C}o_{C(i-1)+k}g_k\left(x_i\right)\right],
$$

and

$$
\mathbb{E}_{\tilde{\boldsymbol{o}}_n}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{n}\tilde{o}_i\max_{1\leqslant k\leqslant C}g_k\left(x_i\right)\right]\leqslant\mathbb{E}_{\boldsymbol{o}_{Cn}}\left[\sup_{g\in\mathcal{G}}\sum_{i=1}^{n}\sum_{k=1}^{C}o_{C(i-1)+k}g_k\left(x_i\right)\right].
$$

Proof Consider the following Gaussian processes:

$$
\left\{A_g=\sum_{i=1}^{n}\hat{o}_i g_{y_i}(x_i):\ g\in\mathcal{G}\right\}
$$

and

$$
\left\{B_g=\sum_{i=1}^{n}\sum_{k=1}^{C}o_{C(i-1)+k}g_k\left(x_i\right):\ g\in\mathcal{G}\right\}.
$$

According to a comparison theorem for Gaussian processes, Theorem 1 in [89], for the proof it suffices to demonstrate that for any $g,g'\in\mathcal{G}$,

$$
\mathbb{E}_{\tilde{\boldsymbol{o}}_n}\left[\left(A_g-A_{g'}\right)^2\right]\leqslant\mathbb{E}_{\boldsymbol{o}_{Cn}}\left[\left(B_g-B_{g'}\right)^2\right]. \tag{A.4}
$$

Based on the definition of an orthogaussian sequence, we have that

$$\begin{aligned}\mathbb{E}_{\hat{\boldsymbol{o}}_n}\left[\left(\sum_{i=1}^{n}\hat{o}_i g_{y_i}(x_i)-\sum_{i=1}^{n}\hat{o}_i g'_{y_i}(x_i)\right)^2\right] &= \mathbb{E}_{\hat{\boldsymbol{o}}_n}\left[\sum_{i=1}^{n}\hat{o}_i^2\left(g_{y_i}(x_i)-g'_{y_i}(x_i)\right)^2\right]\\ &= \sum_{i=1}^{n}\left(g_{y_i}(x_i)-g'_{y_i}(x_i)\right)^2\\ &\leqslant \sum_{i=1}^{n}\sum_{k=1}^{C}\left(g_k(x_i)-g'_k(x_i)\right)^2. \qquad \text{(A.5)}\end{aligned}$$

Similarly,

$$\mathbb{E}_{\boldsymbol{o}_{Cn}}\left[\left(\sum_{i=1}^{n}\sum_{k=1}^{C}o_{(i-1)C+k}\left(g_k\left(x_i\right)-g'_k\left(x_i\right)\right)\right)^2\right]=\sum_{i=1}^{n}\sum_{k=1}^{C}\left(g_k\left(x_i\right)-g'_k\left(x_i\right)\right)^2. \qquad \text{(A.6)}$$

From (A.5) and (A.6) it follows that Inequality (A.4) holds.

The second inequality is proved similarly based on the fact that

$$\begin{aligned}\mathbb{E}_{\tilde{\boldsymbol{o}}_n}\left[\left(\sum_{i=1}^{n}\tilde{o}_i\max_{1\leqslant k\leqslant C}g_k\left(x_i\right)-\sum_{i=1}^{n}\tilde{o}_i\max_{1\leqslant k\leqslant C}g'_k\left(x_i\right)\right)^2\right] &= \sum_{i=1}^{n}\left(\max_{1\leqslant k\leqslant C}g_k\left(x_i\right)-\max_{1\leqslant k\leqslant C}g'_k\left(x_i\right)\right)^2\\ &\leqslant \sum_{i=1}^{n}\max_{1\leqslant k\leqslant C}\left(g_k\left(x_i\right)-g'_k\left(x_i\right)\right)^2,\end{aligned}$$

the right-hand side being less than (A.6). ■

Theorem 28 (After Theorem 2 in [11]) *Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ to $[-M_{\mathcal{F}}, M_{\mathcal{F}}]$ with $M_{\mathcal{F}} \in \mathbb{R}_+$. Let $d = (2\epsilon)$-dim$(\mathcal{F})$. Assume that $d > 4$. Then, for all $\epsilon \in (0, M_{\mathcal{F}}]$,*

$$d \leqslant 8 \ln \mathcal{N}_1\left(\frac{\epsilon}{2}, \mathcal{F}, d\right).$$

Proof Let $S = \{t_i : 1 \leqslant i \leqslant d\} \subset \mathcal{T}$ be the set of maximal cardinality d 2ϵ-shattered by $\mathcal{F}$. Let $\mathbf{t}_d = (t_i)_{1\leqslant i\leqslant d}$ be a sequence of points in S. Let $\mathcal{F}'|_S$ be a set of functions in $\mathcal{F}$ restricted to S and outputting by width 2ϵ around the level function $u : \mathcal{T} \to \mathbb{R}$ in all 2^d ways. Fix $\bar{f}' \in \mathcal{F}'|_S$. For any $f' \in \mathcal{F}'|_S$, let $d_{f',\bar{f}'}$ be the number of points t in S for which $|\bar{f}'(t) - f'(t)| \geqslant 4\epsilon$. Then,

$$\frac{1}{d}\sum_{i=1}^{d}|f'(t_i)-\bar{f}'(t_i)| \geqslant \frac{4\epsilon d_{f',\bar{f}'}}{d} \qquad \text{(A.7)}$$

From (A.7) one can see that f' is ϵ-close to $\bar{f}'$ with respect to $d_{1,\mathbf{t}_d}$, i.e.,

$$\frac{1}{d}\sum_{i=1}^{d}|f'(t_i)-\bar{f}'(t_i)| < \epsilon,$$

if and only if $\frac{4\epsilon d_{f',\bar{f}'}}{d} < \epsilon$ or $d_{f',\bar{f}'} < \frac{d}{4}$. Thenf, the number $N_{\bar{f}'}$ of functions f' ϵ-close to $\bar{f}'$ can be computed by the following combinatorial formula:

$$N_{\bar{f}'} = \sum_{l=1}^{\lfloor d_{f',\bar{f}'} \rfloor} \binom{d}{l}. \tag{A.8}$$

This can be upper bounded according to Chernoff-Okamoto inequality [26]:

$$\sum_{l=0}^{m} \binom{n}{l} p^l (1-p)^{n-l} \leqslant \exp\left(\frac{-(np-m)^2}{2np(1-p)}\right),$$

where $p \leqslant 1/2$ and $m \leqslant np$. Let $p = 1/2$. Then,

$$\sum_{l=0}^{m} \binom{n}{l} \leqslant 2^n \exp\left(\frac{-2(n/2-m)^2}{n}\right).$$

Applying it to the right-hand side of Inequality (A.8), we obtain

$$N_{\bar{f}'} \leqslant 2^d \exp\left(\frac{-2\left(\frac{d}{2} - \lfloor d_{f',\bar{f}'} \rfloor\right)^2}{d}\right).$$

Substitute it in the upper bound on $d_{f',\bar{f}'}$ to obtain

$$\begin{aligned} N_{\bar{f}'} &< 2^d \exp\left(\frac{-2\left(\frac{d}{2} - \lfloor \frac{d}{4} \rfloor\right)^2}{d}\right) \\ &\leqslant 2^d \exp\left(\frac{-2\left(\frac{d}{2} - \frac{d}{4}\right)^2}{d}\right) \\ &= 2^d \exp\left(-\frac{d}{8}\right). \end{aligned}$$

Since there are 2^d functions in $\mathcal{F}'|_S$, the ratio $2^d/N_{\bar{f}'}$ lower bounds $\mathcal{N}\left(\epsilon, \mathcal{F}'|_S, d_{1,\mathbf{t}_d}\right)$. Thus, we have

$$\mathcal{N}\left(\epsilon, \mathcal{F}'|_S, d_{1,\mathbf{t}_d}\right) \geqslant \exp\left(\frac{d}{8}\right).$$

To switch to the covering number of $\mathcal{F}$, we use the triangle inequality which gives

$$\mathcal{N}\left(\epsilon, \mathcal{F}'|_S, d_{1,\mathbf{t}_d}\right) \leqslant \mathcal{N}\left(\frac{\epsilon}{2}, \mathcal{F}, d_{1,\mathbf{t}_d}\right).$$

Finally, the claim follows from

$$\mathcal{N}_1\left(\epsilon, \mathcal{F}, d\right) \leqslant \sup_{\mathbf{t}_d \in \mathcal{T}^d} \mathcal{N}\left(\epsilon, \mathcal{F}, d_{1,\mathbf{t}_d}\right).$$

