Partial Information and Distribution-Dependence in Supervised Learning Models

Thesis submitted for the degree of "Doctor of Philosophy"

by

Sivan Sabato

Submitted to the Senate of the Hebrew University

July 2012

This work was carried out under the supervision of

Prof. Naftali Tishby

To Moshe

Acknowledgments

First and foremost, I wish to express my deep gratitude to my advisor, Naftali Tishby. Tali has patiently guided me in my research, while also supporting me in developing my independence. He has always been there to ask the right question at the right time, and his grand view of the field has influenced me greatly. Part of this thesis is the result of collaboration with Nati Srebro. I have learned a great deal from Nati, and I am grateful to have had the chance to collaborate with him.

I have had invaluable discussions with Nati Linial and with Boaz Nadler, which have helped me improve my work. During my PhD I have also worked with Shai Ben-David, whose crisp sense of abstraction has been an inspiration to me. I have also had the pleasure of collaborating with my fellow students at the Hebrew University, Ohad Shamir, Alon Gonen, Amit Daniely, Dvir Aran, Alon Zweig and Amit Gruber.

My first serious exposure to the field of Machine Learning has been at IBM Research. I am indebted to Shai Fine, whose conviction and enthusiasm have lead me to join the Machine Learning research team at IBM, during an exciting time of growth and advancement. My experience at IBM has lead me to realize that my future was in this field.

A special thanks is extended to Shai Shalev-Shwartz, who I had been lucky to collaborate with even before my PhD, and have continued to work with all along the way. More than a gifted researcher that I have learned a lot from, Shai has also been a great friend, supporting and encouraging me since the very beginning.

I have been very fortunate to have been an Adams fellow during my PhD studies. The support of the Adams foundation has allowed me to focus on my research. I am grateful to Marcel Adams, for his contribution to science in Israel. I would also like to thank the people at the Israel Academy of Sciences, and especially Batsheva Shor, for their organization and support.

I have had a great time at the learning lab at the Hebrew University, thanks to the pleasant atmosphere of cooperation and friendship between everyone—students, faculty and administrative staff alike. Thank you all for making this such a wonderful experience.

None of this would have come to pass without the constant support and love of my family. I am grateful to my parents, my sister and my brother for their endless faith in me. Moshe, thank you for being there for me every tiny step of the way. Tamar, thank you for being you.

Abstract

In this thesis we study two important supervised learning settings: linear classifiers with a margin, and Multiple-Instance Learning, and provide novel results concerning the ability to learn in each of these settings.

In supervised learning, the goal is to learn to classify objects into one of several classes, using only examples of objects, along with the class that they belong to (also termed their *label*). We focus on binary supervised learning, in which each object should be classified into one of two classes. As an example, consider the task of predicting whether a patient will present with diabetes, based on the patient's blood test results. In this example, one class represents patients who will present with diabetes and the other class represents patients who will not present with diabetes. The learner is given a set of examples, where each example is constituted of the blood test results of a patient, along with information on whether this patient has presented with diabetes or not. We term this set of examples the *training set*, or the *training sample*. The training set is used by the learner to infer a *classification rule*, which can be used to predict whether a new patient will present with diabetes, based on this patient's blood test results. The goal of the learner is to find a classification rule which is as accurate as possible in its predictions.

An important measure of the effectiveness of learning is how many labeled examples are needed in order to achieve a certain degree of classification accuracy. The *sample complexity* of a learning problem is the size of a training set required to guarantee a given accuracy on this problem. Equivalently, it is the accuracy that can be guaranteed for the learner, given the size of the training sample. We distinguish between the sample complexity, which is a statistical measure of the difficulty of learning, and computational complexity, which measures the amount of computation required to implement a learning strategy.

The "No free lunch" theorem for supervised learning [Wolpert and Macready, 1997] shows that no single supervised learning algorithm can provide a high-accuracy classification rule for all learning problems using the same sample size. In other words, the sample complexity of supervised learning without additional assumptions is unbounded. It follows that in order to have guarantees on learning, we need to consider more restricted classes of learning problems.

Most commonly, we restrict the set of learning problems that we consider by introducing the notion of a *hypothesis class*. This is a set of classification rules to which we wish to compare the result of our learning algorithm. In a specific learning context, the hypothesis class can represent our beliefs on the true nature of the classification rule for the problem. For instance, if we believe that diabetes can be identified via a boolean function using at most two values in a patient's blood test results, we can use the hypothesis class consisting only of such boolean functions. If our learning problem can indeed be classified with maximum accuracy using one of the classification rules in our hypothesis class, then we say that the problem is *realizable*. In this case, we can hope that our learning algorithm will achieve high accuracy, in absolute terms, when given enough labeled examples. If the problem is not realizable, then we say that we are in the *agnostic* setting. In this case, we hope that our learning algorithm will achieve a small *relative* accuracy. That is, we hope that its classification accuracy will be close to that of the best classifier in our hypothesis class.

The sample complexity of learning relative to a specific hypothesis class strongly depends on its *complexity*. Loosely speaking, the complexity of a class is related to the number of mappings between objects and labels that it allows. There are several popular complexity measures for hypothesis classes, which can be used to derive sample complexity guarantees.

Usually, a sample complexity upper bound is derived for a specific hypothesis class and for a large class of distributions. For instance, the class of *linear classifiers* has upper bounds that depend on the dimension of the input data points, and other upper bounds that depend on their norm. Such upper bounds can be useful for understanding the positive aspects of a learning rule. But it is difficult to understand the deficiencies of a learning rule, or to compare between different rules, based on upper bounds alone. This is because it is possible, and is often the case, that the true sample complexity for a given distribution is much lower than the bound.

Some sample complexity upper bounds are known to be tight or to have an almost-matching lower bound. This means that there exists some distribution in the class covered by the upper bound that actually requires that many examples in order to learn with high accuracy. These results show that there cannot be a better upper bound that holds for the entire class of distributions that the upper bound covers. But they do not imply that the upper bound characterizes the true sample complexity for any *specific* distribution in the class, except for the ones for which the upper bound is indeed tight. For instance, in the case of linear classifiers, although the sample-complexity upper bound that depends on the norm is tight, if the distribution is supported by a low-dimensional sub-space, then the true number of examples required to reach high accuracy is much smaller.

In the first part of this thesis, our goal is to identify a simple quantity, which is a function of the distribution, that *does* precisely characterize the sample complexity of learning this distribution under a specific learning rule. We focus on the popular rule of Margin Error Minimization (MEM), and on the class of linear classifiers. We present a new quantity, termed the *margin-adapted dimension*, and use it to provide a tighter distribution-dependent upper bound, and a matching distribution-dependent lower bound, for MEM algorithms for linear classifiers. The upper bound is universal, and the lower bound holds for a rich class of distributions.

The margin-adapted dimension refines both the dimension and the average norm of the data distribution, and can be easily calculated from the covariance matrix and the mean of the distribution. Our tight characterization, and in particular the distribution-specific lower bound on the sample complexity that we establish, can be used to compare large-margin (L_2 regularized) learning to other learning rules. We provide two such examples: we use our lower bound to rigorously establish a sample complexity gap between L_1 and L_2 regularization previously studied in Ng [2004], and to show a large gap between discriminative and generative learning on a Gaussian-mixture distribution.

Our lower bound hinges on several new results:

- We show that for a convex hypothesis class, fat-shattering is equivalent to shattering with exact margins.
- We link the fat-shattering of a set of vectors with the eigenvalues of the dot-product matrix (the Gram matrix) of the vectors in the set.
- We relate fat-shattering to hardness of learning using MEM.
- We provide a new lower bound for the smallest eigenvalue of a random Gram matrix generated by sub-Gaussian variables, thus extending previous results in analysis of random matrices.

As mentioned above, complexity measures of hypothesis classes are typically analyzed on a case-by-case basis. For instance, the complexity of the class of linear classifiers has been analyzed as a function of parameters such as the dimension of the ambient space and the maximal norm of the separator. In the second part of this thesis, we consider the useful setting of *Multiple Instance Learning*, and propose a generic analysis for this setting, that holds across many different hypothesis classes.

Multiple-Instance Learning (MIL), first introduced in Dietterich et al. [1997], is a special type of a supervised classification problem. As in classical supervised classification, in MIL the learner receives a sample of labeled examples drawn i.i.d. from an arbitrary and unknown distribution, and its objective is to discover a classification rule with a small expected classification error over the same distribution. In MIL additional structure is assumed, whereby the examples are received as *bags* of *instances*, such that each bag is composed of several instances. It is assumed that each instance has a true label, however the learner only observes the labels of the bags. The label of each bag is determined by the hidden labels of the instances in the bag, via some function which is known a-priori. Classical works on MIL assume that the function is the Boolean OR. In this work we consider a more general setting which allows other functions as well.

MIL has been used in numerous applications. In Dietterich et al. [1997] the drug design application motivates this setting. In this application, the goal is to predict which molecules would bind to a specific binding site. Each molecule has several possible conformations (shapes) it can take. If at least one of the conformations binds to the binding site, then the molecule is labeled positive. However, it is not possible to experimentally identify which conformation was the successful one. Thus, a molecule can be thought of as a bag of conformations, where each conformation is an instance in the bag representing the molecule. This application employs the hypothesis class of Axis Parallel Rectangles (APRs), and has made APRs the hypothesis class of choice in several theoretical works. There are many other applications for MIL, including image classification [Maron and Ratan, 1998], web index page recommendation [Zhou et al., 2005] and text categorization [Andrews, 2007].

We propose a formal framework for generalized MIL, which allows analyzing any MIL problem as a function of the underlying hypothesis class: This is the hypothesis class of the possible mappings from single instances to labels. In addition, the analysis depends on the function determining the bag labels based on the instance labels. We provide a generic analysis that bounds the complexity of learning a MIL problem based on the complexity of learning the underlying hypothesis class.

The generic approach has the advantage that it automatically extends all knowledge and methods that apply to non-MIL problems into knowledge and methods that apply to MIL, without requiring specialized analysis for each specific MIL problem. Our results are thus applicable to diverse hypothesis classes and bag labeling functions. Moreover, the generic approach allows a better theoretical understanding of the relationship, in general, between regular learning and Multi-Instance Learning with the same hypothesis class.

Our sample complexity analysis shows that for binary hypotheses and thresholded real-valued hypotheses, the distribution-free sample complexity for generalized MIL grows only logarithmically with the maximal bag size. We also provide poly-logarithmic sample complexity bounds for the case of margin learning. We further provide distribution-dependent sample complexity bounds for more general loss functions. These bound are useful when only the average bag size is bounded. The results imply generalization bounds for previously proposed algorithms for MIL. Addressing the computational feasibility of MIL, we provide a new learning algorithm with provable guarantees for a class of bag-labeling functions that includes the Boolean OR as a special case. Given a non-MIL learning algorithm for the desired hypothesis class, which can handle one-sided errors, we improperly learn MIL with the same hypothesis class. The construction is simple to implement, and provides a computationally efficient PAC-learner for MIL, with only a polynomial dependence of the run time on the bag size. We further show a setting in which MIL can be used to improve the sample complexity of non-MIL learning, by constructing artificial bags. We propose an approach for implementing this paradigm in practice.

Contents

Chapter 1

Introduction

In this thesis we study two important supervised learning settings: linear classifiers with a margin, and Multiple-Instance Learning, and provide novel results concerning the ability to learn in each of these settings. In this chapter we present background on supervised learning, and describe our main contributions.

In supervised learning, the goal is to learn to classify objects into one of several classes, using only examples of objects, along with the class that they belong to. We focus on binary supervised learning, in which each object should be classified into one of two classes. As an example, consider the task of predicting whether a patient will present with diabetes, based on the patient's blood test results. In this example, one class represents patients who will present with diabetes and the other class represents patients who will not present with diabetes. The learner is given a set of examples, where each example is constituted of the blood test results of a patient, along with information on whether this patient has presented with diabetes or not. We term this set of examples the *training set*, or the *training sample*. The training set is used by the learner to infer a *classification rule*, which can be used to predict whether a new patient will present with diabetes, based on this patient's blood test results. The goal of the learner is to find a classification rule which is as accurate as possible in its predictions.

An important measure of the effectiveness of learning is how many labeled examples are needed in order to achieve a certain degree of classification accuracy. The *sample complexity* of a learning problem is the size of a training set required to guarantee a given accuracy on this problem. Equivalently, it is the accuracy that can be guaranteed for the learner, given the size of the training sample. We distinguish between the sample complexity, which is a statistical measure of the difficulty of learning, and computational complexity, which measures the amount of computation required to implement a learning strategy.

The "No free lunch" theorem for supervised learning [Wolpert and Macready, 1997] shows that no single supervised learning algorithm can provide a high-accuracy classification rule for all learning problems using the same sample size. In other words, the sample complexity of supervised learning without additional assumptions is unbounded. It follows that in order to have guarantees on learning, we need to consider more restricted classes of learning problems.

Most commonly, we restrict the set of learning problems that we consider by introducing the notion of a *hypothesis class*. This is a set of classification rules to which we wish to compare the result of our learning algorithm. In a specific learning context, the hypothesis class can represent our beliefs on the true nature of the classification rule for the problem. For instance, if we believe that diabetes can be identified via a boolean function using at most two values in a patient's blood test results, we can use the hypothesis class consisting only of such boolean functions. If our learning problem can indeed be classified with maximum accuracy using one of the classification rules in our hypothesis class, then we say that the problem is *realizable*. In this case, we can hope that our learning algorithm will achieve a small error, in absolute terms, when given enough labeled examples. If the problem is not realizable, then we say that we are in the *agnostic* setting. In this case, we hope that our learning algorithm will achieve a small *relative* error. That is, we hope that its classification accuracy will be close to that of the best classifier in our hypothesis class.

The sample complexity of learning relative to a specific hypothesis class strongly depends on its *complexity*. Loosely speaking, the complexity of a class is related to the number of different mappings between objects and labels that it allows. In this chapter we present several useful complexity measures for hypothesis classes, and the sample complexity guarantees that can be provided for them.

We start with defining our notation in Section 1.1. We then formally define a binary learning problem in Section 1.2. We present some popular and useful measures of learning accuracy in Section 1.3. We then turn to present relevant sample complexity results for binary supervised learning. The classical approach to sample complexity analysis of supervised learning is *distribution-free* analysis. In this approach we are interested in sample-complexity guarantees that hold regardless of the distribution of the labeled objects. We discuss this type of analysis in Section 1.4. We present tools for *distribution-dependent* analysis in Section 1.5. We apply each of the tools to the commonly used hypothesis class of *linear classifiers* in Section 1.6. Finally, we present the main contributions of this thesis in Section 1.7.

1.1 Notation

We denote the set of real numbers by $\mathbb{R}$, and the set of natural numbers by $\mathbb{N}$. We use the function $\text{sign} : \mathbb{R} \to \{\pm 1\}$ where

$$\text{sign}(x) = \begin{cases} 1 & x > 0 \\ -1 & x < 0 \\ 0 & x = 0. \end{cases}$$

For any integer $n \in \mathbb{N}$, we denote by $[n]$ the set $\{1, \ldots, n\}$. For a real number x, we denote $[x]_+ = \max\{0, x\}$ and $[\![x]\!] = \min([x]_+, 1)$. Let A and B be sets and let $f : A \to B$ be a function. Let $F \subseteq B^A$ be a set of functions from A to B. For a subset $X \subseteq A$, we denote the restriction of f to X by $f_{|X}$. The restriction of the set of functions F to X is denoted by $F_{|X} = \{f_{|X} \mid f \in F\}$. For a function $f : \mathbb{R} \to \mathbb{R}$, we denote its first and second derivatives by f' and f'' respectively.

Consider a probability distribution D over some domain. We denote the probability of a predicate p according to a distribution D by $\mathbb{P}_{X \sim D}[p(X)]$, although $X \sim D$ might be omitted when it is clear from context. Similarly, $\mathbb{E}_{X \sim D}[f(X)]$ denotes the expected value of $f(X)$ according to D. For a finite set A, we use A also to denote the uniform distribution over the elements of A.