■

Lemma 32 (Lemma B.2 in [55]) *Let T be a standard Gaussian random variable. For any $p > 0$, the p-th moment of T can be bounded as*

$$\mathbb{E}|T|^p \leqslant (2p)^{\frac{p}{2}+1}.$$

Lemma 33 (After Lemma 1 in [92]) *Let $\mathcal{H}$ denote a Hilbert space with the norm $\|\cdot\|_{\mathcal{H}}$. Let $h \in \mathcal{H}$ be as $\sum_{j=1}^{N} \alpha_j h_j$, where each $\alpha_j \geqslant 0$ and $\sum_{j=1}^{N} \alpha_j \leqslant 1$. Then, for any $K \geqslant 1$, there exist non-negative integers $(k_j)_{1 \leqslant j \leqslant N}$ satisfying $\sum_{j=1}^{N} k_j \leqslant K$, such that*

$$\left\| h - \frac{1}{K} \sum_{j=1}^{N} k_j h_j \right\|_{\mathcal{H}}^2 \leqslant \frac{\max_{1 \leqslant j \leqslant N} \|h_j\|_{\mathcal{H}}^2}{K}.$$

Appendix B

Basic Concentration Inequalities

Concentration inequalities are the results dealing with the deviation of a random variable from its expected value, and they are indispensable tools in the learning theory. This Appendix presents several basic concentration inequalities that we refer to in the main text. In the following, T stands for a non-negative real-valued random variable with law $P_{\mathcal{T}}$, and we assume that $\mathbb{E}[T] < \infty$.

Theorem 29 (Chebyshev's inequality) *For all $\epsilon > 0$,*

$$P_{\mathcal{T}}\left(|T - \mathbb{E}[T]| \geqslant \epsilon\right) \leqslant \frac{Var(T)}{\epsilon^2}. \tag{B.1}$$

Suppose that T has a moment generating function

$$\phi(\lambda) = \mathbb{E}\left[\exp\left(\lambda\left(T - \mathbb{E}T\right)\right)\right] < \infty,$$

for all $\lambda > 0$. The application of Markov's bound to $\exp\left(\lambda\left(T - \mathbb{E}T\right)\right)$ gives the following result.

Theorem 30 (Chernoff's inequality)

$$P_{\mathcal{T}}\left(T - \mathbb{E}[T] \geqslant \epsilon\right) = P_{\mathcal{T}}\left(\exp\left(\lambda\left(T - \mathbb{E}T\right)\right) \geqslant \exp\left(\lambda\epsilon\right)\right) \leqslant \frac{\phi(\lambda)}{\exp\left(\lambda\epsilon\right)}. \tag{B.2}$$

This result can be expressed as an optimization problem:

$$\ln P_{\mathcal{T}}\left(T - \mathbb{E}[T] \geqslant \epsilon\right) \leqslant \inf_{\lambda > 0}\left(\ln\phi(\lambda) - \lambda\epsilon\right). \tag{B.3}$$

When T is a standard Gaussian random variable, substituting its moment generating function

$$\mathbb{E}\left[\exp(\lambda T)\right] = \exp\left(\frac{\lambda^2}{2}\right)$$

into (B.3) gives:

$$\ln P_{\mathcal{T}}\left(T \geqslant \epsilon\right) \leqslant \inf_{\lambda > 0}\left(\frac{\lambda^2}{2} - \lambda\epsilon\right).$$

Taking derivative from the right-hand side and making it equal to zero we get $\lambda = \epsilon$. Thus, for a standard Gaussian random variable the Chernoff bound takes the following form:

$$P_{\mathcal{T}}\left(T \geqslant \epsilon\right) \leqslant \exp\left(-\frac{\epsilon^2}{2}\right) \text{ and } P_{\mathcal{T}}\left(T \leqslant -\epsilon\right) \leqslant \exp\left(-\frac{\epsilon^2}{2}\right).$$

Thus:

Proposition 2 (Gaussian tail probability) *For all $\epsilon > 0$,*

$$P_{\mathcal{T}}\left(|T| \geqslant \epsilon\right) \leqslant 2\exp\left(-\frac{\epsilon^2}{2}\right). \tag{B.4}$$

The moment generating function of a Rademacher random variable is dominated by that of a standard Gaussian random variable as shown below.

Proposition 3 *For any $\lambda > 0$,*

$$\mathbb{E}[\exp\left(\lambda\sigma\right)] \leqslant \exp\left(\frac{\lambda^2}{2}\right). \tag{B.5}$$

Thus, a Rademacher variable is said to be a *sub-Gaussian* random variable and the tail probability (B.4) applies to it. In fact, any bounded random variable is sub-Gaussian based on the following lemma.

Lemma 34 (Lemma 8.1 in [21]) *Let T be a random variable with $\mathbb{E}[T] = 0$ and $T \in [a, b]$ almost surely with $a, b \in \mathbb{R}$. Then, for any $\lambda > 0$,*

$$\mathbb{E}\left[\exp(\lambda T)\right] \leqslant \exp\left(\frac{\lambda^2(b-a)^2}{8}\right).$$

Combining this result with Chebyshev's inequality leads to the tail bound for the sum of independent and identically distributed random variables.

Theorem 31 (Hoeffding's inequality, Theorem 2 in [41]) *Let $(T_i)_{1\leqslant i\leqslant n}$ be a sequence of independent and identically distributed random variables such that $T_i \in [a_i, b_i]$ almost surely. Let $\bar{T} = \frac{1}{n}\sum_{i=1}^{n} T_i$. Then for all $\epsilon > 0$,*

$$P_{\mathcal{T}}\left(\bar{T} - \mathbb{E}\left[\bar{T}\right] \geqslant \epsilon\right) \leqslant \exp\left(\frac{-2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right). \tag{B.6}$$

If all T_i takes their values in $[-a, a]$ with $a > 0$ almost surely, then (B.6) becomes

$$P_{\mathcal{T}}\left(\bar{T} - \mathbb{E}\left[\bar{T}\right] \geqslant \epsilon\right) \leqslant \exp\left(\frac{-n\epsilon^2}{2a^2}\right). \tag{B.7}$$

In the case when the variance of T_i is much less than a, Bernstein's inequality provides a tighter control on the tail probability.

Theorem 32 (Bernstein's inequality [18]) *Let $a > 0$ and let T be a random variable taking its values in an interval $[-a, a]$ almost surely. Let $var(T)$ denote the variance of T. Let $(T_i)_{1 \leqslant i \leqslant n}$ be a sequence of n independent copies of T and let $\bar{T} = \frac{1}{n} \sum_{i=1}^{n} T_i$. Then for all $\epsilon > 0$,*

$$P_{\mathcal{T}} \left(\bar{T} - \mathbb{E}[T] \geqslant \epsilon \right) \leqslant \exp \left(\frac{-n\epsilon^2}{2var(T) + \frac{2a\epsilon}{3}} \right). \tag{B.8}$$

Hoeffding's bound can be generalized to a more complex setting. The following result applies to a function of independent random variables under the condition that it varies no more than a constant c_i when the value of the i-th random variable changes.

Theorem 33 (McDiarmid's inequality [58]) *Let $f : \mathcal{T}^n \to \mathbb{R}$ and let $(T_i)_{1 \leqslant i \leqslant n}$ be a sequence of n independent copies of T. If for all $i \in [\![1, n]\!]$,*

$$\sup_{t_1, \ldots, t_n, t_i' \in \mathcal{T}} |f(t_1, ..., t_i, ..., t_n) - f(t_1, ..., t_i', ..., t_n)| \leqslant c_i,$$

then

$$P_{\mathcal{T}} \left(f(T_1, ..., T_n) - \mathbb{E}[f(T_1, ..., T_n)] > \epsilon \right) \leqslant \exp \left(\frac{-2\epsilon^2}{\sum_{i=1}^{n} c_i^2} \right). \tag{B.9}$$

With the choice $f(T_1, ..., T_n) = \frac{1}{n} \sum_{i=1}^{n} T_i$, we obtain Hoeffding's inequality as the special case.

Appendix C

Symmetrization

The following result is the symmetrization for probability. Its proof is based on that of Lemma 2 of Vapnik and Chervonenkis [88].

Lemma 35 *Let $\mathbf{Z}_m = (Z_i)_{1 \leqslant i \leqslant m} \in \mathcal{Z}^m$ be a sequence of independent random variables having the same distribution as Z. Let $\mathbf{Z}'_m = (Z'_i)_{1 \leqslant i \leqslant m}$ be an independent copy of $\mathbf{Z}_m$. Fix $\epsilon > 0$ and $\gamma \in (0, 1]$. Then, for $m > \frac{2}{\epsilon^2}$,*

$$P^m \left(\sup_{g \in \mathcal{G}} (L_\gamma(g) - L_{\gamma,m}(g)) > \epsilon \right) \leqslant 2P^{2m} \left\{ \sup_{g \in \mathcal{G}} \left(\frac{1}{m} \sum_{i=1}^m \phi_\gamma \left(f_{g,\gamma} \left(Z'_i \right) \right) - \frac{1}{m} \sum_{i=1}^m \phi_\gamma \left(f_{g,\gamma}(Z_i) \right) \right) \geqslant \frac{\epsilon}{2} \right\}.$$

Proof Consider the two independent empirical processes:

$$\left\{ L_\gamma(g) - \frac{1}{m} \sum_{i=1}^m \phi_\gamma \left(f_{g,\gamma}(Z_i) \right) : g \in \mathcal{G} \right\} \text{ and } \left\{ L_\gamma(g) - \frac{1}{m} \sum_{i=1}^m \phi_\gamma (f_{g,\gamma} \left(Z'_i \right)) : g \in \mathcal{G} \right\}.$$

For any $g \in \mathcal{G}$, we have that

$$\begin{aligned} & L_\gamma(g) - \frac{1}{m} \sum_{i=1}^m \phi_\gamma \left(f_{g,\gamma}(z_i) \right) \geqslant \epsilon \text{ and } L_\gamma(g) - \frac{1}{m} \sum_{i=1}^m \phi_\gamma \left(f_{g,\gamma} \left(z'_i \right) \right) \leqslant \frac{\epsilon}{2} \\ \implies & \frac{1}{m} \sum_{i=1}^m \phi_\gamma \left(f_{g,\gamma} \left(z'_i \right) \right) - \frac{1}{m} \sum_{i=1}^m \phi_\gamma \left(f_{g,\gamma}(z_i) \right) \geqslant \frac{\epsilon}{2}. \end{aligned} \tag{C.1}$$

Let $\mathbf{z}_{2m} = (z_i)_{1 \leqslant i \leqslant 2m} \in \mathcal{Z}^{2m}$ and $z'_i = z_{m+i}$ for $i \leqslant m$. Denote $\mathbf{z}'_m = (z'_i)_{1 \leqslant i \leqslant m}$. By definition,

$$\begin{aligned} & P^{2m} \left\{ \sup_{g \in \mathcal{G}} \left(\frac{1}{m} \sum_{i=1}^m \phi_\gamma \left(f_{g,\gamma} \left(Z'_i \right) \right) - \frac{1}{m} \sum_{i=1}^m \phi_\gamma \left(f_{g,\gamma}(Z_i) \right) \right) \geqslant \frac{\epsilon}{2} \right\} \\ & = \int_{\mathcal{Z}^{2m}} \mathbb{1} \left\{ \sup_{g \in \mathcal{G}} \left(\frac{1}{m} \sum_{i=1}^m \phi_\gamma \left(f_{g,\gamma} \left(z'_i \right) \right) - \frac{1}{m} \sum_{i=1}^m \phi_\gamma \left(f_{g,\gamma}(z_i) \right) \right) \geqslant \frac{\epsilon}{2} \right\} dP^{2m}(\mathbf{z}_{2m}). \end{aligned} \tag{C.2}$$

According to Tonelli's theorem, it holds that

$$\begin{aligned}
&\int_{\mathcal{Z}^{2m}} \mathbb{1}\left\{\sup_{g\in\mathcal{G}}\left(\frac{1}{m}\sum_{i=1}^{m}\phi_\gamma\left(f_{g,\gamma}\left(z_i'\right)\right)-\frac{1}{m}\sum_{i=1}^{m}\phi_\gamma\left(f_{g,\gamma}(z_i)\right)\right)\geqslant\frac{\epsilon}{2}\right\}dP^{2m}(\mathbf{z}_{2m})\\
&=\int_{\mathcal{Z}^m}\left(\int_{\mathcal{Z}^m}\mathbb{1}\left\{\sup_{g\in\mathcal{G}}\left(\frac{1}{m}\sum_{i=1}^{m}\phi_\gamma\left(f_{g,\gamma}\left(z_i'\right)\right)-\frac{1}{m}\sum_{i=1}^{m}\phi_\gamma\left(f_{g,\gamma}(z_i)\right)\right)\geqslant\frac{\epsilon}{2}\right\}dP^m(\mathbf{z}_m')\right)dP^m(\mathbf{z}_m).
\end{aligned}\tag{C.3}$$

Now, let $\mathcal{Q}=\left\{\mathbf{z}_m\in\mathcal{Z}^m:\sup_{g\in\mathcal{G}}\left(L_\gamma(g)-\frac{1}{m}\sum_{i=1}^{m}\phi_\gamma\left(f_{g,\gamma}(z_i)\right)\right)>\epsilon\right\}$. By definition of the supremum, for any $\mathbf{z}_m$ in $\mathcal{Q}$, there exists $g^*\in\mathcal{G}$ such that

$$L_\gamma(g^*)-\frac{1}{m}\sum_{i=1}^{m}\phi_\gamma\left(f_{g^*,\gamma}(z_i)\right)\geqslant\epsilon.$$

We can write the following

$$\begin{aligned}
&\int_{\mathcal{Z}^m}\left(\int_{\mathcal{Z}^m}\mathbb{1}\left\{\sup_{g\in\mathcal{G}}\left(\frac{1}{m}\sum_{i=1}^{m}\phi_\gamma\left(f_{g,\gamma}\left(z_i'\right)\right)-\frac{1}{m}\sum_{i=1}^{m}\phi_\gamma\left(f_{g,\gamma}(z_i)\right)\right)\geqslant\frac{\epsilon}{2}\right\}dP^m(\mathbf{z}_m')\right)dP^m(\mathbf{z}_m)\\
&\geqslant\int_{\mathcal{Q}}\left(\int_{\mathcal{Z}^m}\mathbb{1}\left\{\frac{1}{m}\sum_{i=1}^{m}\phi_\gamma\left(f_{g^*,\gamma}\left(z_i'\right)\right)-\frac{1}{m}\sum_{i=1}^{m}\phi_\gamma\left(f_{g^*,\gamma}(z_i)\right)\geqslant\frac{\epsilon}{2}\right\}dP^m(\mathbf{z}_m')\right)dP^m(\mathbf{z}_m).
\end{aligned}\tag{C.4}$$

Now, based on the implication (C.1), for any $\mathbf{z}_m$ in $\mathcal{Q}$, we have

$$\begin{aligned}
&\mathbb{1}\left\{\frac{1}{m}\sum_{i=1}^{m}\phi_\gamma\left(f_{g^*,\gamma}\left(z_i'\right)\right)-\frac{1}{m}\sum_{i=1}^{m}\phi_\gamma\left(f_{g^*,\gamma}(z_i)\right)\geqslant\frac{\epsilon}{2}\right\}\\
&\geqslant\mathbb{1}\left\{L_\gamma(g^*)-\frac{1}{m}\sum_{i=1}^{m}\phi_\gamma\left(f_{g^*,\gamma}(z_i)\right)\geqslant\epsilon\right\}\cdot\mathbb{1}\left\{L_\gamma(g^*)-\frac{1}{m}\sum_{i=1}^{m}\phi_\gamma\left(f_{g^*,\gamma}\left(z_i'\right)\right)\leqslant\frac{\epsilon}{2}\right\}\\
&=1\cdot\mathbb{1}\left\{L_\gamma(g^*)-\frac{1}{m}\sum_{i=1}^{m}\phi_\gamma\left(f_{g^*,\gamma}\left(z_i'\right)\right)\leqslant\frac{\epsilon}{2}\right\}.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\int_{\mathcal{Z}^m}\mathbb{1}\left\{\frac{1}{m}\sum_{i=1}^{m}\phi_\gamma\left(f_{g^*,\gamma}\left(z_i'\right)\right)-\frac{1}{m}\sum_{i=1}^{m}\phi_\gamma\left(f_{g^*,\gamma}(z_i)\right)\geqslant\frac{\epsilon}{2}\right\}dP^m(\mathbf{z}_m')\\
&\geqslant\int_{\mathcal{Z}^m}\mathbb{1}\left\{L_\gamma(g^*)-\frac{1}{m}\sum_{i=1}^{m}\phi_\gamma\left(f_{g^*,\gamma}\left(z_i'\right)\right)\leqslant\frac{\epsilon}{2}\right\}dP^m(\mathbf{z}_m').
\end{aligned}$$