Let d be an integer, and consider the Euclidean space $\mathbb{R}^d$. For a vector $x \in \mathbb{R}^d$, we denote its Euclidean norm by $\|x\|$. For a real matrix $\mathbb{X} \in \mathbb{R}^{d \times n}$, $\|\mathbb{X}\|$ stands for the Euclidean operator norm, that is $\|\mathbb{X}\| = \sup_{x \in \mathbb{R}^n : \|x\| \le 1} \|\mathbb{X}x\|$. We denote an origin-centered ball of radius r in a normed space $(\mathcal{S}, \|\cdot\|)$ by $\mathbb{B}_r(\mathcal{S}) = \{x \in \mathcal{S} \mid \|x\| \le r\}$. For $\mathcal{S} = \mathbb{R}^d$, we write $\mathbb{B}_r^d = \mathbb{B}_r(\mathbb{R}^d)$.

We sometimes represent sets of vectors in $\mathbb{R}^d$ using matrices. We say that $\mathbb{X} \in \mathbb{R}^{m \times d}$ is the matrix of a set $\{x_1, \ldots, x_m\} \subseteq \mathbb{R}^d$ if the rows in the matrix are exactly the vectors in the set. For uniqueness, we assume the rows of $\mathbb{X}$ are sorted according to an arbitrary fixed full order on vectors in $\mathbb{R}^d$. For a PSD matrix $\mathbb{X}$ denote the largest eigenvalue of $\mathbb{X}$ by $\lambda_{\max}(\mathbb{X})$ and the smallest eigenvalue by $\lambda_{\min}(\mathbb{X})$.

We use O-notation as follows: Whenever the expression $O(\xi)$ is used, it stands for $C_1 + C_2 \cdot \xi$ for some constants $C_1, C_2 \ge 0$. Similarly, $\Omega(\xi)$ stands for $C_2 \cdot \xi - C_1$ for some constants $C_1, C_2 \ge 0$. $\widetilde{O}(\xi)$ stands for $\xi \cdot p(\ln(\xi)) + C$ for some polynomial $p(\cdot)$ and some constant $C > 0$. We denote universal constants by C, c or C_1, C_2 etc. The values of these constants may change from statement to statement or even from line to line. The notations are summarized in Table 1.1.

1.2 Binary Supervised Learning

In binary supervised learning there are two possible classes, and we wish to predict which of these classes matches given objects. The two classes are commonly named the *positive class* and the

Table 1.1: Summary of Notation

$\mathbb{R}$	The real numbers
$\mathbb{N}$	The natural numbers
$\mathrm{sign}(x)$	The sign of x
$[n]$	$\{1, \dots, n\}$
$[x]_+$	$\max\{0, x\}$
$\llbracket x \rrbracket$	$\min([x]_+, 1)$
B^A	The functions from A to B
$f_{\vert X}$	The restriction of f to X
$F_{\vert X}$	$\{f_{\vert X} \mid f \in F\}$
f'	The first derivative of a function
f''	The second derivative of a function
$\mathbb{P}$	Probability
$\mathbb{E}$	Expectation
$\lVert x \rVert$	The Euclidean norm of x
$\mathbb{X}$	A matrix
$\lVert \mathbb{X} \rVert$	The Euclidean operator norm of $\mathbb{X}$
$\mathbb{B}_r(\mathcal{S})$	$\{x \in \mathcal{S} \mid \lVert x \rVert \leq r\}$
$\mathbb{B}_r^d$	$\mathbb{B}_r(\mathbb{R}^d)$
$\lambda_{\max}(\mathbb{X})$	The largest eigenvalue of $\mathbb{X}$
$\lambda_{\min}(\mathbb{X})$	The smallest eigenvalue of $\mathbb{X}$
$O(\xi)$	$C_1 + C_2 \cdot \xi$ for some constants $C_1, C_2 \geq 0$
$\Omega(\xi)$	$C_2 \cdot \xi - C_1$ for some constants $C_1, C_2 \geq 0$
$\widetilde{O}(\xi)$	$\xi \cdot p(\ln(z)) + C$ for a polynomial $p(\cdot)$ and $C > 0$
$C, c, C_1, C_2, \dots$	Posivite constants (value may change between expressions)

negative class, and are denoted by *labels* $+1$ and -1 respectively. We say that an object has a particular label if it belongs to the class denoted by that label. A *classification rule*, or a *classifier*, is a function from the domain of possible objects to the set of reals. We interpret the sign of the classifier's output as its prediction on the object—whether it is in the positive class or the negative class. The magnitude of the output is sometimes interpreted as the confidence of the classifier in its prediction.

To measure the accuracy of a classifier, we use the notion of *loss*, measured by a *loss function*. A loss function is a measure of the discrepancy between possible true labels and possible classifier outputs. Formally, it is a function $\ell : \{\pm 1\} \times \mathbb{R} \to \mathbb{R}_+$. When a classifier provides output $\hat{y}$ for a given object with true label y, it incurs loss $\ell(y, \hat{y})$. An accurate classifier is one that incurs a low classification loss.

The components of a binary supervised learning problem are a data domain $\mathcal{X}$, the label set $\{\pm 1\}$, a hypothesis class $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$, a distribution D over $\mathcal{X} \times \{\pm 1\}$, and a loss function ℓ. We assume that $\mathcal{X}$ is equipped with a σ-algebra, and consider only distributions that are measurable with respect to this σ-algebra. We denote by D_X the marginal distribution D induces on $\mathcal{X}$, and by $D_{Y|X}$ the conditional distribution that D induces on $\{\pm 1\}$ given an $X \in \mathcal{X}$.

Each element in the data domain $\mathcal{X}$ represents an object to be classified. For instance, in the diabetes example we can set $\mathcal{X} = \mathbb{R}^d$, where d is the number of measurements in a blood test. Then each element $x \in \mathcal{X}$ is a d-coordinate vector representing a single patient, where each coordinate holds the value of a single measurement in the patients' blood test results. The distribution D here is the probability of having a patient with measurements $x \in \mathbb{R}^d$ and condition $y \in \{\pm 1\}$.

Given the loss function ℓ, we denote the loss of a labeling function h over the distribution D by

$$\ell(h, D) = \mathbb{E}_{(X,Y)\sim D}[\ell(Y, h(X))].$$

Given the distribution D, the best classification rule for a learning problem is well defined: Consider the random pair $(X, Y) \sim D$, so that $X \in \mathcal{X}$ and $Y \in \{\pm 1\}$. The best possible classifier is Bayes' optimal classifier: For a given $x \in \mathcal{X}$, predict $\hat{y} \in \operatorname{argmin}_{y\in\mathbb{R}} \mathbb{E}[\ell(Y, y) \mid X = x]$.

The minimal loss that can be achieved by a classifier in the hypothesis class $\mathcal{H}$ is

$$\ell^*(\mathcal{H}, D) = \inf_{h\in\mathcal{H}} \ell(h, D).$$

If $\ell^*(\mathcal{H}, D) = 0$ then the learning problem is realizable, otherwise it is agnostic. We omit $\mathcal{H}$ and write simply $\ell^*(D)$ if $\mathcal{H}$ is clear from context.

The training sample that the learning algorithm receives as input is a set of m examples $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \subseteq \mathcal{X} \times \{\pm 1\}$, where m is the sample size.[1] Given S, we denote the set of its examples without their labels by $S_X = \{x_1, \ldots, x_m\}$. Crucially, we assume that each pair (x_i, y_i) is drawn independently according to D. A *learning algorithm* is a (possibly non-deterministic) function $\mathcal{A} : \cup_{m=1}^{\infty}(\mathcal{X} \times \{\pm 1\})^m \to \mathbb{R}^{\mathcal{X}}$, that receives a training set, and returns a function for classifying objects in $\mathcal{X}$ into real values. We say that $\mathcal{A}$ is doing *proper learning* of $\mathcal{H}$ if for any possible training set S, $\mathcal{A}(S) \in \mathcal{H}$. The loss of a learning algorithm with input sample S is simply $\ell(\mathcal{A}(S), D)$.

We analyze learning algorithms in the Probably Approximately Correct (PAC) framework [Valiant, 1984]: We bound the loss of the algorithm with a high probability over the random draw of samples. The high-probability loss of an algorithm $\mathcal{A}$ with respect to samples of size m, a distribution D and a confidence parameter $\delta \in (0, 1)$ is

$$\ell(\mathcal{A}, D, m, \delta) = \inf\{\epsilon \geq 0 \mid \mathbb{P}_{S\sim D^m}[\ell(\mathcal{A}(S), D) \geq \epsilon] \leq \delta\}.$$

[1] Samples are in fact multisets, since a labeled example may repeat several times. We use the set notation for simplicity.

In words, we say that with probability at least $1-\delta$ over samples of size m drawn from D^m, $\mathcal{A}$ has a loss of no more than ϵ.

1.3 Common Loss Functions

As mention above, a loss function $\ell : \{\pm 1\} \times \mathbb{R} \to \mathbb{R}_+$ measures the accuracy of a prediction given the true label. Different loss functions represent different measures of accuracy. In this section we discuss several popular and useful loss functions, which will be used throughout this work.

We interpret the sign of the classifier's output as its prediction for the label of the input object, where an output of zero is interpreted as "no prediction". Thus, perhaps the most natural loss function is the one which penalizes the classifier by a constant amount whenever the sign of its output (that is, the classifier's prediction) is different from the true label of the object. This loss is termed the *zero-one loss*, and is defined by $\ell_{0/1}(y, \hat{y}) = \mathbb{I}[y\hat{y} \leq 0]$ (see Figure 1.1). It is easy to see that $\ell_{0/1}(h, D) = \mathbb{P}_{(X,Y)\sim D}[Y \neq \text{sign}(h(X))]$, thus the average zero-one loss is simply the probability that the classifier does not predict the correct label.

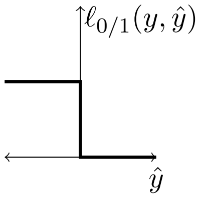

Figure 1.1: The zero-one loss for $y = 1$.

The magnitude of the classifier's output can be interpreted as a measure of its "confidence" in the prediction. Thus, it makes sense to require that the predictor not only output the right label, but do so with a high confidence. This is captured by the *margin loss*: Define a confidence value $\gamma > 0$, also termed the *margin*, and penalize any prediction that is either incorrect, or correct but with a confidence (also margin) of less than γ. Formally, the margin-loss is $\ell_\gamma(y, \hat{y}) = \mathbb{I}[y\hat{y} \leq \gamma]$ (see Figure 1.2).

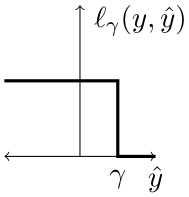

Figure 1.2: The margin loss for $y = 1$ and $\gamma = 0.5$.

While the margin loss and the zero-one loss are very natural, they pose a problem for computationally efficient implementations, since it is NP-hard to minimize them on many useful natural hypothesis classes, such as the class of linear classifiers that we describe in Section 1.6 [Höffgen et al., 1995]. Thus, in many cases a *surrogate loss* is used instead of these losses. A surrogate loss needs to be computationally easy to optimize, while close in some sense to the loss is replaces. A popular choice is the *hinge-loss*, defined by $\ell_{\mathrm{hl}(\gamma)}(y, \hat{y}) = [1 - y\hat{y}/\gamma]_+$ (see Figure 1.3). This loss is convex, which means that it can be minimized efficiently. It is an upper bound for the zero-one loss, and if there are no low-confidence predictions, it is also an upper bound for the margin-loss. Thus, if the hinge-loss is small, then the zero-one loss and perhaps the margin loss are also small. The hinge-loss can be considered as natural even without regarding it as a surrogate for another loss, since it penalizes a classifier more if it is more “confident” in its wrong prediction.

Another useful property of the hinge-loss is that it is *Lipschitz*: In general, a function f from a normed space to a normed space is c-Lipschitz for $c \geq 0$ if $\|f(a) - f(b)\| \leq c\|a - b\|$ for any a, b in the domain. For losses, we say that they are c-Lipschitz if they are c-Lipschitz in their second argument, with respect to the absoulte-value norm. Formally, a loss is c-Lipschitz if

$$\forall y \in \{\pm 1\}, a, b \in \mathbb{R}, \quad |\ell(y, a) - \ell(y, b)| \leq c|a - b|.$$

The hinge-loss with a margin parameter γ is thus $1/\gamma$-Lipschitz. This property implies that the value of the loss is closely coupled with the value of the prediction. As a result, certain sample-complexity analysis tools can be easily applied to this loss.

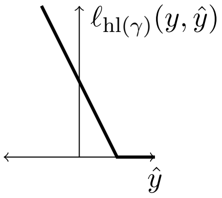

Figure 1.3: The hinge-loss for $y = 1$ and $\gamma = 0.5$.

Finally, we also consider the *ramp-loss*, defined for $\gamma > 0$ by $\mathrm{ramp}_\gamma(y, \hat{y}) = [\![1 - y\hat{y}/\gamma]\!]$ (see Figure 1.4). The ramp-loss is equal to the hinge-loss, except that it is never more than 1. Thus, a classifier is penalized by a constant amount for wrong predictions, and by a smaller amount for the right prediction with a small confidence. As we shall see, the ramp-loss is a useful tool for proving sample complexity bounds, since it is upper-bounded by the margin loss and lower-bounded by the zero-one loss. Furthermore, it is $1/\gamma$-Lipschitz just like the hinge-loss.

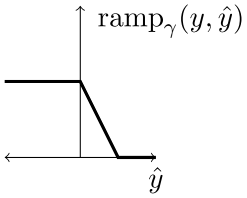

Figure 1.4: The ramp loss for $y = 1$ and $\gamma = 0.5$.

1.4 Distribution-Free Sample Complexity

In distribution-free analysis, we are interested in sample-complexity guarantees that hold regardless of the distribution D. In other words, we are interested in the quantity

$$\inf_{\mathcal{A}} \max_{D} \ell(\mathcal{A}, D, m, \delta),$$

where the infimum on $\mathcal{A}$ is taken over all learning algorithms, and the maximum on D is taken over all the distributions over the σ-algebra of $\mathcal{X} \times \{\pm 1\}$. The critical factor in determining the distribution-free sample-complexity of a supervised learning problem is the complexity of the hypothesis class $\mathcal{H}$. Several complexity measures for hypothesis classes have been proposed, each providing a different type of guarantee.

1.4.1 The VC dimension

We first consider learning with respect to the zero-one loss. We may assume without loss of generality that $\mathcal{H} \subseteq \{0, -1, +1\}^{\mathcal{X}}$, by considering the set $\mathrm{sign} \circ \mathcal{H} = \{\mathrm{sign} \circ h \mid h \in \mathcal{H}\}$, since the zero-one loss is affected only by the sign of the prediction. However, the common view, which we adhere to here, considers $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$, by arbitrarily fixing $\mathrm{sign}(0) = 1$.

For the zero-one loss, it can be shown that the quantity controlling the sample complexity of the best learning algorithm for $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ is the number of different labelings that $\mathcal{H}$ induces on finite sets from the domain $\mathcal{X}$. Intuitively, if there are less possible labelings, then the learner can achieve a high accuracy with fewer training examples, since it needs to distinguish between fewer possibilities. The number of labelings is measured by the *growth function* of $\mathcal{H}$, which is the function $\Pi_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ defined by

$$\Pi_{\mathcal{H}}(n) = \max\{|\mathcal{H}_{|X}| \mid X \subseteq \mathcal{X}, |X| = n\}.$$

We would like to establish an upper bound for distribution-free binary supervised learning, as a function of the growth function. Ths can be shown via a *uniform convergence* argument: For every individual hypothesis $h \in \mathcal{H}$, its loss on a random sample converges fast to its loss on the distribution as the sample size grows. The effective number of hypotheses that need to be considered can be bounded by the growth function. Thus, a union bound can be used to show that with high probability, *all of the hypotheses simultaneously* incur a loss on the random sample that is close to the loss they incur on the distribution. This means that a learning algorithm may *choose the hypothesis that is the most accurate on the sample*, and is guaranteed that its loss on the distribution will also be low. The principle of choosing the hypothesis that does best on the sample is known as *Empirical Risk Minimization (ERM)*. Formally, we define an ERM algorithm as follows.

Definition 1.1 (ERM algorithm)**.** *A learning algorithm $\mathcal{A}$ is an* ERM algorithm *for hypothesis class $\mathcal{H}$ and loss ℓ if*

$$\forall S \subseteq \mathcal{X} \times \{\pm 1\}, \quad \mathcal{A}(S) \in \operatorname*{argmin}_{h \in \mathcal{H}} \ell(h, S).$$

We are now refdy to state the distribution-free upper bound based on the growth function. This result can be derived from Anthony and Bartlett [1999] (Theorem 4.3). The first results of this type are due to Vapnik and Chervonenkis [1971].

Theorem 1.2. *There exists a universal constant C such that the following holds. Let $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ be some hypothesis class. For any ERM algorithm $\mathcal{A}$ for $\ell_{0/1}$, and for any distribution D,*

$$\ell_{0/1}(\mathcal{A}, D, m, \delta) - \ell^*_{0/1}(\mathcal{H}, D) \leq \sqrt{\frac{C \cdot \ln(\frac{4\Pi_{\mathcal{H}}(2m)}{\delta})}{m}}.$$

As it turns out, a single quantity suffices to characterize the growth function, and hence the distribution-free sample complexity of binary supervised learning, up to logarithmic factors. This quantity is the *VC-dimension* of $\mathcal{H}$ [Vapnik and Chervonenkis, 1971]. As we shall now show, the VC-dimension can be used to provide both an upper bound and a lower bound on the distribution-free sample complexity of binary supervised learning. The VC-dimension of a hypothesis class is defined using the notion of *shattering*.

Definition 1.3 (Shattering)**.** *Let $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ be some hypothesis class. A set $X \subseteq \mathcal{X}$ is* shattered by *$\mathcal{H}$ if for every labeling of X, denoted $g : X \to \{\pm 1\}$, there exists a hypothesis $h \in \mathcal{H}$ such that $h_{|X} = g$.*

For instance, suppose the domain $\mathcal{X}$ is $[0, 1]$, and the hypothesis class is the set of functions $h_{[a,b]}$ that map $[a, b]$ to $+1$ and the rest of $[0, 1]$ to -1. Then any set of two different points in $[0, 1]$ is shattered, by choosing an appropriate interval for each possible labeling (see Figure 1.5, top). However, no set of three points is shattered, since there is a labeling that no function $h_{[a,b]}$ can generate (see Figure 1.5, bottom).

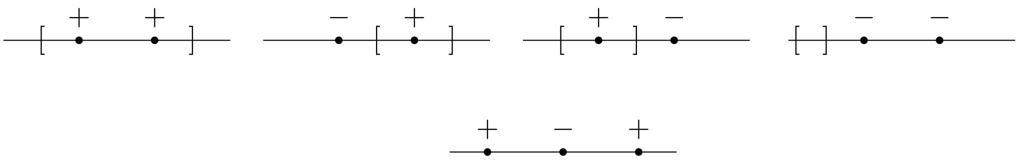

Figure 1.5: Top: A set of two points is shattered. Bottom: A set of three points is not shattered.

The VC-dimension is the largest size of a shattered set in the domain. Formally,

Definition 1.4 (VC-dimension)**.** *Let $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ be some hypothesis class. The* VC-dimension *of $\mathcal{H}$, denoted $\mathrm{VC}(\mathcal{H})$, is the size of the largest subset of $\mathcal{X}$ that is shattered by $\mathcal{H}$.*

It is easy to see that the existence of a shattered set of size n implies that $\Pi_{\mathcal{H}}(m) \geq 2^n$ for any $m \geq n$. Sauer's Lemma [Sauer, 1972, Vapnik and Chervonenkis, 1971] shows that the VC-dimension provides an upper bound to the growth function as well.

Lemma 1.5 (Sauer's Lemma)**.** *Let X be a set of size n, and let A be a set of functions from X to $\{\pm 1\}$. If the VC-dimension of A is d, then $|A| \leq \sum_{i=1}^{d} \binom{n}{i}$.*

For any $m \geq d$, we have [Chari et al., 1994]

$$\sum_{i=1}^{d} \binom{m}{i} < \left(\frac{em}{d}\right)^d.$$

Thus, for any hypothesis class $\mathcal{H}$ with VC-dimension d and any $n \geq d$,

$$2^d \leq \Pi_{\mathcal{H}}(n) \leq \left(\frac{em}{d}\right)^d.$$

The logarithm of the growth function is thus about the same as the VC-dimension, up to logarithmic factors. Therefore, we can conclude a sample-complexity upper bound using the VC-dimension. The following formulation follows Anthony and Bartlett [1999] (Theorem 4.2).

Theorem 1.6. *There are universal constants C and c such that the following holds. Let $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ be some hypothesis class with VC-dimension d. For any ERM algorithm $\mathcal{A}$, and for any distribution D, for any $m \geq d$ and $\delta \in (0,1)$*

$$\ell_{0/1}(\mathcal{A}, D, m, \delta) - \ell^*_{0/1}(\mathcal{H}, D) \leq \sqrt{\frac{C(d \ln(\frac{2em}{d}) + \ln(\frac{c}{\delta}))}{m}}.$$

It is possible to get an improved upper bound using a method known as *chaining*. The following is based on Anthony and Bartlett [1999], Theorem 4.10.

Theorem 1.7. *There is a universal constant C such that the following holds. Let $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ be some hypothesis class with VC-dimension d. For any ERM algorithm $\mathcal{A}$, and for any distribution D, for any $m \geq d$ and $\delta \in (0,1)$*

$$\ell_{0/1}(\mathcal{A}, D, m, \delta) - \ell^*_{0/1}(\mathcal{H}, D) \leq \sqrt{\frac{C(d + \ln(\frac{1}{\delta}))}{m}}.$$

It is also possible to show a lower bound on the distribution-free sample complexity based on the VC-dimension. For a distribution-free lower bound, it suffices to show that for any learning there exists a distribution such that the algorithm would require many examples to learn it accurately. Intuitively, if the distribution D is supported by a shattered set, then the label of any element in the set provides no indication on the labels of the other elements. Based on this idea, the following theorem can be proved.

Theorem 1.8. *There exist universal constants $c, C > 0$ such that For any learning algorithm $\mathcal{A}$ and any integer m, and for every hypothesis class $\mathcal{H}$ with VC-dimension d, there exists a distribution D such that and for any $\delta \leq c$,*

$$\ell_{0/1}(\mathcal{A}, D, m, \delta) - \ell^*_{0/1}(\mathcal{H}, D) \geq \sqrt{\frac{C \cdot d}{m}}.$$

This result has appeared in several places, including Vapnik and Chervonenkis [1974] and Devroye and Lugosi [1995]. Here we follow the formulation of Anthony and Bartlett [1999], theorem 5.2.

Considering Theorem 1.7 and Theorem 1.8 together, we conclude that ERM algorithms achieve the best possible distribution-free sample complexity.

1.4.2 Covering Numbers

We have seen that if the VC-dimension is bounded, then the effective number of hypotheses is bounded, and thus a uniform convergence argument can be used to provide a sample complexity guarantee. When the VC-dimension is not bounded, we cannot use uniform convergence in the same way. In fact, Theorem 1.8 shows that the same type of guarantee as in Theorem 1.7 does not exist in this case. Nonetheless, a guarantee can be provided, if we require slightly less from the learning algorithm.

We will provide a guarantee on the zero-one loss of the learning algorithm relative not to the best achievable margin loss. This is in contrast with the last section, where the guarantee was relative to the best achievable zero-one loss. This will imply a high classification accuracy if it is possible to classify the objects in the domain correctly and *with high confidence* using the given hypothesis class.

When comparing to the margin loss, we can use a uniform convergence argument as follows: Instead of counting the number of different labelings induced by hypotheses in our hypothesis class, we will "bundle together" classifiers that emit similar values, and count only the number of classifiers that are sufficiently far from each other. Whenever this number is bounded, a guarantee relative to the margin loss can be provided.

Formally, we count classifiers that are sufficiently far from each other using the notion of a *covering number*. Let $(\mathcal{B}, \|\cdot\|_\circ)$ be a normed space. A γ*-cover* of this space is a subset $\mathcal{C} \subseteq \mathcal{B}$ such that for any $x \in \mathcal{B}$ there exists a $y \in \mathcal{C}$ such that $\|x - y\|_\circ \leq \gamma$. The covering number for given $\gamma > 0$, $\mathcal{B}$ and $\circ$, denoted by $\mathcal{N}(\gamma, \mathcal{B}, \circ)$, is the size of the smallest such γ-covering for $\mathcal{B}$.

We use covering numbers to measure the "effective size" of the hypothesis class with respect to a given set $X \subseteq \mathcal{X}$. Thus, we consider normed spaces of functions $(\mathcal{F}, \|\cdot\|_{L_p(X)})$, where $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ is a set of real-valued functions, and the norm $L_p(X)$ for $p \geq 1$ is defined by

$$\|f\|_{L_p(X)} = \left(\frac{1}{|X|} \sum_{s \in X} |f(s)|^p \right)^{1/p}.$$

For $p = \infty$, $L_\infty(X)$ is defined by $\|f\|_{L_\infty(X)} = \max_{s \in X} |f(X)|$. The covering number of $\mathcal{F}$ for a sample size m with respect to the L_p norm is

$$\mathcal{N}_m(\gamma, \mathcal{F}, p) = \sup_{X \subseteq \mathcal{X} : |X| = m} \mathcal{N}(\gamma, \mathcal{F}, L_p(X)).$$

As we will see in the next section, a small covering number for a function class implies faster uniform convergence rates, hence a smaller sample complexity for learning.

While the covering number can be much larger than the growth function, the relation between the two quantities can be bounded. By Dudley [1978], for any $\mathcal{H}$ with VC-dimension d, any $X \subseteq \mathcal{X}$, and any $\gamma > 0$,

$$\ln \mathcal{N}(\gamma, \mathcal{H}, L_2(X)) \leq 2d \ln \left(\frac{4e}{\gamma^2} \right). \tag{1.1}$$

1.4.3 The Fat-shattering dimension

The covering number of a hypothesis class can be thought of as a scale-sensitive version of the growth function. Just as the behavior of the growth function can be characterized by the combinatorial notion of a VC-dimension, it is possible to characterize the behavior of certain covering numbers using a scale-sensitive combinatorial notion termed the *fat-shattering dimension*, first introduced in Kearns and Schapire [1994]. The fat-shattering dimension is defined using the scale-sensitive notion of *fat-shattering*.

Definition 1.9 (Fat-shattering)**.** *Let $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ be a hypothesis class, and let $\gamma > 0$. A set $\{x_1, \ldots, x_m\} \subseteq \mathcal{X}$ is γ-**shattered by** $\mathcal{H}$ if there is a vector $r \in \mathbb{R}^m$ such that for every vector $y \in \{\pm 1\}^m$ there is an $h \in \mathcal{H}$ such that*

$$\forall i \in [m], \quad y[i](h(x_i) - r[i]) \geq \gamma.$$

The fat-shattering dimension is the size of the largest set which is fat-shattered.

Definition 1.10 (Fat-shattering dimension)**.** *The γ-**fat-shattering dimension** of $\mathcal{H}$, denoted $\mathrm{Fat}(\gamma, \mathcal{H})$, is the size of the largest subset of $\mathcal{X}$ that is γ-shattered by $\mathcal{H}$.*

The fat-shattering dimension is strongly related to the behavior of the L_∞ covering numbers of $\mathcal{H}$. This can be seen in the following bounds. The first bound is from Bartlett et al. [1997]

Theorem 1.11. *Let F be a set of real-valued functions and let $\gamma > 0$. For $m \geq \mathrm{Fat}(16\gamma, F)$,*

$$e^{\mathrm{Fat}(16\gamma, F)/8} \leq \mathcal{N}_m(\gamma, F, \infty).$$

The reverse bound, listed below, is due to Anthony and Bartlett [1999] (Theorem 12.8), following Alon et al. [1993].

Theorem 1.12. *Let $\mathcal{F}$ be a set of real-valued functions with range in $[0, B]$. Let $\gamma > 0$. Let $d = \mathrm{Fat}(\frac{\gamma}{4}, \mathcal{F})$. For all $m \geq 1$,*

$$\mathcal{N}_m(\gamma, F, \infty) < 2\left(\frac{4B^2 m}{\gamma^2}\right)^{d \log(4eBm/\gamma d)}.$$

We use the term *Margin Error Minimization* (MEM) algorithms to refer to ERM algorithms that minimize the margin loss. Using Theorem 1.12 and a uniform convergence argument, it is possible to derive a sample complexity guarantee for MEM algorithms as a function of the fat-shattering dimension. This is stated in the following theorem, based on Anthony and Bartlett [1999] (Theorem 13.4).

Theorem 1.13. *There are universal constants $C, c > 0$ such that the following holds. Let $\gamma > 0$. Let $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ be some hypothesis class with $\mathrm{Fat}(\gamma/8, \mathcal{H}) = d \geq 1$. Then for any integer m and $\delta \in (0,1)$, and for any distribution D, for any margin-error minimization algorithm $\mathcal{A}$ for $\mathcal{H}$,*

$$\ell_{0/1}(\mathcal{A}, D, m, \delta) - \ell^*_\gamma(\mathcal{H}, D) \leq \sqrt{\frac{C(d\ln(\frac{c \cdot m}{d})\ln(c \cdot m) + \ln(\frac{c}{\delta}))}{m}}.$$

1.4.4 The Pseudo-dimension

By taking the margin γ to zero, we can get a guarantee for the zero-one loss relative to the best zero-one loss $\ell^*_{0/1}$, since $\ell_{0/1} = \lim_{\gamma \to 0} \ell_\gamma$. The γ-shattering dimension of $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ for $\gamma \to 0$ is termed the *pseudo-dimension* [Pollard, 1984] of $\mathcal{H}$. This dimension is equal to the VC-dimension of the class $T_{\mathcal{H}} = \{(x,z) \mapsto \mathrm{sign}(h(x) - z) \mid h \in \mathcal{H}\}$, where $x \in \mathcal{X}$ and $z \in \mathbb{R}$. Alternatively, the pseudo-dimension can be defined directly on the class $\mathcal{H}$ as follows.

Definition 1.14 (Pseudo-shattering)**.** *Let $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ be a class of real-valued functions. A set $\{x_1, \ldots, x_m\} \subseteq \mathcal{X}$ is **pseudo-shattered** by $\mathcal{H}$ if there is a vector $r \in \mathbb{R}^m$ such that for every $y \in \{\pm 1\}^m$ there is an $h \in \mathcal{H}$ such that*

$$\forall i \in [m], \quad \mathrm{sign}(h(x_i) - r[i]) = y[i].$$

Definition 1.15 (Pseudo-dimension)**.** *Let $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ be some hypothesis class. The* pseudo-dimension *of $\mathcal{H}$ is the size of the largest subset of $\mathcal{X}$ that is pseudo-shattered by $\mathcal{H}$.*

The sample complexity guarantees of Theorem 1.13 hold also with d standing for the pseudo-dimension and $\ell^*_{0/1}$ instead of ℓ^*_γ.

The L_2 covering number of a function class can be bounded using the pseudo dimension as follows [see e.g. Bartlett, 2006, Theorem 3.1]: There are constants C_1 and C_2 such that if the pseudo-dimension of $\mathcal{H} \subseteq [0,1]^{\mathcal{X}}$ is d, then

$$\mathcal{N}(\gamma, \mathcal{H}, L_2(S)) \leq C_1 \left(\frac{C_2}{\gamma^2}\right)^d. \tag{1.2}$$

1.5 Distribution Dependence and General Losses

In the previous section we showed distribution-free sample-complexity bounds when the target loss is the zero-one loss. Can we bound the sample complexity required to achieve a low loss for other target losses? Furthermore, is it possible to get better upper bounds than distribution-free bounds, based on the properties of the specific distribution in our learning problem?