Taking this into account in (C.4) and by transitivity from (C.2)-(C.3), we deduce

$$\begin{aligned}
&P^{2m}\left\{\sup_{g\in\mathcal{G}}\left(\frac{1}{m}\sum_{i=1}^{m}\phi_\gamma\left(f_{g,\gamma}\left(Z_i'\right)\right)-\frac{1}{m}\sum_{i=1}^{m}\phi_\gamma\left(f_{g,\gamma}(Z_i)\right)\right)\geqslant\frac{\epsilon}{2}\right\}\\
&\geqslant\int_{\mathcal{Q}}\left(\int_{\mathcal{Z}^m}\mathbb{1}\left\{L_\gamma(g^*)-\frac{1}{m}\sum_{i=1}^{m}\phi_\gamma\left(f_{g^*,\gamma}\left(z_i'\right)\right)\leqslant\frac{\epsilon}{2}\right\}dP^m(\mathbf{z}_m')\right)dP^m(\mathbf{z}_m).
\end{aligned}\tag{C.5}$$

We now focus on the probability

$$\int_{\mathcal{Z}^m} \mathbb{1}\left\{L_\gamma(g^*) - \frac{1}{m}\sum_{i=1}^m \phi_\gamma\left(f_{g^*,\gamma}\left(z_i'\right)\right) \leqslant \frac{\epsilon}{2}\right\} dP^m(\mathbf{z}_m') = P^m\left\{L_\gamma(g^*) - \frac{1}{m}\sum_{i=1}^m \phi_\gamma\left(f_{g^*,\gamma}\left(Z_i'\right)\right) \leqslant \frac{\epsilon}{2}\right\}.$$

Since for all $i \in [\![1, m]\!]$, Z_i' admits the same distribution as Z, we have

$$L_\gamma(g^*) = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^m \phi_\gamma\left(f_{g^*,\gamma}\left(Z_i'\right)\right)\right],$$

and thus the tail probability $P^m\left(L_\gamma(g^*) - \frac{1}{m}\sum_{i=1}^m \phi_\gamma\left(f_{g^*,\gamma}\left(Z_i'\right)\right) \leqslant \frac{\epsilon}{2}\right)$ can be bounded using a concentration inequality. Chebyshev's inequality (B.1) yields a tighter bound. Since the variance of $\phi_\gamma\left(f_{g^*,\gamma}\left(Z\right)\right)$ has a bounded range, $[0, 1]$, by Popoviciu's inequality [70] we obtain

$$\operatorname{var}\left(\phi_\gamma\left(f_{g^*,\gamma}\left(Z\right)\right)\right) \leqslant \frac{1}{4}(1-0) = \frac{1}{4}.$$

Consequently, under the assumption that $m > \frac{2}{\epsilon^2}$,

$$\begin{aligned} P^m\left(L_\gamma(g^*) - \frac{1}{m}\sum_{i=1}^m \phi_\gamma\left(f_{g^*,\gamma}\left(Z_i'\right)\right) \leqslant \frac{\epsilon}{2}\right) &= 1 - P^m\left(L_\gamma(g^*) - \frac{1}{m}\sum_{i=1}^m \phi_\gamma\left(f_{g^*,\gamma}\left(Z_i'\right)\right) > \frac{\epsilon}{2}\right) \\ &\geqslant 1 - \frac{4m}{4m^2\epsilon^2} \\ &\geqslant \frac{1}{2}. \end{aligned}$$

Substituting it into (C.5) and taking into account that

$$\int_{\mathcal{Q}} 1 \cdot dP^m(\mathbf{z}_m) = P^m\left\{\sup_{g\in\mathcal{G}}\left(L_\gamma(g) - L_{\gamma,m}(g)\right) > \epsilon\right\},$$

we get the desired result. ■

Appendix D

Perceptron Mistake Bound

The following is the perceptron mistake bound extended to a Hilbert space.

Proposition 4 *Let $\mathcal{F}$ of a set of functions from a Hilbert space $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ to $\mathbb{R}$ defined as $h \mapsto \langle h, w\rangle$ with $w \in \mathbb{H}$. Let $((h_i, y_i))_{1\leqslant i\leqslant n} \in (\mathbb{H} \times \{-1, +1\})^n$ with $\max_{1\leqslant i\leqslant n} \|h_i\|_{\mathbb{H}} \leqslant R$. For $\alpha \in \mathbb{R}_+$, assume that there exists $f_\alpha \in \mathcal{F}$ such that*

$$\forall i \in [\![1, n]\!], \quad y_i f_\alpha(h_i) = y_i \langle w_\alpha, h_i\rangle \geqslant \alpha.$$

Let $f^* \in \mathcal{F}$ be such that

$$\forall i \in [\![1, n]\!], \quad y_i f^*(h_i) > 0.$$

Let $(i_k)_{1\leqslant k\leqslant m}$ be a sequence of indices of examples on which the perceptron makes an update before converging to f^*. Then, m is bounded by

$$m \leqslant \frac{\|w_\alpha\|_{\mathbb{H}}^2 R^2}{\alpha^2}, \tag{D.1}$$

and the function f^* takes the form

$$f^*(\cdot) = \left\langle \sum_{k=1}^{m} y_{i_k} h_{i_k}, \cdot \right\rangle. \tag{D.2}$$

Proof The perceptron loops over all examples and makes an update

$$w^{(k)} = w^{(k-1)} + \mu y_i h_i, \tag{D.3}$$

with a learning rate $\mu \in (0, 1]$ and the initial condition $w^{(0)} = 0$, whenever sign $\left(\langle w^{(k-1)}, h_i\rangle\right) \neq y_i$. We set $\mu = 1$ and combine the changes in the inner product $\langle w^{(m)}, w_\alpha\rangle$ and the squared norm $\left\|w^{(m)}\right\|_{\mathbb{H}}^2$ after m steps. By our assumption, after k steps,

$$\langle w^{(k)}, w_\alpha\rangle - \langle w^{(k-1)}, w_\alpha\rangle = \langle w^{(k-1)} + y_i h_i, w_\alpha\rangle - \langle w^{(k-1)}, w_\alpha\rangle = y_i \langle w_\alpha, h_i\rangle \geqslant \alpha$$

for some $i \in [\![1, n]\!]$. Thus, after m steps we have

$$\langle w^{(m)}, w_\alpha \rangle \geqslant m\alpha. \tag{D.4}$$

The change in the squared norm after k steps; on the other hand, is

$$\begin{aligned}
\left\|w^{(k)}\right\|_{\mathbb{H}}^2 - \left\|w^{(k-1)}\right\|_{\mathbb{H}}^2 &= \left\|w^{(k-1)} + y_i h_i\right\|_{\mathbb{H}}^2 - \left\|w^{(k-1)}\right\|_{\mathbb{H}}^2 \\
&= \left\|w^{(k-1)}\right\|_{\mathbb{H}}^2 + 2\langle w^{(k-1)}, y_i h_i \rangle + \|h_i\|_{\mathbb{H}}^2 - \left\|w^{(k-1)}\right\|_{\mathbb{H}}^2 \\
&\leqslant \|h_i\|_{\mathbb{H}}^2 ,
\end{aligned}$$

where the inequality follows from $\langle w^{(k-1)}, y_i h_i \rangle \leqslant 0$. Therefore, after m steps,

$$\left\|w^{(m)}\right\| \leqslant \left(\sum_{k=1}^{m} \|h_{i_k}\|_{\mathbb{H}}^2\right)^{\frac{1}{2}} .$$

Combining the above inequality with (D.4) using Cauchy-Schwarz inequality, we obtain

$$m\alpha \leqslant \langle w^{(m)}, w_\alpha \rangle \leqslant \left\|w^{(m)}\right\|_{\mathbb{H}} \|w_\alpha\|_{\mathbb{H}} \leqslant \|w_\alpha\|_{\mathbb{H}} \left(\sum_{k=1}^{m} \|h_{i_k}\|_{\mathbb{H}}^2\right)^{\frac{1}{2}} .$$

Thus, Inequality (D.1) follows from $\sum_{k=1}^{m} \|h_{i_k}\|_{\mathbb{H}}^2 \leqslant mR^2$ and Inequality (D.2), on the other hand, from (D.3). ■

Appendix E

L_p-norm Combinatorial Bound

Here we have gathered the results used in the proof of our extension of the L_2-norm combinatorial bound of Mendelson and Vershynin [61].