These two questions can be answered in the affirmative, using the tool of *Rademacher complexity* [Bartlett and Mendelson, 2002]. Let us start with necessary definitions. Let $\mathcal{Z}$ be some domain. The *empirical Rademacher complexity* of a class of functions $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{Z}}$ with respect to a set $S = \{z_i\}_{i\in[m]} \subseteq \mathcal{Z}$ is

$$\mathcal{R}(\mathcal{F}, S) = \frac{1}{m}\mathbb{E}_\sigma[|\sup_{f\in\mathcal{F}} \sum_{i\in[m]} \sigma_i f(z_i)|],$$

where $\sigma = (\sigma_1, \ldots, \sigma_m)$ are m independent uniform $\{\pm 1\}$-valued variables. The *average Rademacher complexity* of $\mathcal{F}$ with respect to a distribution D over $\mathcal{Z}$ and a sample size m is

$$\mathcal{R}_m(\mathcal{F}, D) = \mathbb{E}_{S\sim D^m}[\mathcal{R}(\mathcal{F}, S)].$$

Assume a hypothesis class $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ and a loss function $\ell : \{\pm 1\} \times \mathbb{R} \to \mathbb{R}$. For a hypothesis $h \in \mathcal{H}$, we introduce the function $h_\ell : \mathcal{X} \times \{\pm 1\} \to \mathbb{R}$, defined by $h_\ell(x, y) = \ell(y, h(x))$. We further define the function class $\mathcal{H}_\ell = \{h_\ell \mid h \in \mathcal{H}\} \subseteq \mathbb{R}^{\mathcal{X}\times\{\pm 1\}}$.

As shown in Bartlett and Mendelson [2002], Rademacher complexities can be used to derive sample complexity bounds for general bounded losses: Assume that the range of $\mathcal{H}_\ell$ is in $[0, 1]$. For any $\delta \in (0, 1)$, with probability of $1 - \delta$ over the draw of samples $S \subseteq \mathcal{X} \times \{\pm 1\}$ of size m according to D, every $h \in \mathcal{H}$ satisfies

$$\ell(h, D) \le \ell(h, S) + 2\mathcal{R}_m(\mathcal{H}_\ell, D) + \sqrt{\frac{8\ln(2/\delta)}{m}}. \tag{1.3}$$

This *distribution-dependent* guarantee can be used, for instance, to bound the loss of an ERM algorithm $\mathcal{A}$ for $\mathcal{H}$ and ℓ, relative to the best loss $\ell^*(\mathcal{H}, D)$, as follows. From Eq. (1.3) we have that with probability $1 - \delta/2$ over the samples S of size m,

$$\ell(\mathcal{A}(S), D) \le \ell(\mathcal{A}(S), S) + 2\mathcal{R}_m(\mathcal{H}_\ell, D) + \sqrt{\frac{8\ln(2/\delta)}{m}}. \tag{1.4}$$

Set $h^* \in \mathcal{H}$ such that $\ell(h^*, D) = \ell^*(\mathcal{H}, D)$.[2] Since $\mathcal{A}$ is an ERM algorithm, $\ell(\mathcal{A}(S), S) \le \ell(h^*, S)$. Now, by Hoeffding's inequality, since the range of $\mathcal{H}_\ell$ is in $[0, 1]$, with probability at least $1 - \delta/2$

$$\ell(h^*, S) \le \ell(h^*, D) + \sqrt{\frac{\ln(2/\delta)}{2m}}.$$

[2] If no element in $\mathcal{H}$ achieves the infimum $\ell^*(\mathcal{H}, D)$, a similar yet more arduous argument can be carried out by setting h^* such that $\ell(h^*, D) \le \ell^*(\mathcal{H}, D) + \epsilon$, for any $\epsilon > 0$.

Therefore we conclude that

$$\ell(\mathcal{A}, D, m, \delta) \leq \ell^*(\mathcal{H}, D) + 2\mathcal{R}_m(\mathcal{H}_\ell, D) + \sqrt{\frac{14\ln(2/\delta)}{m}}. \tag{1.5}$$

To get distribution-free results with general losses, one can use the *worst-case Rademacher complexity*, defined as follows for any integer m:

$$\mathcal{R}_m^{\sup}(\mathcal{F}) = \sup_{S \subseteq \mathcal{Z}^m} \mathcal{R}(\mathcal{F}, S).$$

Thus, the upper bounds shown above can be turned into distribution-free bounds by replacing $\mathcal{R}_m(\mathcal{H}_\ell, D)$ with $\mathcal{R}_m^{\sup}(\mathcal{H}_\ell)$.

A closely related complexity measure, termed the *Gaussian complexity*, can be defined analogously to the Rademacher complexity. The empirical Gaussian complexity is

$$\mathcal{G}(\mathcal{F}, S) = \frac{1}{m}\mathbb{E}_s[|\sup_{f \in \mathcal{F}} \sum_{i \in [m]} s_i f(x_i, y_i)|],$$

Where $s = (s_1, \ldots, s_m)$ are independent standard normal variables. Similarly, $\mathcal{G}_m(\mathcal{F}, D)$ is the expectation of $\mathcal{G}(\mathcal{F}, S)$ over samples of size m. The Gaussian complexity and the Rademacher complexity are related as follows [Tomczak-Jaegermann, 1989]: There are constants $c, C > 0$ such that for all function classes $\mathcal{F}$ and distributions D,

$$c \cdot \mathcal{R}_m(\mathcal{F}, D) \leq \mathcal{G}_m(\mathcal{F}, D) \leq C \cdot \ln(m)\mathcal{R}_m(\mathcal{F}, D). \tag{1.6}$$

1.5.1 Relationships with other complexity measures

The Rademacher complexity can be related to the other complexity measures we have defined in previous sections. In this section we survey some useful relationships.

First, the Rademacher complexity can be bounded by the VC-dimension as follows [Bartlett and Mendelson, 2002]. For any distribution D over $\mathcal{X} \times \{\pm 1\}$,

$$\mathcal{R}_m(\mathcal{H}_{\ell_{0/1}}, D) \leq O\left(\sqrt{\frac{VC(\mathcal{H})}{m}}\right).$$

It is also easy to see, by considering the upper bound in Eq. (1.3) and the lower bound in Theorem 1.8, that there exists a distribution D such that

$$\mathcal{R}_m(\mathcal{H}_{\ell_{0/1}}, D) \geq O\left(\sqrt{\frac{VC(\mathcal{H})}{m}}\right).$$

For classes of real-valued functions, the Rademacher complexity can be bounded as a function of the fat-shattering dimension of the class, but this depends on the entire behavior of the fat-shattering dimension as a function of γ [Mendelson, 2002]. For the other direction, the worst-case Rademacher complexity can be tied to the fat-shattering dimension via the following result [See e.g. Mendelson, 2002, Theorem 4.11].

Theorem 1.16. *Let $m \geq 1$ and $\gamma \geq 0$. If $\mathcal{R}_m^{\sup}(\mathcal{F}) \leq \gamma$ then the γ-fat-shattering dimension of $\mathcal{F}$ is at most m.*

The Rademacher complexity is strongly related to the L_2 covering number of the function class. First, we have an upper bound for the L_2 covering number based on the Rademacher complexity. Sudakov's minoration theorem (Sudakov 1971, and see also Ledoux and Talagrand, 1991) states that there exists a constant $C > 0$ such that for any $\eta > 0$

$$\ln \mathcal{N}(\eta, \mathcal{F}, L_2(S)) \leq \frac{Cm}{\eta^2}\mathcal{G}^2(\mathcal{F}, S). \tag{1.7}$$

Due to Eq. (1.6), this implies a bound on the Rademacher complexity as well.

To bound for the Rademacher complexity from above using covering numbers, one needs to consider the behavior of the covering number as a function of γ. A classical result is Dudley's entropy integral [Dudley, 1967], which states that

$$\mathcal{R}(\mathcal{F}, S) \leq \frac{12}{\sqrt{m}} \int_0^\infty \sqrt{\ln \mathcal{N}(\gamma, \mathcal{F}, L_2(S))}\, d\gamma. \tag{1.8}$$

If the integral is unbounded, the following refinement can be used [Srebro et al., 2010, Lemma A.3]: For all $\epsilon \in (0, 1]$, for all real function classes $\mathcal{F}$ with range $[0, 1]$ and for all sets S,

$$\mathcal{R}(\mathcal{F}, S) \leq 4\epsilon + \frac{10}{\sqrt{m}} \int_\epsilon^1 \sqrt{\ln \mathcal{N}(\gamma, \mathcal{F}, L_2(S))}\, d\gamma. \tag{1.9}$$

Lastly, instead of an integral, one can bound the Rademacher complexity also using a finite sum as follows [Mendelson, 2002, Lemma 3.7]: Let $\epsilon_i = 2^{-i}$. Then

$$\sqrt{m}\mathcal{R}(\textsc{ramp}_\gamma, S) \leq C \sum_{i \in [N]} \epsilon_{i-1} \sqrt{\ln \mathcal{N}(\epsilon_i, \textsc{ramp}_\gamma, L_2(S))} + 2\epsilon_N \sqrt{m}. \tag{1.10}$$

1.6 Linear Classifiers

When defining a learning problem, the hypothesis class should represent our prior knowledge or beliefs about what classifiers might be good predictors in this problem. In this way, the learning algorithm can enjoy a hypothesis class of low complexity while keeping the best loss $\ell^*(\mathcal{H}, D)$ low as well, thus a small sample will suffice to achieve a good prediction accuracy.

While this prior knowledge can be specific to a problem, it turns out that some hypothesis classes can be used very successfully on a vast range of problems. In particular, a common and successful approach is to set the data domain to be $\mathcal{X} \subseteq \mathbb{R}^d$ for some integer d, so that each object is represented by a vector $x \in \mathbb{R}^d$, and to learn a predictor relative to the hypothesis class of *linear classifiers*. A linear classifier is the function h_w, for some vector $w \in \mathbb{R}^d$, defined by $h_w(x) = \langle w, x\rangle$. The label predicted by such a classifier is $\mathrm{sign} \circ h_w(x) = \mathrm{sign}(\langle w, x\rangle)$. The use of linear classifiers has proved very useful in practice, and is at the core of popular learning algorithms such as the Perceptron [Rosenblatt, 1958] and Support Vector Machines (SVMs) [Boser et al., 1992, Cortes and Vapnik, 1995, Vapnik, 1995].

It should be noted that the class of linear classifiers is sometimes defined with a bias: $h_{w,b}(x) = \langle w, x\rangle - b$ for $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$. However, the two formulations are practically equivalent, since any biased classifier can be turned into an unbiased classifier by adding another dimension $d + 1$ and setting $x[d + 1] = 1$ for all the objects in the domain. We will adhere to the formulation of linear classifiers without a bias, also termed *homogeneous linear classifiers*.

In this work we focus on homogeneous linear classifiers that can be described by vectors in the unit ball. For a normed space $\mathcal{S}$, denote $\mathcal{W}(\mathcal{S}) = \{h_w \mid w \in \mathcal{S}, \|w\| \leq 1\}$. We write simply $\mathcal{W}$ when $\mathcal{S}$ is clear from context. There are well-known bounds for the complexity terms of hypothesis classes of this form. First, the VC-dimension and the pseudo-dimension for linear classifiers in Euclidean space can be calculated exactly [Dudley, 1978, Pollard, 1984].

Theorem 1.17. *The VC-dimension of $\mathcal{W}(\mathbb{R}^d)$ is exactly d.*

Theorem 1.18. *The pseudo-dimension of $\mathcal{W}(\mathbb{R}^d)$ is exactly d.*

By Theorem 1.7, this implies that a bounded relative zero-one loss can be achieved by an ERM, using a sample of size $O(d/\epsilon^2)$. Now, suppose the dimension d is very large. In that case, the sample complexity guarantees based on the VC-dimension might be meaningless. It turns out that by using a margin formulation, it is possible to get guarantees that are independent of the dimensionality of the space. Thus, we can use even an infinite-dimensional space: Instead of a Euclidean space $\mathbb{R}^d$, consider a real-valued *Hilbert space*: This is a vector space $\mathcal{X}$ with an associated real-valued inner product $\langle\cdot,\cdot\rangle : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The norm on the Hilbert space is defined by $\|x\| = \sqrt{\langle x, x\rangle}$. Linear classifiers can be defined on Hilbert spaces similarly to their definition on $\mathbb{R}^d$, using the inner product of the Hilbert space. The dimension of a Hilbert space can be either finite or infinite. A real-valued finite-dimensional Hilbert space with dimension d is isomorphic to the Euclidean space $\mathbb{R}^d$. A Hilbert space with a countable dimension is termed a *separable* Hilbert space. We have the following bound on the fat-shattering dimension of linear classifiers in a Hilbert space, originally from Gurvits [1997].

Theorem 1.19. *Let $\mathcal{S}$ be a separable Hilbert space. Let $B > 0$ such that $\mathcal{X} \subseteq \mathbb{B}_B(\mathcal{S})$. Then the γ-fat-shattering dimension of $\mathcal{W}(\mathcal{S})$ is at most $\frac{B^2}{\gamma^2}$.*

Thus, we can conclude from Theorem 1.13 that a bounded zero-one loss (relative to the γ-margin loss) can be achieved by a MEM algorithm, using a sample of size $\tilde{O}(B^2/\gamma^2\epsilon^2)$. For Lipschitz losses, such as the hinge-loss and the ramp-loss, a sample complexity bound that depends on the average squared norm of the data can be derived using Rademacher complexities, as the following result shows [Bartlett and Mendelson, 2002].

Theorem 1.20. *Let $\mathcal{S}$ be a separable Hilbert space. Let ℓ be a c-Lipschitz loss function. Then for any distribution D over $\mathcal{S} \times \{\pm 1\}$, $\mathcal{R}_m(\mathcal{W}(\mathcal{S})_\ell, D) \leq \sqrt{\frac{c^2\mathbb{E}[\|x\|^2]}{m}}$, where the expectation is over the marginal of D on $\mathcal{S}$.*

This theorem, combined with Eq. (1.5), allows deriving sample complexity upper bounds for learning algorithms that minimize the hinge-loss or the ramp-loss. Consider first the hinge-loss—this is the loss that is minimized in soft-margin SVM [Cortes and Vapnik, 1995]. Since the hinge-loss with margin γ is $1/\gamma$-Lipschitz, we get a Rademacher complexity upper bound of $\sqrt{\frac{\mathbb{E}[\|x\|^2]}{m\gamma^2}}$. This implies that a sample size of $O(\mathbb{E}[\|x\|^2]/\gamma^2)$ suffices to achieve a small relative hinge-loss, compared to the best achievable hinge-loss. This can be done using an ERM algorithm for the hinge-loss, such as soft-margin SVM. Since the hinge-loss is an upper bound on the zero-one loss, this implies a guarantee also on the zero-one loss of the classifier emitted by the algorithm, although this guarantee is with respect to the best achievable hinge-loss.

For the ramp-loss, an even stronger result can be derived. The ramp-loss is also $1/\gamma$-Lipschitz, thus it has the same Rademacher complexity upper bound of $\sqrt{\frac{\mathbb{E}[\|x\|^2]}{m\gamma^2}}$. By Eq. (1.3), it follows that a sample size of $O(\mathbb{E}[\|x\|^2]/\gamma^2)$ suffices so that all linear classifiers in the unit ball have their ramp-loss on the distribution not much larger than their ramp-loss on the sample. The ramp-loss is lower-bounded by the zero-one loss and upper-bounded by the margin-loss ℓ_γ. Therefore, we can conclude that a sample size of $O(\mathbb{E}[\|x\|^2]/\gamma^2)$ suffices so that all linear classifiers in the unit ball have their zero-one loss on the distribution not much larger than their γ-margin loss on the sample. It follows that a MEM algorithm will also require only that many examples to achieve a low zero-one loss relative to the best margin loss. The following lower bound shows that this guarantee is tight.