Lemma 36 *For all $p \in \mathbb{N}^* \setminus \{1, 2\}$,*

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

Proof By Formula (8.5) in [17, page 119],

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

where $\psi_p(u) = \sum_{j=0}^{p-1} (-1)^j \binom{p}{j+1} (u+1)^j \psi_{(p-1)-j}(u)$ is an Eulerian polynomial in u of degree $p-1$ with $\psi_0(u) = \psi_1(u) = 1$ (see page 116 in [17] for explicit form of this polynomial for smaller values of p). Thus for $u = 2$,

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

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

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

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

$\psi_k(-2) < k^k/2$. Then,

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

We have that

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

Applying it in (E.1), we obtain

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

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

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

Consequently,

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

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

The results demonstrated hereafter are the generalizations of those in [61]. In the following, we denote $K_p = \sum_{k=1}^{\infty} \frac{k^p}{2^k}$ with $p \in \mathbb{N}^* \setminus \{1, 2\}$.

Lemma 37 (After Lemma 5 of [61]) *Let X be a bounded random variable. Let $M_p(X) = (\mathbb{E}|X|^p])^{1/p}$. Then, there exist numbers $a \in \mathbb{R}$ and $\beta \in (0, 1/2]$, such that*

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

or vice versa.

Proof The proof closely follows that of Lemma 5 of [61] where the variance of X is replaced by its higher moments.

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

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

by setting

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

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

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

and

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

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

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

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

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

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

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

and

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

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

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

and

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

Thus, the lemma holds. This proves (E.2). Now, by induction from (E.2) we get that

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

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

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

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

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

By a similar procedure, it can be proved that

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

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

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

Lemma 38 (After Lemma 6 of [61]) *Let $\mathcal{F}$ be a finite class of functions from $\mathcal{T}$ to $[0, M_{\mathcal{F}}]$ with $M_{\mathcal{F}} \in \mathbb{R}_+$ and $|\mathcal{F}| > 1$. Assume that for some $\epsilon \in (0, M_{\mathcal{F}}]$, $\mathcal{F}$ is ϵ-separated in the pseudo-metric $d_{p,\mathbf{t}_n}$. Then there exist $i \in [\![1, n]\!]$, $a \in \mathbb{R}$ and $\beta \in (0, 1/2]$ such that*

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

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

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

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

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

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

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

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

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

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

Thus,

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

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

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

Next, we apply Lemma 37 to the random element f and take into account that

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

and that

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

Then, it follows that

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

and, similarly,

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

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

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

Definition 24 *Let $\mathcal{F}$ be a class of functions on $\mathcal{T}$. A tree $T(\mathcal{F})$ is a finite collection of subsets of $\mathcal{F}$, such that its any two elements are either disjoint or one of them contains the other. A son of $\bar{\mathcal{F}} \in T(\mathcal{F})$ is its maximal (with respect to inclusion) proper subset. An element of $T(\mathcal{F})$ with no sons is called a leaf. Let $\epsilon > 0$. If every $\bar{\mathcal{F}} \in T(\mathcal{F})$ which is not a leaf has exactly two sons $\bar{\mathcal{F}}_+, \bar{\mathcal{F}}_-$ and*

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

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

Proposition 5 (After Proposition 8 in [61]) *Let $\mathcal{F}$ be a finite class of functions from $\mathcal{T}$ to $[0, M_{\mathcal{F}}]$ with $M_{\mathcal{F}} \in \mathbb{R}_+$. Assume that for some $\epsilon \in (0, M_{\mathcal{F}}]$, $\mathcal{F}$ is ϵ-separated in the pseudo-metric $d_{p,\mathbf{t}_n}$. Then, there is a $\epsilon/4(4K_p)^{1/p}$-separating tree of $\mathcal{F}$ with at least $|\mathcal{F}|^{1/2}$ leaves.*

Proof By Lemma 38, $\mathcal{F}$ has two subsets $\mathcal{F}_+$ and $\mathcal{F}_-$ such that

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

which implies

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

The rest of the proof is based on induction on the cardinality of $\mathcal{F}$ and is exactly as in [61], except that the tree is now $\epsilon/4(4K_p)^{1/p}$-separated. ■

Proposition 6 (After Proposition 10 in [61]) *Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ to a finite set B of integers. Let $S \subseteq \mathcal{T}$ and let $v : S \to B$. The number of pairs (S, v) strongly shattered by $\mathcal{F}$ is at least the number of leaves in any 1-separating tree of $\mathcal{F}$.*

Proof The proof follows exactly the one of Proposition 10 in [61], with a few minor technical changes. Let $\bar{\mathcal{F}}$ be a node in a 1-separating tree of $\mathcal{F}$. Let $N(A)$ denote the number of pairs strongly shattered by a set A. For the proof it suffices to show that if $\bar{\mathcal{F}}_+$ and $\bar{\mathcal{F}}_-$ are two sons of $\bar{\mathcal{F}}$, then

$$N(\bar{\mathcal{F}}) \geqslant N(\bar{\mathcal{F}}_+) + N(\bar{\mathcal{F}}_-). \tag{E.3}$$

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

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

It follows that

$$\exists b \in B, \forall (f_+, f_-) \in \left(\bar{\mathcal{F}}_+, \bar{\mathcal{F}}_-\right), \quad \begin{cases} f_+(t_{i_0}) > b \\ f_-(t_{i_0}) < b. \end{cases} \tag{E.4}$$

If a pair is strongly shattered either by $\bar{\mathcal{F}}_+$ or $\bar{\mathcal{F}}_-$, then it is also strongly shattered by $\bar{\mathcal{F}}$. On the other hand, if a pair (S, v) is strongly shattered both by $\bar{\mathcal{F}}_+$ and $\bar{\mathcal{F}}_-$, then $t_{i_0} \notin S$. Otherwise, there would exist $(f'_+, f'_-) \in \left(\bar{\mathcal{F}}_+, \bar{\mathcal{F}}_-\right)$ satisfying $f'_+(t_{i_0}) \leqslant v(t_{i_0}) - 1$ and $f'_-(t_{i_0}) \geqslant v(t_{i_0}) + 1$. Combining it with (E.4) yields a contradiction:

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

Now, consider a pair $(S \cup \{t_{i_0}\}, v')$, where $v'(t_i) = v(t_i)$ for all $t_i \in S$ and $v'(t_{i_0}) = b$. This pair is shattered by $\bar{\mathcal{F}}$, but neither by $\bar{\mathcal{F}}_+$ or $\bar{\mathcal{F}}_-$. As S is shattered both by $\bar{\mathcal{F}}_+$ and $\bar{\mathcal{F}}_-$, then from (E.4) it follows that,

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

similarly,

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

It proves the claim that $\bar{\mathcal{F}}$ shatters the pair $(S \cup \{t_{i_0}\}, v')$. Therefore, in both cases we get (E.3). ■

The next result is obtained by combining Propositions 5 and 6.

Corollary 7 (After Corollary 11 in [61]) *Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ to a finite set B of integers. Let $S \subseteq \mathcal{T}$ and let $v : S \to B$. If $\mathcal{F}$ is $4(4K_p)^{1/p}$-separated in the pseudo-metric $d_{p,\mathbf{t}_n}$, then it strongly shatters at least $|\mathcal{F}|^{1/2}$ pairs (S, v) .*

Proposition 7 (After Proposition 12 in [61]) *Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ to $[\![0, b]\!]$. Let $d_s = S\text{-}dim(\mathcal{F})$. Assume $\mathcal{F}$ is $4(4K_p)^{1/p}$-separated in the pseudo-metric $d_{p,\mathbf{t}_n}$. Then for any $d \geqslant d_s$,*

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

Proof By Corollary 7, $\mathcal{F}$ strongly shatters at least $|\mathcal{F}|^{1/2}$ pairs (S, v). On the other hand, the total number of such pairs for which the cardinality of S is at most d_s is bounded above by

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

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

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

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

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

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

Appendix F

Rademacher Complexity Bounds for Linear Classifiers

The proof of Theorem 19 makes use of a result concerning the strong convexity and strong smoothness due to Kakade and coauthors [42]. To introduce it, we need to give several definitions. Let $\mathbb{R}^* = \mathbb{R} \cup \{\infty\}$. Let the metric space $\mathcal{T}$ be equipped with an inner product $\langle \cdot, \cdot \rangle$ and let $f : \mathcal{T} \to \mathbb{R}^*$ be a convex function. Let $\beta > 0$. f is said to be β-strongly convex with respect to the $\|\cdot\|_{\mathcal{T}}$-norm on $\mathcal{T}$ if for all $t, t' \in \mathcal{T}$ and for all $\alpha \in (0, 1)$,

$$f(\alpha t + (1-\alpha)t') \leqslant \alpha f(t) + (1-\alpha) f(t') - \frac{1}{2}\beta\alpha(1-\alpha)\left\|t - t'\right\|_{\mathcal{T}}.$$