Theorem 1.21. *There are constants C, c such that the following holds. Let $\mathcal{S}$ be a separable Hilbert space of infinite dimension. Let $B > 0$. For any learning algorithm $\mathcal{A}$ and any integer m, there is a distribution D over $\mathbb{B}_B(\mathcal{S}) \times \{\pm 1\}$ such that for any $\delta \leq c$,*

$$\ell_{0/1}(\mathcal{A}, D, m, \delta) - \ell^*_\gamma(\mathcal{W}(\mathcal{S}), D) \geq \sqrt{\frac{C \cdot B^2}{\gamma^2 m}}.$$

Proof. Assume to the contrary that there exists an algorithm $\mathcal{A}$ such that for all distributions D over $\mathbb{B}_B(\mathcal{S}) \times \{\pm 1\}$,

$$\ell_{0/1}(\mathcal{A}, D, m, \delta) - \ell^*_\gamma(\mathcal{W}(\mathcal{S}), D) < \sqrt{\frac{C \cdot B^2}{\gamma^2 m}}.$$

Let $x_i = B \cdot e_i \in \mathbb{B}_B(\mathcal{S})$ where e_i is the i'th vector in an orthonormal basis for $\mathcal{S}$. The set $X = \{x_1, \ldots, x_n\}$, for $n = \lfloor \frac{B^2}{\gamma^2} \rfloor$, is γ-shattered by $\mathcal{W}(\mathcal{S})$, since for any labeling $y_1, \ldots, y_n \in \{\pm 1\}^n$, we can set $w = \frac{\gamma}{B} \sum_{i=1}^n y_i e_i$, and we get that for all $i \in [n]$, $y_i \langle w, x_i \rangle = \gamma$. In addition, $\|w\| = \frac{\gamma}{B}\sqrt{n} \leq 1$, hence $h_w \in \mathcal{W}(\mathcal{S})$. It follows that for any distribution D with support in $X \times \{\pm 1\}$, we have $\ell^*_\gamma(\mathcal{W}(\mathcal{S}), D) \leq \ell^*_{0/1}(\mathcal{W}(\mathcal{S}), D)$. Thus we have

$$\ell_{0/1}(\mathcal{A}, D, m, \delta) - \ell^*_{0/1}(\mathcal{W}(\mathcal{S}), D) < \sqrt{\frac{C \cdot B^2}{\gamma^2 m}} \leq \sqrt{\frac{C \cdot (n+1)}{m}}.$$

But the set X is shattered by $\mathrm{sign} \circ \mathcal{H}$, thus the VC-dimension of $\mathcal{H}_{|X}$ is n. Therefore, by Theorem 1.8, there exists a universal constant C such that no algorithm can satisfy this inequality. We have thus reached a contradiction. □

Thus we can conclude the following: The distribution-free sample-complexity of learning homogeneous linear classifiers in the unit ball is proportional to d, and can be achieved by an ERM for the zero-one loss. The distribution-free sample-complexity with respect to the margin loss is proportional to $\min(d, \frac{B^2}{\gamma^2})$, and can be achieved by an MEM algorithm. Thus, in high or infinite dimensions, the sample-complexity of binary classification might be prohibitive, while the sample complexity of margin learning can be reasonable.

In this work we do not specifically address Hilbert spaces, and work for convenience with linear separators in $\mathbb{R}^d$. However, our results are not specific to Euclidean spaces, and they can be easily adapted to general separable Hilbert spaces.

1.7 Main Contributions

As we have shown in the previous sections, many complexity measures allow bounding the sample complexity of various hypothesis classes and algorithms. These complexity measures are usually used to provide upper-bounds for the sample complexity of a specific hypothesis class. These upper bounds typically hold for a large class of distributions. For instance, consider homogeneous linear classifiers in the unit ball, in the Euclidean space $\mathbb{R}^d$. As shown in Section 1.6, the distribution-free sample complexity of learning with this class, for any data distribution, is proportional to d. In addition, the sample complexity upper bound of large-margin classification is proportional to B^2/γ^2, where B^2 is the average squared norm of the data and γ is the size of the margin.

Such upper bounds can be useful for understanding the positive aspects of a learning rule. But it is difficult to understand the deficiencies of a learning rule, or to compare between different rules, based on upper bounds alone. This is because it is possible, and is often the case, that the true sample complexity for a given data distribution is much lower than the bound.

As we have shown above, some sample complexity upper bounds are known to be tight or to have an almost-matching lower bound. For instance, the VC-dimension lower bound in Theorem 1.8 shows that there exists a distribution in the class covered by the VC-dimension upper bound, for which this bound is tight. This holds in particular for linear classifiers in the unit ball. The lower bound for margin learning of linear classifiers, in Theorem 1.21, shows a similar result for the margin-based upper bound.

These results show that there cannot be a better upper bound that holds for the same class of distributions that the upper bound covers. But they do not imply that the upper bound characterizes the true sample complexity for any *specific* distribution in the class, except for the ones for which the upper bound is indeed tight. For instance, although the sample-complexity upper bound of $O(B^2/\gamma^2)$ for margin-learning is tight, Theorem 1.6 and Theorem 1.17 imply that if the distribution is supported by a low-dimensional sub-space, then the true number of examples required to reach a low error is much smaller.

In the first part of this thesis, our goal is to identify a simple quantity, which is a function of the distribution, that *does* precisely characterize the sample complexity of learning this distribution under a specific learning rule. We focus on the popular rule of Margin Error Minimization (MEM), defined in Section 1.4.3, and on the class of homogeneous linear classifiers. We present a new quantity, termed the *margin-adapted dimension*, and use it to provide a tighter distribution-dependent upper bound, and a matching distribution-dependent lower bound, for MEM algorithms for linear classifiers. The upper bound is universal, and the lower bound holds for a rich class of distributions.

The margin-adapted dimension, which we denote by k_γ for a margin of γ, refines both the dimension and the average norm of the data distribution, and can be easily calculated from the covariance matrix and the mean of the distribution. We provide a sample-complexity upper bound showing that $\tilde{O}(\frac{k_\gamma}{\epsilon^2})$ examples suffice in order to learn any distribution with a margin-adapted dimension of k_γ. We then show that for a rich family of 'light tailed' distributions, q the number of samples required for learning by minimizing the margin error is also lower-bounded by $\Omega(k_\gamma)$.

Our lower bound hinges on several new results:

- We relate fat-shattering to hardness of learning using MEM.
- We show that for a convex hypothesis class, fat-shattering is equivalent to shattering with exact margins.
- We link the fat-shattering of a set of vectors with the eigenvalues of the dot-product matrix (the Gram matrix) of the vectors in the set.
- We provide a new lower bound for the smallest eigenvalue of a random Gram matrix generated by sub-Gaussian variables. This bound extends previous results in analysis of random matrices.

Some of the results in this part have appeared in Sabato et al. [2010b].

As mentioned above, complexity measures of hypothesis classes are typically analyzed on a case-by-case basis. For instance, the complexity of the class of linear classifiers has been analyzed as a function of parameters such as the dimension of the ambient space and the maximal norm of the separator. In the second part of this thesis, we consider the useful setting of *Multiple Instance Learning*, and propose a generic analysis for this setting, that holds across many different hypothesis classes.

Multiple-Instance Learning (MIL), first introduced in Dietterich et al. [1997], is a special type of a supervised classification problem. As in classical supervised classification, in MIL the learner receives a sample of labeled examples drawn i.i.d. from an arbitrary and unknown distribution, and its objective is to discover a classification rule with a small expected classification error over the same distribution. In MIL additional structure is assumed, whereby the examples are received as *bags* of *instances*, such that each bag is composed of several instances. It is assumed that each instance has a true label, however the learner only observes the labels of the bags, which is is determined by the hidden labels of the instances via some function which is known a-priori. Classical works on MIL assume that the function is the Boolean OR. In this work we consider a more general setting which allows other functions as well.

We propose a formal framework for generalized MIL, which allows analyzing any MIL problem as a function of the underlying hypothesis class: : This is the hypothesis class of the possible mappings from single instances to labels. In addition, the analysis depends on the function determining the bag labels based on the instance labels. We provide a generic analysis that bounds the complexity of learning a MIL problem based on the complexity of learning the underlying hypothesis class. Our main contributions are:

- Bounding the sample complexity of MIL as a function of the complexity of the underlying hypothesis class. We provide bounds for the following complexity measures.
    - VC-dimension
    - Pseudo-dimension
    - Covering numbers
    - Fat-shattering dimension
    - Rademacher complexity
- A generic learning algorithm, which operates by using a regular supervised learning algorithm for the underlying hypothesis class as an oracle. The algorithm is computationally efficient if the oracle is an efficient learner in the agnostic setting.
- We present and analyze a setting in which MIL can be used to improve the sample complexity of non-MIL learning, by constructing artificial bags.

Some of these results have appeared in Sabato and Tishby [2009], Sabato et al. [2010a].

To make this dissertation coherent and due to the lack of space, some of my research work was omitted from this thesis. For example, I have worked on the generalization ability of the Information Bottleneck method [Shamir et al., 2010, 2008], on multiclass learnability [Daniely et al., 2011] and on active learning [Gonen et al., 2011].

Part I

Margin Learning

Chapter 2

Introduction (Part I)

In this part we pursue a tight characterization of the sample complexity of learning a classifier under a particular data distribution, and using a particular learning rule. Specifically, we treat the case where the data domain is $\mathcal{X} = \mathbb{R}^d$, and the hypothesis class is the homogeneous classifiers in the unit ball, $\mathcal{H} = \mathcal{W}(\mathbb{R}^d)$. We obtain a tight distribution-specific characterization of the sample complexity of large-margin learning.

Denote by $m(\epsilon, \gamma, D)$ the number of examples required to achieve an excess error of no more than ϵ relative to the best possible γ-margin error for a specific distribution D, using a MEM algorithm. Our main result shows that for a rich family of 'light-tailed' distributions,

$$\Omega(k_\gamma(D)) \leq m(\epsilon, \gamma, D) \leq \tilde{O}\left(\frac{k_\gamma(D)}{\epsilon^2}\right).$$

The upper bound is in fact universal and holds for any distribution, while the lower bound holds for a family of distributions that we define below.

As can be seen in this bound, we do not tightly characterize the dependence of the sample complexity on the desired error [as done e.g. in Steinwart and Scovel, 2007], thus our bounds are not tight for asymptotically small error levels. Our results are most significant if the desired error level is a constant well below chance but bounded away from zero. This is in contrast to classical statistical asymptotics that are also typically tight, but are valid only for very small ϵ. As was recently shown by Liang and Srebro 2010, the sample complexity for very small ϵ (in the classical statistical asymptotic regime) depends on quantities that can be very different from those that control the sample complexity for moderate error rates, which are more relevant for machine learning.

Our tight characterization, and in particular the distribution-specific lower bound on the sample complexity that we establish, can be used to compare large-margin (L_2 regularized) learning to other learning rules. We provide two such examples: we use our lower bound to rigorously establish a sample complexity gap between L_1 and L_2 regularization previously studied in Ng [2004], and to show a large gap between discriminative and generative learning on a Gaussian-mixture distribution.

We start by discussing related work in Section 2.1. We then present the problem setting and notation in Section 2.2. We introduce the margin-adapted dimension in Section 2.3. The sample-complexity upper bound is proved in Chapter 3. Chapter 4 is dedicated to the proof of the lower bound. In Chapter 5 we give examples of implications, and also show that any non-trivial sample-complexity lower bound for more general distributions must employ properties other than the covariance matrix of the distribution.

2.1 Related Work

Most work on "sample complexity lower bounds" is directed at proving that under some set of assumptions, there exists a data distribution for which one needs at least a certain number of examples to learn with required error and confidence [for instance Antos and Lugosi, 1998, Ehrenfeucht et al., 1988, Gentile and Helmbold, 1998]. This type of a lower bound does not, however, indicate much on the sample complexity of other distributions under the same set of assumptions.

For distribution-specific lower bounds, the classical analysis of Vapnik [Vapnik, 1995, Theorem 16.6] provides not only sufficient but also necessary conditions for the learnability of a hypothesis class with respect to a specific distribution. The essential condition is that the metric entropy of the hypothesis class with respect to the distribution be sub-linear in the limit of an infinite sample size. In some sense, this criterion can be seen as providing a "lower bound" on learnability for a specific distribution. However, we are interested in finite-sample convergence rates, and would like those to depend on simple properties of the distribution. The asymptotic arguments involved in Vapnik's general learnability claim do not lend themselves easily to such analysis.

Benedek and Itai [1991] show that if the distribution is known to the learner, a specific hypothesis class is learnable if and only if there is a finite ϵ-cover of this hypothesis class with respect to the distribution. Ben-David et al. [2008] consider a similar setting, and prove sample complexity lower bounds for learning with any data distribution, for some binary hypothesis classes on the real line. Vayatis and Azencott [1999] provide distribution-specific sample complexity upper bounds for hypothesis classes with a limited VC-dimension, as a function of how balanced the hypotheses are with respect to the considered distributions. These bounds are not tight for all distributions, thus they also do not fully characterize the distribution-specific sample complexity. Caramanis and Mannor [2007] provide lower bounds for the margin error of separating hyperplanes on nearly log-concave distributions. These bounds are related to the fact that such distributions do not satisfy large-margin separation. In contrast, our bounds hold for all distributions, including ones that can be separated with zero margin error.

2.2 Problem setting and definitions

In this chapter we consider the domain $\mathcal{X} = \mathbb{R}^d$, and the function class of linear separators with a unit norm, $\mathcal{H} = \mathcal{W}(\mathbb{R}^d)$. We write simply w to denote the function $h_w = \langle w, x\rangle$ for some $w \in \mathbb{R}^d$. Our goal is to bound the zero-one loss $\ell_{0/1}$. We consider MEM algorithms relative to the margin loss ℓ_γ for some $\gamma > 0$. We denote such an algorithm for a margin of γ by $\mathcal{A}_\gamma$. For a distribution D over $\mathcal{X} \times \{\pm 1\}$, we denote by D_X the marginal distribution of D on $\mathcal{X}$.

The distribution-specific sample complexity for MEM algorithms is defined as follows:

Definition 2.1 (Distribution-specific sample complexity)**.** *For* $\gamma > 0$, $\epsilon, \delta \in [0, 1]$, *and a distribution* D, *the* distribution-specific sample complexity, *denoted by* $m(\epsilon, \gamma, D, \delta)$, *is the minimal sample size such that for any MEM algorithm* $\mathcal{A}$, *and for any* $m \geq m(\epsilon, \gamma, D, \delta)$,

$$\ell_{0/1}(\mathcal{A}_\gamma, D, m, \delta) - \ell^*_\gamma(\mathcal{H}, D) \leq \epsilon.$$

Note that while we are considering a specific distribution, we require that *all* possible MEM algorithms do well on this distribution. This is because we are interested in the MEM strategy in general, and thus we study the guarantees that can be provided regardless of the specific MEM implementation.

In the rest of the chapter we write simply $\ell^*_\gamma(D)$ and omit the fixed term $\mathcal{H}$. We also sometimes omit δ and write simply $m(\epsilon, \gamma, D)$, indicating that δ is assumed to be some fixed small constant.

2.3 The margin-adapted dimension

The sample complexity of MEM for linear classifiers with unit norm can be upper-bounded in terms of the average norm relative to the margin $\mathbb{E}[\|X\|^2]/\gamma^2$, or alternatively in terms of the dimensionality d (see Section 1.6). Although both of these bounds are tight in the worst-case sense, i.e., they are the best bounds that rely only on the norm or only on the dimensionality respectively, neither is tight in a distribution-specific sense: If the average norm is unbounded while the dimension is small, then there can be an arbitrarily large gap between the true distribution-dependent sample complexity and the bound that depends on the average norm. If the converse holds, that is, the dimension is arbitrarily large while the average-norm is bounded, then the dimensionality bound is loose.

Seeking a tight distribution-specific analysis, one simple approach to tighten these bounds is to consider their minimum, which is proportional to $\min(d, \mathbb{E}[\|X\|^2]/\gamma^2)$. Trivially, this is an upper bound on the sample complexity as well. However, this simple combination is also not tight: Consider a distribution in which there are a few directions with very high variance, but the combined variance in all other directions is small (see Figure 2.1). We will show that in such situations the

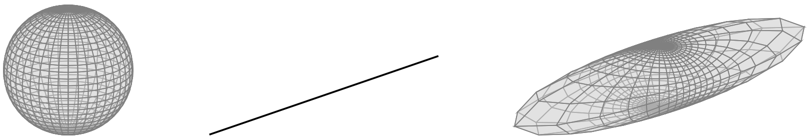

Figure 2.1: Illustrating covariance matrix ellipsoids. left: norm bound is tight; middle: dimension bound is tight; right: neither bound is tight.

sample complexity is characterized not by the minimum of dimension and norm, but by the sum of the number of high-variance dimensions and the average squared norm in the other directions. This behavior is captured by the *margin-adapted dimension*. We define it using the following property of a distribution.