The Fenchel conjugate $f^* : \mathcal{T} \to \mathbb{R}$ of f is defined as

$$\forall t' \in \mathcal{T}, \quad f^*(t') = \sup_t \left(\langle t, t' \rangle - f(t)\right).$$

The dual $\|\cdot\|_*$ of $\|\cdot\|_{\mathcal{T}}$ is defined as

$$\forall t' \in \mathcal{T}, \quad \left\|t'\right\|_* = \sup_{\|t\| \leqslant 1} \langle t, t' \rangle.$$

Lemma 39 (After Corollary 4 in [42]) *Let $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ be a Hilbert space. Let $f : \mathcal{H} \to \mathbb{R}$ be β-strongly convex with respect to the $\|\cdot\|_{\mathcal{H}}$-norm. Denote by f^* the Fenchel conjugate of f, and assume that $f^*(\mathbf{0}) = 0$. Then, for any $(\mathbf{v}_i)_{1 \leqslant i \leqslant n} \in \mathcal{H}^n$ and for any $\mathbf{r} \in \mathcal{H}$,*

$$\sum_{i=1}^{n} \langle \mathbf{v}_i, \mathbf{r} \rangle \leqslant f(\mathbf{r}) + \sum_{i=1}^{n} \langle \nabla f^*\left(\mathbf{v}_{1:i-1}\right), \mathbf{v}_i \rangle + \frac{1}{2\beta} \sum_{i=1}^{n} \|\mathbf{v}_i\|_*^2,$$

where $\mathbf{v}_{1:l} = \sum_{i=1}^{l} \mathbf{v}_i$, $\|\cdot\|_*$ is the dual norm of $\|\cdot\|_{\mathcal{H}}$ and ∇f is the gradient of f.

Proof of Theorem 19 Let $p \leqslant q \leqslant 2$ and let $f_q(\cdot) = \frac{1}{2}\|\cdot\|^2_{\mathcal{H}_{\kappa,q}}$. Let q^* denote the Hölder conjugate of q:

$$\frac{1}{q} + \frac{1}{q^*} = 1.$$

Then, the norm $\|\cdot\|_{\mathcal{H}_\kappa,q^*}$ is the dual of $\|\cdot\|_{\mathcal{H}_\kappa,q}$, and the function f_q is $\beta = \frac{1}{q^*}$-strongly convex with respect to $\|\cdot\|_{\mathcal{H}_\kappa,q}$. Now, in Inequality (3.7) in Chapter 3, instantiate $\mathcal{G}$ with $\mathcal{B}_{q,\Lambda}$. For all $i \in [\![1, m]\!]$, let $\mathbf{v}_i = \left(o_{C(i-1)+k}\kappa_{x_i}\right)_{1\leqslant k\leqslant C}$. Fix $\lambda > 0$. Then, applying Lemma 39 to the right-hand side of Inequality (3.7),

$$\begin{aligned}
\lambda \sup_{h\in\mathcal{B}(q,\Lambda)} \sum_{i=1}^{m}\sum_{k=1}^{C} o_{C(i-1)+k} h_k(x_i) &= \sup_{h\in\mathcal{B}(q,\Lambda)} \sum_{i=1}^{m}\sum_{k=1}^{C} \langle h_k, \lambda o_{C(i-1)+k}\kappa_{x_i}\rangle \\
&= \sup_{h\in\mathcal{B}(q,\Lambda)} \sum_{i=1}^{n} \langle \lambda \mathbf{v}_i, h\rangle \\
&\leqslant \sup_{h\in\mathcal{B}(q,\Lambda)} f_q(h) + \sum_{i=1}^{m} \langle \nabla f^*\left(\mathbf{v}_{1:i-1}\right), \lambda\mathbf{v}_i\rangle \\
&\quad + \frac{\lambda^2 q^*}{2}\sum_{i=1}^{m} \|\mathbf{v}_i\|^2_{\mathcal{H}_\kappa,q^*}.
\end{aligned}$$

Now, by the assumption,

$$\forall h \in \mathcal{B}_{q,\Lambda}, \quad f_q(h) \leqslant \frac{\Lambda^2}{2}.$$

Then, based on this, as well as the fact that $\boldsymbol{o}_{Cm}$ is an orthogaussian sequence and thus that

$$\mathbb{E}_{\boldsymbol{o}_{Cm}} \langle \nabla f^*\left(\mathbf{v}_{1:i-1}\right), \lambda\mathbf{v}_i\rangle = 0,$$

we have

$$\mathbb{E}_{\boldsymbol{o}_{Cm}} \sup_{h\in\mathcal{B}_{q,\Lambda}} \sum_{i=1}^{m}\sum_{k=1}^{C} o_{C(i-1)+k} h_k(x_i) \leqslant \frac{\Lambda^2}{\lambda} + \frac{\lambda q^*}{2}\sum_{i=1}^{m} \mathbb{E}_{\boldsymbol{o}_{Cm}} \|\mathbf{v}_i\|^2_{\mathcal{H}_\kappa,q^*}.$$

Minimizing the right-hand side with respect to λ yields

$$\mathbb{E}_{\boldsymbol{o}_{Cm}} \sup_{h\in\mathcal{B}_{q,\Lambda}} \sum_{i=1}^{m}\sum_{k=1}^{C} o_{C(i-1)+k} h_k(x_i) \leqslant \Lambda\sqrt{q^*\sum_{i=1}^{m} \mathbb{E}_{\boldsymbol{o}_{Cm}} \|\mathbf{v}_i\|^2_{\mathcal{H}_\kappa,q^*}}. \tag{F.1}$$

The goal now is to control the term $\sum_{i=1}^{m} \mathbb{E}_{\boldsymbol{o}_{Cm}} \|\mathbf{v}_i\|^2_{\mathcal{H}_\kappa,q^*}$. It follows that

$$\begin{aligned}
\sum_{i=1}^{m} \mathbb{E}_{\boldsymbol{o}_{Cm}} \|\mathbf{v}_i\|^2_{\mathcal{H}_\kappa,q*} &= \sum_{i=1}^{m} \mathbb{E}_{\boldsymbol{o}_{Cm}} \left[\sum_{k=1}^{C} \left\|o_{C(i-1)+k}\kappa_{x_i}\right\|^{q^*}_{\mathcal{H}_\kappa}\right]^{\frac{2}{q^*}} \\
&= \sum_{i=1}^{m} \kappa(x_i, x_i)\, \mathbb{E}_{\boldsymbol{o}_{Cn}} \left[\sum_{k=1}^{C} \left|o_{C(i-1)+k}\right|^{q^*}\right]^{\frac{2}{q^*}} \\
&= \mathbb{E}_{\tilde{\boldsymbol{o}}_C} \left[\sum_{k=1}^{C} |\tilde{o}_k|^{q^*}\right]^{\frac{2}{q^*}} \sum_{i=1}^{m} \kappa(x_i, x_i),
\end{aligned} \tag{F.2}$$

where the last equality is obtained based on the fact that

$$\forall i \in [\![1, m]\!],\ \mathbb{E}_{\boldsymbol{o}_{Cm}} \left[\sum_{k=1}^{C} \left| o_{C(i-1)+k} \right| \right] = \mathbb{E}_{\tilde{\boldsymbol{o}}_C} \left[\sum_{k=1}^{C} |\tilde{o}_k| \right].$$

Next, by Jensen's Inequality,

$$\mathbb{E}_{\tilde{\boldsymbol{o}}_C} \left[\sum_{k=1}^{C} |\tilde{o}_k|^{q^*} \right]^{\frac{2}{q^*}} \leqslant \left[\sum_{k=1}^{C} \mathbb{E}_{\tilde{\boldsymbol{o}}_C} |\tilde{o}_k|^{q^*} \right]^{\frac{2}{q^*}} \leqslant \left(C M_{q^*} \right)^{\frac{2}{q^*}}, \tag{F.3}$$

where $M_{q^*} = \mathbb{E}\,|\tilde{o}_1|^{q^*}$. Substituting the last result in (F.2) and using the fact that $\sum_{i=1}^{m} k\left(x_i, x_i\right) \leqslant m\Lambda_{\mathcal{X}}^2$, give

$$\sum_{i=1}^{m} \mathbb{E}_{\boldsymbol{o}_{Cm}} \left\| \mathbf{v}_i \right\|_{\mathcal{H},q*}^2 \leqslant m\Lambda_{\mathcal{X}}^2 \left(C M_{q^*} \right)^{\frac{2}{q^*}}. \tag{F.4}$$