Definition 2.2. *Let $b > 0$ and let k be a positive integer. A distribution D_X over $\mathbb{R}^d$ is* (b, k)-limited *if there exists a sub-space $V \subseteq \mathbb{R}^d$ of dimension $d - k$ such that $\mathbb{E}_{X \sim D_X}[\|\mathbb{O}_V \cdot X\|^2] \leq b$, where $\mathbb{O}_V$ is an orthogonal projection onto V.*

Definition 2.3 (margin-adapted dimension)**.** *The* margin-adapted dimension *of a distribution D_X, denoted by $k_\gamma(D_X)$, is the minimum k such that the distribution is $(\gamma^2 k, k)$-limited.*

We sometimes drop the argument of k_γ when it is clear from context. It is easy to see that for any distribution D_X over $\mathbb{R}^d$, $k_\gamma(D_X) \leq \min(d, \mathbb{E}[\|X\|^2]/\gamma^2)$. Moreover, k_γ can be much smaller than this minimum. For example, consider a random vector $X \in \mathbb{R}^{1001}$ with mean zero and statistically independent coordinates, such that the variance of the first coordinate is 1000, and the variance in each remaining coordinate is 0.001. We have $k_1 = 1$ but $d = \mathbb{E}[\|X\|^2] = 1001$.

$k_\gamma(D_X)$ can be calculated from the uncentered covariance matrix $\mathbb{E}_{X \sim D_X}[XX^T]$ as follows: Let $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_d \geq 0$ be the eigenvalues of this matrix. Then

$$k_\gamma = \min\{k \mid \sum_{i=k+1}^{d} \lambda_i \leq \gamma^2 k\}. \tag{2.1}$$

A quantity similar to this definition of k_γ was studied previously in Bousquet [2002]. The eigenvalues of the *empirical* covariance matrix were used to provide sample complexity bounds, for instance in Schölkopf et al. [1999]. However, k_γ generates a different type of bound, since it is defined based on the eigenvalues of the distribution and not of the sample. We will see that for small finite samples, the latter can be quite different from the former.

Chapter 3

A Distribution-Dependent Upper Bound

In this chapter we prove an upper bound on the sample complexity of learning with MEM. To do that, we will use the ramp-loss ramp_γ which was defined in Section 1.3. We show uniform convergence of the training error and test error with respect to this loss. The ramp-loss is lower-bounded by the zero-one loss and upper-bounded by the margin loss. Thus, the uniform convergence result will allow us to bound the true zero-one loss of MEM as a function of the best margin error on the distribution. We denote

$$\textsc{ramp}_\gamma = \mathcal{H}_{\mathrm{ramp}_\gamma} = \{(x,y) \mapsto \mathrm{ramp}_\gamma(w,x,y) \mid w \in \mathbb{B}_1^d\}.$$

We will show uniform convergence over $\textsc{ramp}_\gamma$ by bounding the Rademacher complexity of this class as a function of the data distribution. We will bound $\mathcal{R}_m(\textsc{ramp}_\gamma, D)$ on any (B^2, k)-limited distribution, by restating the functions in $\textsc{ramp}_\gamma$ as sums of two functions, each selected from a function class with bounded complexity. The first function class will be bounded because of the norm bound on the subspace V, and the second function class will have a bounded pseudo-dimension. However, the second function class will depend on the choice of the first function in the sum. Therefore, we require the following lemma, which allows combining covering numbers of different function classes. We use the notion of a *Hausdorff distance* between two sets $\mathcal{G}_1, \mathcal{G}_2 \subseteq \mathcal{X}$, defined as $\Delta_H(\mathcal{G}_1, \mathcal{G}_2) = \sup_{g_1 \in \mathcal{G}_1} \inf_{g_2 \in \mathcal{G}_2} \|g_1 - g_2\|_\circ$.

Lemma 3.1. *Let $(\mathcal{X}, \|\cdot\|_\circ)$ be a normed space. Let $\mathcal{F} \subseteq \mathcal{X}$ be a set, and let $\mathcal{G} : \mathcal{X} \to 2^{\mathcal{X}}$ be a mapping from objects in $\mathcal{X}$ to sets of objects in $\mathcal{X}$. Assume that $\mathcal{G}$ is c-Lipschitz with respect to the Hausdorff distance on sets, that is*

$$\forall f_1, f_2 \in \mathcal{X}, \Delta_H(\mathcal{G}(f_1), \mathcal{G}(f_2)) \leq c\|f_1 - f_2\|_\circ.$$

Let $\mathcal{F}_\mathcal{G} = \{f + g \mid f \in \mathcal{F}, g \in \mathcal{G}(f)\}$*. Then*

$$\mathcal{N}(\eta, \mathcal{F}_\mathcal{G}, \circ) \leq \mathcal{N}(\eta/(2+c), \mathcal{F}, \circ) \cdot \sup_{f \in \mathcal{F}} \mathcal{N}(\eta/(2+c), \mathcal{G}(f), \circ).$$

Proof. For any set $A \subseteq \mathcal{X}$, denote by $\mathcal{C}_A$ a minimal η-covering for A with respect to $\| \cdot \|_\circ$, so that $|\mathcal{C}_A| = \mathcal{N}(\eta, A, \circ)$. Let $f + g \in \mathcal{F}_\mathcal{G}$ such that $f \in \mathcal{F}, g \in \mathcal{G}(f)$. There is a $\hat{f} \in \mathcal{C}_\mathcal{F}$ such that $\|f - \hat{f}\|_\circ \leq \eta$. In addition, by the Lipschitz assumption there is a $\tilde{g} \in \mathcal{G}(\hat{f})$ such that $\|g - \tilde{g}\|_\circ \leq c\|f - \hat{f}\|_\circ \leq c\eta$. Lastly, there is a $\hat{g} \in \mathcal{C}_{\mathcal{G}(\hat{f})}$ such that $\|\tilde{g} - \hat{g}\|_\circ \leq \eta$. Therefore

$$\|f + g - (\hat{f} + \hat{g})\|_\circ \leq \|f - \hat{f}\|_\circ + \|g - \tilde{g}\|_\circ + \|\tilde{g} - \hat{g}\|_\circ \leq (2 + c)\eta.$$

Thus the set $\{f + g \mid f \in \mathcal{C}_\mathcal{F}, g \in \mathcal{C}_{\mathcal{G}(f)}\}$ is a $(2+c)\eta$ cover of $\mathcal{F}_\mathcal{G}$. The size of this cover is at most $|\mathcal{C}_\mathcal{F}| \cdot \sup_{f \in \mathcal{F}} |\mathcal{C}_{\mathcal{G}(f)}| \leq \mathcal{N}(\eta, \mathcal{F}, \circ) \cdot \sup_{f \in \mathcal{F}} \mathcal{N}(\eta, \mathcal{G}(f), \circ)$. □

The following lemma shows a useful class of mappings that are Lipschitz with respect to the Hausdorff distance as required by Lemma 3.1.

Lemma 3.2. *Let* $f : \mathcal{X} \to \mathbb{R}$ *be a function and let* $Z \subseteq \mathbb{R}^\mathcal{X}$ *be a function class over some domain* $\mathcal{X}$*. Let* $\mathcal{G} : \mathbb{R}^\mathcal{X} \to 2^{\mathbb{R}^\mathcal{X}}$ *be the function defined by*

$$\mathcal{G}(f) \triangleq \{x \mapsto [\![f(x) + z(x)]\!] - f(x) \mid z \in Z\}. \tag{3.1}$$

Then $\mathcal{G}$ *is* 1*-Lipschitz with respect to the Hausdorff distance.*

Proof. For a function $f : \mathcal{X} \to \mathbb{R}$ and a $z \in Z$, define the function $G[f, z]$ by

$$\forall x \in \mathcal{X}, \quad G[f, z](x) = [\![f(x) + z(x)]\!] - f(x).$$

Let $f_1, f_2 \in \mathbb{R}^\mathcal{X}$ be two functions, and let $g_1 = G[f_1, z] \in \mathcal{G}(f_1)$ for some $w_b \in \bar{V}$. Then, since $G[f_2, z] \in \mathcal{G}(f_2)$, we have $\inf_{g_2 \in \mathcal{G}(f_2)} \|g_1 - g_2\|_{L_2(S)} \leq \|G[f_1, z] - G[f_2, z]\|$. Now, for all $x \in \mathbb{R}$,

$$\begin{aligned} |G[f_1, z](x) - G[f_2, z](x)| &= |[\![f_1(x) + z(x)]\!] - f_1(x) - [\![f_2(x) + z(x)]\!] + f_2(x)| \\ &\leq |f_1(x) - f_2(x)|. \end{aligned}$$

Thus

$$\begin{aligned} \|G[f_1, z] - G[f_2, z]\|^2_{L_2(S)} &= \mathbb{E}_{X \sim S}(G[f_1, z](X) - G[f_2, z](X))^2 \\ &\leq \mathbb{E}_{X \sim S}(f_1(X) - f_2(X))^2 = \|f_1 - f_2\|^2_{L_2(S)}. \end{aligned}$$

It follows that $\inf_{g_2 \in \mathcal{G}(f_2)} \|g_1 - g_2\|_{L_2(S)} \leq \|f_1 - f_2\|_{L_2(S)}$. This holds for any $g_1 \in \mathcal{G}(f_1)$, thus $\Delta_H(\mathcal{G}(f_1), \mathcal{G}(f_2)) \leq \|f_1 - f_2\|_{L_2(S)}$. □

We will also require the following lemma, which uses the pseudo-dimension of a function class to bound the pseudo-dimension of a different class that is derived from it.

Lemma 3.3. *Let $f : \mathcal{X} \to \mathbb{R}$ be a function and let $Z \subseteq \mathbb{R}^{\mathcal{X}}$ be a function class over some domain $\mathcal{X}$. Let $\mathcal{G}(f)$ be defined as in Eq. (3.1). Then the pseudo-dimension of $\mathcal{G}(f)$ is at most the pseudo-dimension of Z.*

Proof. Let k be the pseudo-dimension of $\mathcal{G}(f)$, and let $\{x_1, \ldots, x_k\} \subseteq \mathcal{X}$ be a set which is pseudo-shattered by $\mathcal{G}(f)$. We show that the same set is pseudo-shattered by Z as well, thus proving the lemma. Since $\mathcal{G}(f)$ is pseudo-shattered, there exists a vector $r \in \mathbb{R}^k$ such that for all $y \in \{\pm 1\}^k$ there exists a $g_y \in \mathcal{G}(f)$ such that $\forall i \in [m], \operatorname{sign}(g_y(x_i) - r[i]) = y[i]$. Therefore for all $y \in \{\pm 1\}^k$ there exists a $z_y \in Z$ such that

$$\forall i \in [k], \operatorname{sign}(\llbracket f(x_i) + z_y(x_i) \rrbracket - f(x_i) - r[i]) = y[i].$$

By considering the case $y[i] = 1$, we have

$$0 < \llbracket f(x_i) + z_y(x_i) \rrbracket - f(x_i) - r[i] \leq 1 - f(x_i) - r[i].$$

By considering the case $y[i] = -1$, we have

$$0 > \llbracket f(x_i) + z_y(x_i) \rrbracket - f(x_i) - r[i] \geq -f(x_i) - r[i].$$

Therefore $0 < f(x_i) + r[i] < 1$. Now, let $y \in \{\pm 1\}^k$ and consider any $i \in [k]$. If $y[i] = 1$ then

$$\llbracket f(x_i) + z_y(x_i) \rrbracket - f(x_i) - r[i] > 0$$

It follows that

$$\llbracket f(x_i) + z_y(x_i) \rrbracket > f(x_i) + r[i] > 0,$$

thus

$$f(x_i) + z_y(x_i) > f(x_i) + r[i].$$

In other words, $\operatorname{sign}(z_y(x_i) - r[i]) = 1 = y[i]$. If $y[i] = -1$ then

$$\llbracket f(x_i) + z_y(x_i) \rrbracket - f(x_i) - r[i] < 0.$$

It follows that

$$[\![f(x_i) + z_y(x_i)]\!] < f(x_i) + r[i] < 1,$$

thus

$$f(x_i) + z_y(x_i) < f(x_i) + r[i].$$

in other words, $\mathrm{sign}(z_y(x_i) - r[i]) = -1 = y[i]$. We conclude that Z shatters $\{x_1, \ldots, x_k\}$ as well, using the same vector $r \in \mathbb{R}^k$. Thus the pseudo-dimension of Z is at least k. □

The bound on the Rademacher complexity of $\textsc{ramp}_\gamma$ is provided in the following theorem. We then state a corollary that uses Theorem 3.4 to derive a sample-complexity upper bound for MEM that depends only on k_γ.

Theorem 3.4. *Let D be a distribution over $\mathbb{R}^d \times \{\pm 1\}$, and assume D_X is (B^2, k)-limited. Then*

$$\mathcal{R}(\textsc{ramp}_\gamma, D) \leq \sqrt{\frac{O(k + B^2/\gamma^2)\ln(m)}{m}}.$$

Proof. In this proof all absolute constants are assumed to be positive and are denoted by C or C_i for some integer i. Their values may change from line to line or even within the same line.

Consider the distribution $\tilde{D}$ which results from drawing $(X, Y) \sim D$ and emitting $(Y \cdot X, 1)$. It too is (B^2, k)-limited, and $\mathcal{R}(\textsc{ramp}_\gamma, D) = \mathcal{R}(\textsc{ramp}_\gamma, \tilde{D})$. Therefore, we assume without loss of generality that for all (X, Y) drawn from D, $Y = 1$. Accordingly, we henceforth omit the y argument from $\mathrm{ramp}_\gamma(w, x, y)$ and write simply $\mathrm{ramp}_\gamma(w, x) \triangleq \mathrm{ramp}_\gamma(w, x, 1)$.

Let $\mathbb{O}_V$ be an orthogonal projection onto a sub-space V of dimension $d - k$ such that $\mathbb{E}_{X \sim D_X}[\|\mathbb{O}_V \cdot X\|^2] \leq B^2$. Let $\bar{V}$ be the complementary sub-space to V. Let $S = \{x_1, \ldots, x_m\} \subseteq \mathbb{R}^d$, and denote $B(S) = \sqrt{\mathbb{E}_{X \sim S}[\|\mathbb{O}_V \cdot X\|^2]}$. For a function $f : \mathbb{R}^d \to \mathbb{R}$, the $L_2(S)$ norm of f is $\|f\|_{L_2(S)} = \sqrt{\mathbb{E}_{X \sim S}[f(X)^2]}$.

We will bound the Rademacher complexity of $\textsc{ramp}$ by first bounding the covering number of $\textsc{ramp}_\gamma$ with respect to $L_2(S)$, and then using Eq. (1.10). To bound $\mathcal{N}(\eta, \textsc{ramp}_\gamma, L_2(S))$ for $\eta > 0$, note that $\mathrm{ramp}_\gamma(w, x) = [\![1 - \langle w, x\rangle/\gamma]\!] = 1 - [\![\langle w, x\rangle/\gamma]\!]$. Since shifting by a constant and negating do not change the covering number of a function class, $\mathcal{N}(\eta, \textsc{ramp}_\gamma, L_2(S))$ is equal to the covering number of $\{x \mapsto [\![\langle w, x\rangle/\gamma]\!] \mid w \in \mathbb{B}_1^d\}$. Moreover, let

$$\textsc{ramp}'_\gamma = \{x \mapsto [\![\langle w_a + w_b, x\rangle/\gamma]\!] \mid w_a \in \mathbb{B}_1^d \cap V, w_b \in \bar{V}\}.$$

Then $\{x \mapsto [\![\langle w, x\rangle/\gamma]\!] \mid w \in \mathbb{B}_1^d\} \subseteq \textsc{ramp}'_\gamma$, thus it suffices to bound the covering number of $\textsc{ramp}'_\gamma$. To do that, we show that $\textsc{ramp}'_\gamma$ satisfies the assumptions of Lemma 3.1 for the the space $(\mathbb{R}^{\mathbb{R}^d}, \|\cdot\|_{L_2(S)})$.

Let $\mathcal{F} = \{x \mapsto \langle w_a, x\rangle/\gamma \mid w_a \in \mathbb{B}_1^d \cap V\}$. Let $\mathcal{G} : \mathbb{R}^{\mathbb{R}^d} \to 2^{\mathbb{R}^{\mathbb{R}^d}}$ be the mapping defined by

$$\mathcal{G}(f) \triangleq \{x \mapsto [\![f(x) + \langle w_b, x\rangle/\gamma]\!] - f(x) \mid w_b \in \bar{V}\}.$$

Clearly, $\mathcal{F}_{\mathcal{G}} = \{f + g \mid f \in \mathcal{F}, g \in \mathcal{G}(f)\} = \text{RAMP}'_\gamma$. Furthermore, by Lemma 3.2, $\mathcal{G}$ is 1-Lipschitz as required by Lemma 3.1. Thus, by Lemma 3.1

$$\mathcal{N}(\eta, \text{RAMP}'_\gamma, L_2(S)) \le \mathcal{N}(\eta/3, \mathcal{F}, L_2(S)) \cdot \sup_{f \in \mathcal{F}} \mathcal{N}(\eta/3, \mathcal{G}(f), L_2(S)). \tag{3.2}$$

We now proceed to bound the two covering numbers on the right hand side. First, consider $\mathcal{N}(\eta/3, \mathcal{G}(f), L_2(S))$. By Lemma 3.3, the pseudo-dimension of $\mathcal{G}(f)$ is the same as the pseduo-dimension of $\{x \mapsto \langle w, x\rangle/\gamma \mid w \in \bar{V}\}$, which is exactly k, the dimension of $\bar{V}$. Therefore, by Eq. (1.2),

$$\mathcal{N}(\eta/3, \mathcal{G}(f), L_2(S)) \le C_1 \left(\frac{C_2}{\eta^2}\right)^k. \tag{3.3}$$

Second, consider $\mathcal{N}(\eta/3, \mathcal{F}, L_2(S))$. By Sudakov's minoration, stated in Eq. (1.7) for any $\eta > 0$

$$\ln \mathcal{N}(\eta, \mathcal{F}, L_2(S)) \le \frac{Cm}{\eta^2}\mathcal{G}^2(\mathcal{F}, S).$$

The right-hand side can be bounded as follows:

$$\begin{aligned}
\gamma \cdot m \cdot \mathcal{G}(\mathcal{F}, S) &= \gamma \cdot \mathbb{E}_s[\sup_{f \in \mathcal{F}} |\sum_{i=1}^m s_i f(x_i)|] = \mathbb{E}_s[\sup_{w \in \mathbb{B}_1^d \cap V} |\langle w, \sum_{i=1}^m s_i x_i\rangle|] \\
&\le \mathbb{E}_s[\| \sum_{i=1}^m s_i \mathbb{O}_V x_i \|] \le \sqrt{\mathbb{E}_s[\| \sum_{i=1}^m s_i \mathbb{O}_V x_i \|^2]} = \sqrt{\sum_{i \in [m]} \|\mathbb{O}_V x_i\|^2} = \sqrt{m} B(S).
\end{aligned}$$

Therefore $\ln \mathcal{N}(\eta, \mathcal{F}, L_2(S)) \le C\frac{B^2(S)}{\gamma^2\eta^2}$. Substituting this and Eq. (3.3) for the right-hand side in Eq. (3.2) and adjusting constants we get

$$\ln \mathcal{N}(\eta, \text{RAMP}_\gamma, L_2(S)) \le \ln \mathcal{N}(\eta, \text{RAMP}'_\gamma, L_2(S)) \le C_1(1 + k\ln(\frac{C_2}{\eta}) + \frac{B^2(S)}{\gamma^2\eta^2}),$$

To finalize the proof, we plug this inequality into Eq. (1.10) to get

$$\begin{aligned}
&\sqrt{m}\mathcal{R}(\textsc{ramp}_\gamma, S) \le C_1 \sum_{i\in[N]} \epsilon_{i-1}\sqrt{1 + k\ln(C_2/\epsilon_i) + \frac{B^2(S)}{\gamma^2\epsilon_i^2}} + 2\epsilon_N\sqrt{m} \\
&\le C_1\left(\sum_{i\in[N]} \epsilon_{i-1}\left(1 + \sqrt{k\ln(C_2/\epsilon_i)} + \sqrt{\frac{B^2(S)}{\gamma^2\epsilon_i^2}}\right)\right) + 2\epsilon_N\sqrt{m} \\
&= C_1\left(\sum_{i\in[N]} 2^{-i+1} + \sqrt{k}\sum_{i\in[N]} 2^{-i+1}\ln(C_2/2^{-i}) + \sum_{i\in[N]}\frac{B(S)}{\gamma}\right) + 2^{-N+1}\sqrt{m} \\
&\le C\left(1 + \sqrt{k} + \frac{B(S)\cdot N}{\gamma}\right) + 2^{-N+1}\sqrt{m}.
\end{aligned}$$

In the last inequality we used the fact that $\sum_i i2^{-i+1} \le 4$. Setting $N = \ln(2m)$ we get

$$\mathcal{R}(\textsc{ramp}_\gamma, S) \le \frac{C}{\sqrt{m}}\left(1 + \sqrt{k} + \frac{B(S)\ln(2m)}{\gamma}\right).$$

Taking expectation over both sides, and noting that $\mathbb{E}[B(S)] \le \sqrt{\mathbb{E}[B^2(S)]} \le B$, we get

$$\mathcal{R}(\textsc{ramp}_\gamma, S) \le \frac{C}{\sqrt{m}}(1 + \sqrt{k} + \frac{B\ln(2m)}{\gamma}) \le \sqrt{\frac{O(k + B^2\ln(2m)/\gamma^2)}{m}}.$$

□

Corollary 3.5 (Sample complexity upper bound)**.** *Let D be a distribution over $\mathbb{R}^d \times \{\pm 1\}$. Then*

$$m(\epsilon,\gamma,D) \le \tilde{O}\left(\frac{k_\gamma(D_X)}{\epsilon^2}\right).$$

Proof. Let $\mathcal{A}$ be a MEM algorithm, and let $w^* \in \operatorname{argmin}_{w\in\mathbb{B}_1^d} \ell_\gamma(w, D)$. By Eq. (1.4), with probability $1-\delta/2$

$$\operatorname{ramp}_\gamma(\mathcal{A}_\gamma(S), D) \le \operatorname{ramp}_\gamma(\mathcal{A}_\gamma(S), S) + 2\mathcal{R}_m(\textsc{ramp}_\gamma, D) + \sqrt{\frac{8\ln(2/\delta)}{m}}.$$

Set $h^* \in \mathcal{H}$ such that $\ell_\gamma(h^*, D) = \ell^*_\gamma(\mathcal{H}, D)$. We have

$$\operatorname{ramp}_\gamma(\mathcal{A}_\gamma(S), S) \le \ell_\gamma(\mathcal{A}_\gamma(S), S) \le \ell_\gamma(h^*, S).$$

The first inequality follows since the ramp loss is upper bounded by the margin loss. The second inequality follows since $\mathcal{A}$ is a MEM algorithm. Now, by Hoeffding's inequality, since the range of ramp_γ is in $[0, 1]$, with probability at least $1 - \delta/2$

$$\ell_\gamma(h^*, S) \leq \ell_\gamma(h^*, D) + \sqrt{\frac{\ln(2/\delta)}{2m}}.$$

It follows that with probability $1 - \delta$

$$\mathrm{ramp}_\gamma(\mathcal{A}_\gamma(S), D) \leq \ell^*_\gamma(\mathcal{H}, D) + 2\mathcal{R}_m(\textsc{ramp}_\gamma, D) + \sqrt{\frac{14\ln(2/\delta)}{m}}. \tag{3.4}$$

By definition of $k_\gamma(D_X)$, D_X is $(\gamma^2 k_\gamma, k_\gamma)$-limited. Therefore, by Theorem 3.4,

$$\mathcal{R}_m(\textsc{ramp}_\gamma, D) \leq \sqrt{\frac{O(k_\gamma(D_X))\ln(m)}{m}}.$$

In addition, $\ell_{0/1} \leq \mathrm{ramp}_\gamma$. Combining these with Eq. (3.4) we conclude that

$$\ell_{0/1}(\mathcal{A}_\gamma, D, m, \delta) \leq \ell^*_\gamma(\mathcal{H}, D) + \sqrt{\frac{O(k_\gamma(D_X)\ln(m) + \ln(1/\delta))}{m}}.$$

Bounding the second right-hand term by ϵ, we conclude that $m(\epsilon, \gamma, D) \leq \tilde{O}(k_\gamma/\epsilon^2)$. □

Chapter 4

A Distribution-Dependent Lower Bound

The new upper bound presented in Cor. 3.5 can be tighter than both the norm-only and the dimension-only upper bounds. But does the margin-adapted dimension characterize the true sample complexity of the distribution, or is it just another upper bound? To answer this question, we first need tools for deriving sample complexity lower bounds. Section 4.1 relates the smallest eigenvalue of a Gram-matrix to a lower bound on sample complexity. In Section 4.2 the family of sub-Gaussian product distributions is presented. We prove a sample-complexity lower bound for this family in Section 4.3.

4.1 A sample complexity lower bound with Gram-matrix eigenvalues

The ability to learn is closely related to the probability of a sample to be shattered, as evident in Vapnik's formulations of learnability as a function of the ϵ-entropy [Vapnik, 1995]. It is well known that the maximal size of a shattered set dictates a sample-complexity upper bound. We show that for some hypothesis classes it also implies a lower bound in Theorem 4.1 below. The theorem states that if a sample drawn from a data distribution is fat-shattered (see Def. 1.9) with a reasonably high probability, then MEM can fail to learn a good classifier for this distribution. We then relate the fat-shattering of a sample to the minimal eigenvalue of its Gram matrix. Therefore, a lower bound on the smallest eigenvalue of the Gram-matrix implies a lower-bound on the sample complexity. We say that a set is γ*-shattered at the origin* if it is γ-shattered when r in Def. 1.9 is set to the zero vector.

The following theorem shows that a high probability of γ-shattering implies hardness of margin learning. This holds not only for linear classifiers, but more generally for all *symmetric* hypothesis classes. Given a domain $\mathcal{X}$, we say that a hypothesis class $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ is symmetric if for all $h \in \mathcal{H}$, $-h \in \mathcal{H}$ as well. This clearly holds for the class of linear classifiers $\mathcal{H}$.

Theorem 4.1. *Let $\mathcal{X}$ be some domain, and assume that $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ is a symmetric hypothesis class. Let D be a distribution over $\mathcal{X} \times \{\pm 1\}$. If the probability of a sample of size m drawn from D_X^m to be γ-shattered at the origin is at least η, then $m(\epsilon, \gamma, D, \eta/2) \geq \lfloor m/2 \rfloor$ for all $\epsilon < 1/2 - \ell^*_\gamma(D)$.*

Proof. Let $\epsilon \leq \frac{1}{2} - \ell^*_\gamma(D)$. We show a MEM algorithm $\mathcal{A}$ such that

$$\ell_{0/1}(\mathcal{A}_\gamma, D, \lfloor m/2 \rfloor, \eta/2) \geq \frac{1}{2} > \ell^*_\gamma(D) + \epsilon,$$

thus proving the desired lower bound on $m(\epsilon, \gamma, D, \eta/2)$.

Assume for simplicity that m is even (otherwise replace m with $m - 1$). Consider two sets $S, \tilde{S} \subseteq \mathcal{X} \times \{\pm 1\}$, each of size $m/2$, such that $S_X \cup \tilde{S}_X$ is γ-shattered at the origin. Then there exists a hypothesis $h_1 \in \mathcal{H}$ such that the following holds:

- For all $x \in S_X \cup \tilde{S}_X$, $|h_1(x)| \geq \gamma$.
- For all $(x, y) \in S$, $\text{sign}(h_1(x)) = y$.
- For all $(x, y) \in \tilde{S}$, $\text{sign}(h_1(x)) = -y$.

It follows that $\ell_\gamma(h_1, S) = 0$. In addition, let $h_2 = -h_1$. We have $h_2 \in \mathcal{H}$ due to the symmetry of $\mathcal{H}$. It follows that $\ell_\gamma(h_2, \tilde{S}) = 0$. In addition, h_1 and h_2 never predict the same label. Thus $\ell_{0/1}(h_1, D) + \ell_{0/1}(h_2, D) \geq 1$. It follows that for at least one of $i \in \{1, 2\}$, we have $\ell_{0/1}(h_i, D) \geq \frac{1}{2}$. Denote the set of hypotheses with a high zero-one loss by

$$\mathcal{H}_\otimes = \{h \in \mathcal{H} \mid \ell_{0/1}(h, D) \geq \frac{1}{2}\}.$$

We have just shown that if $S_X \cup \tilde{S}_X$ is γ-shattered then at least one of the following holds: (1) $h_1 \in \mathcal{H}_\otimes \cap \text{argmin}_{h \in \mathcal{H}} \ell_\gamma(h, S)$ or (2) $h_2 \in \mathcal{H}_\otimes \cap \text{argmin}_{h \in \mathcal{H}} \ell_\gamma(h, \tilde{S})$.

Now, consider a MEM algorithm $\mathcal{A}$ such that whenever possible, it returns a hypothesis from $\mathcal{H}_\otimes$. Formally, given the input sample S, if $\mathcal{H}_\otimes \cap \text{argmin}_{h \in \mathcal{H}} \ell_\gamma(h, S) \neq \emptyset$, then $\mathcal{A}(S) \in \mathcal{H}_\otimes \cap \text{argmin}_{h \in \mathcal{H}} \ell_\gamma(h, S)$. It follows that

$$\begin{aligned}
&\mathbb{P}_{S \sim D^{m/2}}[\ell_{0/1}(\mathcal{A}(S), D) \geq \tfrac{1}{2}] \geq \mathbb{P}_{S \sim D^{m/2}}[\mathcal{H}_\otimes \cap \operatorname*{argmin}_{h \in \mathcal{H}} \ell_\gamma(h, S) \neq \emptyset] \\
&= \frac{1}{2}(\mathbb{P}_{S \sim D^{m/2}}[\mathcal{H}_\otimes \cap \operatorname*{argmin}_{h \in \mathcal{H}} \ell_\gamma(h, S) \neq \emptyset] + \mathbb{P}_{\tilde{S} \sim D^{m/2}}[\mathcal{H}_\otimes \cap \operatorname*{argmin}_{h \in \mathcal{H}} \ell_\gamma(h, \tilde{S}) \neq \emptyset]) \\
&\geq \frac{1}{2}(\mathbb{P}_{S, \tilde{S} \sim D^{m/2}}[\mathcal{H}_\otimes \cap \operatorname*{argmin}_{h \in \mathcal{H}} \ell_\gamma(h, S) \neq \emptyset \text{ OR } \mathcal{H}_\otimes \cap \operatorname*{argmin}_{h \in \mathcal{H}} \ell_\gamma(h, \tilde{S}) \neq \emptyset]) \\
&\geq \frac{1}{2}\mathbb{P}_{S, \tilde{S} \sim D^{m/2}}[S_X \cup \tilde{S}_X \text{ is } \gamma\text{-shattered at the origin }].
\end{aligned}$$

The last inequality follows from the argument above regarding h_1 and h_2. The last expression is simply half the probability that a sample of size m from D_X is shattered. By assumption, this probability is at least η. Thus we conclude that $\mathbb{P}_{S\sim D^{m/2}}[\ell_{0/1}(\mathcal{A}(S), D) \geq \frac{1}{2}] \geq \eta/2$. It follows that $\ell_{0/1}(\mathcal{A}_\gamma, D, m/2, \eta/2) \geq \frac{1}{2}$. □

As a side note, it is interesting to observe that Theorem 4.1 does not hold in general for non-symmetric hypothesis classes. For example, assume that the domain is $\mathcal{X} = [0, 1]$, and the hypothesis class is the set of all functions that label a finite number of points in $[0, 1]$ by $+1$ and the rest by -1. Consider the distribution which is uniform over $[0, 1]$ and labels all of the domain with -1. For any $m > 0$ and $\gamma \in (0, 1)$, a sample of size m is γ-shattered at the origin with probability 1. However, any learning algorithm that returns a hypothesis from the hypothesis class will incur zero error.

We now return to the case of homogeneous linear classifiers, and link high-probability fat-shattering to properties of the distribution. First, we provide a sufficient condition for the fat-shattering of a sample, based on the minimum eigenvalue of its Gram matrix. Theorem 4.2 stated below presents an equivalent and simpler characterization of fat-shattering for linear classifiers. We use it to prove the sufficient condition in Cor. 4.5.