Thus, from (F.1)-(F.4) we get:

$$\hat{R}_m \left(\mathcal{F}_{\mathcal{B}_{q,\Lambda,\gamma}} \right) \leqslant \frac{\Lambda\Lambda_{\mathcal{X}} \sqrt{q^*} \left(C M_{q^*} \right)^{\frac{1}{q^*}}}{\sqrt{m}} \sqrt{\frac{\pi}{2}}. \tag{F.5}$$

Now, the goal is to minimize the right-hand side of (F.5) with respect to q^*. Note that while $(M_{q^*})^{\frac{1}{q^*}}$ is an increasing function of q^*, $\sqrt{q^*} C^{\frac{1}{q^*}}$ attains its minimum at $q^* = 2\ln C$. With this value of q^*, we have

$$(4\ln C)^{\frac{1}{2\ln C}} \leqslant 2 \quad \text{and} \quad C^{\frac{1}{2\ln C}} \leqslant \sqrt{e}.$$

Then, using Lemma 32 in Appendix A,

$$(M_{2\ln C})^{\frac{1}{2\ln C}} \leqslant (4\ln C)^{\frac{1}{2}+\frac{1}{2\ln C}} \leqslant (4\ln C)^{\frac{1}{2}} (4\ln C)^{\frac{1}{2\ln C}} \leqslant 4\sqrt{\ln C}.$$

Substituting these values in Inequality (F.5), we have that when $p \leqslant q \leqslant \dfrac{2\ln C}{2\ln C - 1}$,

$$\hat{R}_m \left(\mathcal{F}_{\mathcal{B}_{q,\Lambda,\gamma}} \right) \leqslant 4\sqrt{e\pi}\Lambda\Lambda_{\mathcal{X}} \frac{\ln C}{\sqrt{m}}. \tag{F.6}$$

On the other hand, when $q^* = 2$ then according to Lemma 32,

$$(M_2)^{\frac{1}{2}} \leqslant (2 \cdot 2)^{2 \cdot \frac{1}{2}} = 4$$

and thus for $p = q = 2$,

$$\hat{R}_m \left(\mathcal{F}_{\mathcal{B}_{q,\Lambda,\gamma}} \right) \leqslant 4\sqrt{\pi}\Lambda\Lambda_{\mathcal{X}} \sqrt{\frac{C}{m}}.$$

To conclude the proof, note that based on the constraint $p \leqslant q \leqslant 2$, we have $f_p(h) \geqslant f_q(h)$ for all $h \in \mathcal{H}^C$ and thus $\mathcal{B}_{p,\Lambda} \subseteq \mathcal{B}_{q,\Lambda}$. This implies $\hat{R}_m \left(\mathcal{F}_{\mathcal{B}_{p,\Lambda,\gamma}} \right) \leqslant \hat{R}_m \left(\mathcal{F}_{\mathcal{B}_{q,\Lambda,\gamma}} \right)$.

Proof of Theorem 20 The proof follows that of Theorem 19, where we replace all orthogaussian sequences with Rademacher ones. Then, based on the fact that the expectation of the absolute value of the Rademacher variable is one, Inequality (F.3) is replaced by

$$\mathbb{E}_{\tilde{\boldsymbol{\sigma}}_C}\left[\sum_{k=1}^{C}|\tilde{\sigma}_k|^{q^*}\right]^{\frac{2}{q^*}} \leqslant \left[\sum_{k=1}^{C}\mathbb{E}_{\tilde{\boldsymbol{\sigma}}_C}|\tilde{\sigma}_k|^{q^*}\right]^{\frac{2}{q^*}} \leqslant C^{\frac{2}{q^*}},$$

where $\tilde{\boldsymbol{\sigma}}_C = (\tilde{\sigma}_k)_{1\leqslant k\leqslant C}$ is the Rademacher sequence. Consequently, Inequality (F.5) is replaced by

$$\hat{R}_m\left(\mathcal{F}_{\mathcal{B}_{q,\Lambda,\gamma}}\right) \leqslant \frac{\Lambda\Lambda_{\mathcal{X}}\sqrt{2q^*}C^{\frac{1}{q^*}}}{\sqrt{m}},$$

where the gain is by a factor $M_{q^*}^{\frac{1}{q^*}}\sqrt{\frac{\pi}{4}}$. In the above inequality, substituting the value $q^* = 2\ln C$ for $p \leqslant 1$ and using the fact that $C^{\frac{1}{2\ln C}} \leqslant \sqrt{e}$ gives the first bound. For second one, we set $q^* = 2$, and the claim follows.

Appendix G

Technical Results

This Appendix gathers several technical results used in the main text.

Lemma 40 (Lemma 3.14 in [3]) *For any $\alpha > 0$ and for any $x > 0$,*

$$\log_a x \leqslant \log_a \left(\frac{1}{\alpha a \ln a} \right) + \alpha x.$$

Lemma 41 (Appears as a partial result in the proof of Theorem 17 in [7]) *For any $a, b \geqslant 1$ and for any $x \geqslant 1$,*

$$a \log^2 (bx) \leqslant \frac{x}{2} + 16a \log^2 (16ab),$$

where $\log(\cdot)$ *stands for the logarithm of any base.*

Lemma 42 *Let M be a positive integer. The number s of non-negative integer solutions of the inequality*

$$\sum_{j=1}^{n} x_j \leqslant M$$

can be bounded as

$$s \leqslant (n+1)^M.$$

Proof To estimate s we appeal to the classical formula of stars and bars [29]:

$$s = \sum_{l=0}^{M} \binom{l+n-1}{n-1}.$$

The right-hand side can be upper bounded as follows. For $l > 0$, we have

$$\binom{l+n-1}{n-1} = \frac{(l+n-1)!}{(n-1)!l!} = \frac{\prod_{i=1}^{l}(i+n-1)}{l!}.$$

Since for all $i \leqslant l$, $n + i - 1 \leqslant ni$, we have

$$\binom{l+n-1}{n-1} \leqslant \frac{n^l \prod_{i=1}^{l} i}{l!} = n^l.$$

Thus,

$$s \leqslant 1 + \sum_{l=1}^{M} n^l \leqslant (n+1)^M.$$

■

Bibliography

[1] 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.

[2] B. Samy anManikd D. Krzysztof, J. Thorsten, K. Marius, and V. Extreme Classification (Dagstuhl Seminar 18291). *Dagstuhl Reports*, 8(7):62–80, 2019.

[3] M. Anthony. *Uniform convergence and learnability.* PhD thesis, London School of Economics and Political Science (United Kingdom), 1991.

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

[5] Rohit Babbar. *Machine Learning Strategies for Large-scale Taxonomies.* PhD thesis, Université de Grenoble, 2014.

[6] K.I. Babenko. On the entropy of a class of analytic functions. *Nauchn. Dokl. Vysshei Shkoly Ser. Fiz. Mat. Nauk*, (2):9–16, 1958.

[7] 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.

[8] P.L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48(1):85–113, 2002.

[9] P.L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[10] P.L. Bartlett, D.J. Foster, and M. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017.

[11] P.L. Bartlett, S.R. Kulkarni, and S.E. Posner. Covering numbers for real-valued function classes. *IEEE Transactions on Information Theory*, 43(5):1721–1724, 1997.

[12] P.L. Bartlett and P.M. Long. More theorems about scale-sensitive dimensions and learning. In *COLT'95*, pages 392–401, 1995.

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

[14] 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.

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

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

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

[18] S. Bernstein. *The Theory of Probabilities*. Gastehizdat Publishing House, Moscow, 1946.

[19] L. Breiman, J.H. Friedman, R. Olshen, and C.J. Stone. Classification and regression trees. 1984.

[20] C. Cortes and V.N. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[21] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New York, 1996.

[22] Ü. 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.

[23] 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.

[24] R. M. Dudley. A course on empirical processes. 1982.

[25] R.M. Dudley. Universal Donsker classes and metric entropy. *The Annals of Probability*, 15(4):1306–1326, 1987.

[26] R.M. Dudley. *Uniform central limit theorems*. Cambridge University Press, 1999.

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

[28] M. Farooq and I. Steinwart. Learning rates for kernel-based expectile regression. *arXiv preprint arXiv:1702.07552*, 2017.

[29] W. Feller. *An introduction to probability theory and its applications*, volume 2. John Wiley & Sons, 2008.

[30] E. Giné and J. Zinn. Some limit theorems for empirical processes. *The Annals of Probability*, 12(4):929–989, 1984.

[31] V. Goodman. Some probability and entropy estimates for Gaussian measures. In *Probability in Banach Spaces 6*, pages 150–156. Springer, 1990.

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

[33] A.J. Grove, N. Littlestone, and D. Schuurmans. General convergence results for linear discriminant updates. *Machine Learning*, 43(3):173–210, 2001.

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

[35] Y. Guermeur. A generic model of multi-class support vector machine. *International Journal of Intelligent Information and Database Systems*, 6(6):555–577, 2012.

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

[37] Y. Guermeur. Combinatorial and Structural Results for gamma-Psi-dimensions. *ArXiv e-prints*, 2018.

[38] Y. Guo, P.L. Bartlett, J. Shawe-Taylor, and R.C. Williamson. Covering numbers for support vector machines. *IEEE Transactions on Information Theory*, 48(1):239–250, 2002.

[39] Anupam Gupta, Robert Krauthgamer, and James R Lee. Bounded geometries, fractals, and low-distortion embeddings. page 534. IEEE, 2003.

[40] D. Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995.

[41] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[42] S.M. Kakade, S. Shalev-Shwartz, and A. Tewari. Regularization techniques for learning with matrices. *Journal of Machine Learning Research*, 13(Jun):1865–1890, 2012.

[43] 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.

[44] A.N. Kolmogorov. Certain asymptotic characteristics of completely bounded metric spaces. *Doklady Akademii Nauk SSSR*, 108(3):385–388, 1956.

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

[46] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

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

[48] V.I Koltchinskii. On the central limit theorem for empirical measures. *Theory of probability and mathematical statistics*, 24:71–82, 1981.

[49] A. Kontorovich and R. Weiss. Maximum margin muliclass nearest neighbors. In *ICML'14*, 2014.

[50] A. Krizhevsky, I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[51] J. Kuelbs and W.V. Li. Metric entropy and the small ball problem for gaussian measures. *Journal of Functional Analysis*, 116(1):133–157, 1993.

[52] T. Kühn. Covering numbers of Gaussian reproducing kernel Hilbert spaces. *Journal of Complexity*, 27(5):489–499, 2011.

[53] V. Kuznetsov, M. Mohri, and U. Syed. Multi-class deep boosting. In *NIPS 27*, pages 2501–2509, 2014.

[54] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, Berlin, 1991.

[55] Y. Lei, Ü. 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.

[56] P. Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse* 6^e *série*, 9(2):245–303, 2000.

[57] A. Maurer. A vector-contraction inequality for Rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pages 3–17. Springer, 2016.

[58] C. McDiarmid. On the method of bounded differences. In J. Siemons, editor, *Surveys in Combinatorics*, volume 141 of *London Mathematical Society Lecture Note Series*, pages 148–188. Cambridge University Press, Cambridge, 1989.

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

[60] S. Mendelson. A few notes on statistical learning theory. Technical report, 2003.

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

[62] M. Minsky and S. Papert. The perceptron: Principles of computational geometry. *MIT press. McCulloch, WS and Pitts, W.,(1943) A Logical Calculus of the Ideas Imminent in Nervous Activity, Bulletin of Mathematical Biophysics*, 5:115–133, 1969.

[63] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. The MIT Press, Cambridge, MA, 2012.

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

[65] B.K. Natarajan. On learning sets and functions. *Machine Learning*, 4(1):67–97, 1989.

[66] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 5947–5956, 2017.

[67] B. Neyshabur, S. Bhojanapalli, and N. Srebro. A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017.

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

[69] D. Pollard. Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics*. JSTOR, 1990.

[70] T. Popoviciu. Sur les équations algébriques ayant toutes leurs racines réelles. *Mathematica*, 9:129–145, 1935.

[71] J.R. Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[72] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

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

[74] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

[75] S. Shelah. A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41(1):247–261, 1972.

[76] N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. In *Advances in neural information processing systems*, pages 2199–2207, 2010.

[77] I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[78] I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. *The Annals of Statistics*, 35(2):575–607, 2007.

[79] S. Szarek. On the best constants in the Khinchin inequality. *Studia Mathematica*, 2(58):197–208, 1976.

[80] M. Talagrand. The Glivenko-Cantelli problem, ten years later. *Journal of Theoretical Probability*, 9(2):371–384, 1996.

[81] M. Talagrand. New concentration inequalities in product spaces. *Inventiones mathematicae*, 126(3):505–563, 1996.

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

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

[84] V.M. Tikhomirov. Certain asymptotic characteristics of completely bounded metric spaces. *Doklady Akademii Nauk SSSR*, 117:191–194, 1957.

[85] A.W. van der Vaart and J.H. van Zanten. Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *The Annals of Statistics*, pages 2655–2675, 2009.

[86] 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.

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

[88] 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.

[89] A.R. Vitale. Some comparisons for Gaussian processes. *Proceedings of the American Mathematical Society*, pages 3043–3046, 2000.

[90] A.G. Vitushkin. The absolute ε-entropy of metric spaces. In *Doklady Akademii Nauk*, volume 117, pages 745–747. Russian Academy of Sciences, 1957.

[91] R.C. Williamson, A.J. Smola, and B. Schölkopf. Generalization performance of regularization networks and support vector machines *via* entropy numbers of compact operators. *IEEE Transactions on Information Theory*, 47(6):2516–2532, 2001.

[92] T. Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550, 2002.

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

[94] Y. Zhang, J. Lee, M. Wainwright, and M.I. Jordan. On the learnability of fully-connected neural networks. In *Artificial Intelligence and Statistics*, pages 83–91, 2017.

[95] D. Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.

Résumé

Cette thèse porte sur la théorie de la discrimination multi-classe à marge. Elle a pour cadre la théorie statistique de l'apprentissage de Vapnik et Chervonenkis. L'objectif est d'établir des bornes de généralisation possédant une dépendances explicite au nombre C de catégories, à la taille m de l'échantillon et au paramètre de marge γ, lorsque la fonction de perte considérée est une fonction de perte à marge possédant la propriété d'être lipschitzienne. La borne de généralisation repose sur la performance empirique du classifieur ainsi que sur sa "capacité". Dans cette thèse, les mesures de capacité considérées sont les suivantes : la complexité de Rademacher, les nombres de recouvrement et la dimension fat-shattering. Nos principales contributions sont obtenues sous l'hypothèse que les classes de fonctions composantes calculées par le classifieur ont des dimensions fat-shattering polynomiales et que les fonctions composantes sont indépendantes.

Dans le contexte du schéma de calcul introduit par Mendelson, qui repose sur les relations entre les mesures de capacité évoquées plus haut, nous étudions l'impact que la décomposition au niveau de l'une de ces mesures de capacité a sur les dépendances (de la borne de généralisation) à C, m et γ. En particulier, nous démontrons que la dépendance à C peut être considérablement améliorée par rapport à l'état de l'art si la décomposition est reportée au niveau du nombre de recouvrement ou de la dimension fat-shattering. Ce changement peut affecter négativement le taux de convergence (dépendance à m), ce qui souligne le fait que l'optimisation par rapport aux trois paramètres fondamentaux se traduit par la recherche d'un compromis.

Mots-clés: apprentissage, théorie de l'apprentissage, discrimination multi-classe, risques garantis, classifieurs à marge

Abstract

This thesis deals with the theory of margin multi-category classification, and is based on the statistical learning theory founded by Vapnik and Chervonenkis. We are interested in deriving generalization bounds with explicit dependencies on the number C of categories, the sample size m and the margin parameter γ, when the loss function considered is a Lipschitz continuous margin loss function. Generalization bounds rely on the empirical performance of the classifier as well as its "capacity". In this work, the following scale-sensitive capacity measures are considered: the Rademacher complexity, the covering numbers and the fat-shattering dimension. Our main contributions are obtained under the assumption that the classes of component functions implemented by a classifier have polynomially growing fat-shattering dimensions and that the component functions are independent.

In the context of the pathway of Mendelson, which relates the Rademacher complexity to the covering numbers and the latter to the fat-shattering dimension, we study the impact that decomposing at the level of one of these capacity measures has on the dependencies on C, m and γ. In particular, we demonstrate that the dependency on C can be substantially improved over the state of the art if the decomposition is postponed to the level of the metric entropy or the fat-shattering dimension. On the other hand, this impacts negatively the rate of convergence (dependency on m), an indication of the fact that optimizing the dependencies on the three basic parameters amounts to looking for a trade-off.

Keywords: statistical learning theory, multi-category classification, risk bounds, margin classifiers