Theorem 4.2. *Let $\mathbb{X} \in \mathbb{R}^{m\times d}$ be the matrix of a set of size m in $\mathbb{R}^d$. The set is γ-shattered at the origin if and only if $\mathbb{X}\mathbb{X}^T$ is invertible and for all $y \in \{\pm 1\}^m$, $y^T(\mathbb{X}\mathbb{X}^T)^{-1}y \leq \gamma^{-2}$.*

To prove Theorem 4.2 we require two auxiliary lemmas. The first lemma, stated below, allows substituting γ-shattering with shattering with exact γ-margins, by showing that the two notions are equivalent if the function class is convex.

Lemma 4.3. *Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ be a class of functions, and assume $\mathcal{F}$ is convex, that is*

$$\forall f_1, f_2 \in \mathcal{F}, \forall \lambda \in [0, 1], \quad \lambda f_1 + (1-\lambda)f_2 \in \mathcal{F}.$$

If $S = \{x_1, \dots, x_m\} \subseteq \mathcal{X}$ is γ-shattered by $\mathcal{F}$ with witness $r \in \mathbb{R}^m$, then for every $y \in \{\pm 1\}^m$ there is an $f \in \mathcal{F}$ such that for all $i \in [m]$, $y[i](f(x_i) - r[i]) = \gamma$.

The proof of this lemma is provided in Section 4.4.1. The second lemma allows converting the representation of the Gram-matrix to a different feature space while keeping the separation properties intact. For a matrix $\mathbb{M}$, $\mathbb{M}^+$ denotes its pseudo-inverse.

Lemma 4.4. *Let $\mathbb{X} \in \mathbb{R}^{m\times d}$ be a matrix such that $\mathbb{X}\mathbb{X}^T$ is invertible, and let $\mathbb{Y} \in \mathbb{R}^{m\times k}$ such that $\mathbb{X}\mathbb{X}^T = \mathbb{Y}\mathbb{Y}^T$. Let $r \in \mathbb{R}^m$ be some real vector. If there exists a vector $\widetilde{w} \in \mathbb{R}^k$ such that $\mathbb{Y}\widetilde{w} = r$, then there exists a vector $w \in \mathbb{R}^d$ such that $\mathbb{X}w = r$ and $\|w\| = \|\mathbb{Y}^T(\mathbb{Y}^T)^+\widetilde{w}\| \leq \|\widetilde{w}\|$.*

Proof. Denote $\mathbb{K} = \mathbb{X}\mathbb{X}^T = \mathbb{Y}\mathbb{Y}^T$. Let $\mathbb{S} = \mathbb{Y}^T\mathbb{K}^{-1}\mathbb{X}$ and let $w = \mathbb{S}^T\widetilde{w}$. We have $\mathbb{X}w = \mathbb{X}\mathbb{S}^T\widetilde{w} = \mathbb{X}\mathbb{X}^T\mathbb{K}^{-1}\mathbb{Y}\widetilde{w} = \mathbb{Y}\widetilde{w} = r$. In addition, $\|w\| = w^T w = \widetilde{w}^T\mathbb{S}\mathbb{S}^T\widetilde{w}$. By definition of $\mathbb{S}$,

$$\mathbb{S}\mathbb{S}^T = \mathbb{Y}^T\mathbb{K}^{-1}\mathbb{X}\mathbb{X}^T\mathbb{K}^{-1}\mathbb{Y} = \mathbb{Y}^T\mathbb{K}^{-1}\mathbb{Y} = \mathbb{Y}^T(\mathbb{Y}\mathbb{Y}^T)^{-1}\mathbb{Y} = \mathbb{Y}^T(\mathbb{Y}^T)^{+}.$$

Denote $\mathbb{O} = \mathbb{Y}^T(\mathbb{Y}^T)^{+}$. $\mathbb{O}$ is an orthogonal projection matrix: by the properties of the pseudo-inverse, $\mathbb{O} = \mathbb{O}^T$ and $\mathbb{O}^2 = \mathbb{O}$. Therefore $\|w\| = \widetilde{w}^T\mathbb{S}\mathbb{S}^T\widetilde{w} = \widetilde{w}^T\mathbb{O}\widetilde{w} = \widetilde{w}^T\mathbb{O}\mathbb{O}^T\widetilde{w} = \|\mathbb{O}\widetilde{w}\| \leq \|\widetilde{w}\|$. □

Proof of Theorem 4.2. We prove the theorem for 1-shattering. The case of γ-shattering follows by rescaling X appropriately. Let $\mathbb{X}\mathbb{X}^T = \mathbb{U}\Lambda\mathbb{U}^T$ be the SVD of $\mathbb{X}\mathbb{X}^T$, where $\mathbb{U}$ is an orthogonal matrix and Λ is a diagonal matrix. Let $\mathbb{Y} = \mathbb{U}\Lambda^{\frac{1}{2}}$. We have $\mathbb{X}\mathbb{X}^T = \mathbb{Y}\mathbb{Y}^T$. We show that the specified conditions are sufficient and necessary for the shattering of the set.

Sufficient: If $\mathbb{X}\mathbb{X}^T$ is invertible, then Λ is invertible, thus so is $\mathbb{Y}$. For any $y \in \{\pm 1\}^m$, Let $w_y = \mathbb{Y}^{-1}y$. Then $\mathbb{Y}w_y = y$. By Lemma 4.4, there exists a separator w such that $\mathbb{X}w = y$ and $\|w\| \leq \|w_y\| = \sqrt{y^T(\mathbb{Y}\mathbb{Y}^T)^{-1}y} = \sqrt{y^T(\mathbb{X}\mathbb{X}^T)^{-1}y} \leq 1$.

Necessary: If $\mathbb{X}\mathbb{X}^T$ is not invertible then the vectors in S are linearly dependent, thus S cannot be shattered using linear separators [see e.g. Vapnik, 1995]. The first condition is therefore necessary. Assume S is 1-shattered at the origin and show that the second condition necessarily holds. By Lemma 4.3, for all $y \in \{\pm 1\}^m$ there exists a $w_y \in \mathbb{B}_1^d$ such that $\mathbb{X}w_y = y$. Thus by Lemma 4.4 there exists a $\widetilde{w}_y$ such that $\mathbb{Y}\widetilde{w}_y = y$ and $\|\widetilde{w}_y\| \leq \|w_y\| \leq 1$. $\mathbb{X}\mathbb{X}^T$ is invertible, thus so is $\mathbb{Y}$. Therefore $\widetilde{w}_y = \mathbb{Y}^{-1}y$. Thus $y^T(\mathbb{X}\mathbb{X}^T)^{-1}y = y^T(\mathbb{Y}\mathbb{Y}^T)^{-1}y = \|\widetilde{w}_y\| \leq 1$. □

Corollary 4.5. *Let $\mathbb{X} \in \mathbb{R}^{m\times d}$ be the matrix of a set of size m in $\mathbb{R}^d$. If $\lambda_{\min}(\mathbb{X}\mathbb{X}^T) \geq m\gamma^2$ then the set is γ-shattered at the origin.*

Proof. If $\lambda_{\min}(\mathbb{X}\mathbb{X}^T) \geq m\gamma^2$ then $\mathbb{X}\mathbb{X}^T$ is invertible and $\lambda_{\max}((\mathbb{X}\mathbb{X}^T)^{-1}) \leq (m\gamma^2)^{-1}$. For any $y \in \{\pm 1\}^m$ we have $\|y\| = \sqrt{m}$ and

$$y^T(\mathbb{X}\mathbb{X}^T)^{-1}y \leq \|y\|^2\lambda_{\max}((\mathbb{X}\mathbb{X}^T)^{-1}) \leq m(m\gamma^2)^{-1} = \gamma^{-2}.$$

By Theorem 4.2 the sample is γ-shattered at the origin. □

Cor. 4.5 generalizes the requirement of linear independence for shattering with no margin: A set of vectors is shattered with no margin if the vectors are linearly independent, that is if $\lambda_{\min} > 0$. The corollary shows that for γ-fat-shattering, we can require instead $\lambda_{\min} \geq m\gamma^2$. We can now conclude the following theorem, which states that if it is highly probable that the smallest eigenvalue of the sample Gram matrix is large, then MEM might fail to learn a good classifier for the given distribution. Its proof is immediate by combining Theorem 4.1. and Cor. 4.5.

Theorem 4.6. *Let D be a distribution over $\mathbb{R}^d \times \{\pm 1\}$. Let $m > 0$ and let $\mathbb{X}$ be the matrix of a sample drawn from D_X^m. Let $\eta = \mathbb{P}[\lambda_{\min}(\mathbb{X}\mathbb{X}^T) \geq m\gamma^2]$. Then for all $\epsilon < 1/2 - \ell_\gamma^*(D)$, $m(\epsilon, \gamma, D, \eta/2) \geq \lfloor m/2 \rfloor$.*

Theorem 4.6 generalizes the case of learning a linear separator without a margin: If a sample of size $2m$ is linearly independent with high probability, then there is no hope of using m points to predict the label of the other points. The theorem extends this observation to the case of learning with a margin, by requiring a stronger condition than just linear independence of the points in the sample.

Recall that our upper-bound on the sample complexity from Chapter 3 is $\tilde{O}(k_\gamma)$. We now define the family of sub-Gaussian product distributions, and show that for this familym the lower bound that can be deduced from Theorem 4.6 is also linear in k_γ.

4.2 Sub-Gaussian distributions

In order to derive a lower bound on distribution-specific sample complexity in terms of the covariance of $X \sim D_X$, we must assume that X is not too heavy-tailed. This is because for any data distribution there exists another distribution which is almost identical and has the same sample complexity, but has arbitrarily large covariance values. This can be achieved by mixing the original distribution with a tiny probability for drawing a vector with a huge norm. We thus restrict the discussion to multidimensional sub-Gaussian distributions. This ensures light tails of the distribution in all directions, while still allowing a rich family of distributions, as we presently see. Sub-Gaussianity is defined for scalar random variables as follows.

Definition 4.7 (Sub-Gaussian random variables, see e.g. Buldygin and Kozachenko [1998])**.** *A random variable $X \in \mathbb{R}$ is* sub-Gaussian with moment B, *for $B \geq 0$, if*

$$\forall t \in \mathbb{R}, \quad \mathbb{E}[\exp(tX)] \leq \exp(t^2 B^2/2).$$

In this work we further say that X is sub-Gaussian with relative moment $\rho > 0$ *if X is sub-Gaussian with moment $\rho\sqrt{\mathbb{E}[X^2]}$, i.e.*

$$\forall t \in \mathbb{R}, \quad \mathbb{E}[\exp(tX)] \leq \exp(t^2 \rho^2 \mathbb{E}[X^2]/2).$$

Note that a sub-Gaussian variable with moment B and relative moment ρ is also sub-Gaussian with moment B' and relative moment ρ' for any $B' \geq B$ and $\rho' \geq \rho$.

The family of sub-Gaussian distributions is quite extensive: For instance, it includes any bounded, Gaussian, or Gaussian-mixture random variable with mean zero. Specifically, if X is a mean-zero Gaussian random variable, $X \sim N(0, \sigma^2)$, then X is sub-Gaussian with relative moment 1 and the inequalities in the definition above hold with equality. As another example, if X is a uniform random variable over $\{\pm b\}$ for some $b \geq 0$, then X is sub-Gaussian with relative moment 1, since

$$\mathbb{E}[\exp(tX)] = \frac{1}{2}(\exp(tb) + \exp(-tb)) \leq \exp(t^2 b^2/2) = \exp(t^2 \mathbb{E}[X^2]/2). \tag{4.1}$$

Let $\mathbb{B} \in \mathbb{R}^{d \times d}$ be a symmetric PSD matrix. A random vector $X \in \mathbb{R}^d$ is a *sub-Gaussian random vector* with moment matrix $\mathbb{B}$ if for all $u \in \mathbb{R}^d$, $\mathbb{E}[\exp(\langle u, X \rangle)] \leq \exp(\langle \mathbb{B}u, u \rangle/2)$. The following lemma provides a useful property of the norm of a sub-Gaussian random vector. The proof is given in Section 4.4.2.

Lemma 4.8. *Let $X \in \mathbb{R}^d$ be a sub-Gaussian random vector with moment matrix $\mathbb{B}$. Then for all $t \in (0, \frac{1}{4\lambda_{\max}(\mathbb{B})}]$, $\mathbb{E}[\exp(t\|X\|^2)] \leq \exp(2t \cdot \mathrm{trace}(\mathbb{B}))$.*

Our lower bound holds for the family of sub-Gaussian product distributions, defined as follows.

Definition 4.9 (Sub-Gaussian product distributions)**.** *A distribution D_X over $\mathbb{R}^d$ is a* sub-Gaussian product distribution *with moment B and relative moment ρ if there exists some orthonormal basis $a_1, \ldots, a_d \in \mathbb{R}^d$, such that for $X \sim D_X$, $\langle a_i, X \rangle$ are independent sub-Gaussian random variables, each with moment B and relative moment ρ.*

Note that a sub-Gaussian product distribution has mean zero, thus its covariance matrix is equal to its uncentered covariance matrix. For any fixed $\rho \geq 0$, we denote by $\mathcal{D}^{\mathrm{sg}}_{\rho}$ the family of all sub-Gaussian product distributions with relative moment ρ, in arbitrary dimension. For instance, all multivariate Gaussian distributions and all uniform distributions on the corners of a centered hyper-rectangle are in $\mathcal{D}^{\mathrm{sg}}_{1}$. All uniform distributions over a full centered hyper-rectangle are in $\mathcal{D}^{\mathrm{sg}}_{3/2}$. Note that if $\rho_1 \leq \rho_2$, $\mathcal{D}^{\mathrm{sg}}_{\rho_1} \subseteq \mathcal{D}^{\mathrm{sg}}_{\rho_2}$.

We provide a lower bound for all distributions in $\mathcal{D}^{\mathrm{sg}}_{\rho}$. This lower bound is linear in the margin-adapted dimension of the distribution, thus it matches the upper bound provided in Cor. 3.5. The constants in the lower bound depend only on the value of ρ, which we regard as a constant.

4.3 A sample-complexity lower bound for sub-Gaussian product distributions

As shown in Section 4.1, to obtain a sample complexity lower bound it suffices to have a lower bound on the value of the smallest eigenvalue of a random Gram matrix. The distribution of the smallest eigenvalue of a random Gram matrix has been investigated under various assumptions. The cleanest results are in the asymptotic case where the sample size and the dimension approach infinity, the ratio between them approaches a constant, and the coordinates of each example are identically distributed.

Theorem 4.10 (Bai and Silverstein 2010, Theorem 5.11)**.** *Let $\{\mathbb{X}_i\}_{i=1}^{\infty}$ be a series of matrices of sizes $m_i \times d_i$, whose entries are i.i.d. random variables with mean zero, variance σ^2 and finite fourth moments. If $\lim_{i\to\infty} \frac{m_i}{d_i} = \beta < 1$, then $\lim_{i\to\infty} \lambda_{\min}(\frac{1}{d_i}\mathbb{X}_i\mathbb{X}_i^T) = \sigma^2(1-\sqrt{\beta})^2$.*

This asymptotic limit can be used to approximate an asymptotic lower bound on $m(\epsilon, \gamma, D)$, if D_X is a product distribution of i.i.d. random variables with mean zero, variance σ^2, and finite fourth moment. Let $\mathbb{X} \in \mathbb{R}^{m\times d}$ be the matrix of a sample of size m drawn from D_X. We can find $m = m_\circ$ such that $\lambda_{m_\circ}(\mathbb{X}\mathbb{X}^T) \approx \gamma^2 m_\circ$, and use Theorem 4.6 to conclude that $m(\epsilon, \gamma, D) \geq m_\circ/2$. If d and m are large enough, we have by Theorem 4.10 that for $\mathbb{X}$ drawn from D_X^m:

$$\lambda_{\min}(\mathbb{X}\mathbb{X}^T) \approx d\sigma^2(1-\sqrt{m/d})^2 = \sigma^2(\sqrt{d}-\sqrt{m})^2.$$

Solving the equality $\sigma^2(\sqrt{d}-\sqrt{m_\circ})^2 = m_\circ\gamma^2$ we get $m_\circ = d/(1+\gamma/\sigma)^2$. The margin-adapted dimension for D_X is $k_\gamma \approx d/(1+\gamma^2/\sigma^2)$, thus $\frac{1}{2}k_\gamma \leq m_\circ \leq k_\gamma$. In this case, then, the sample complexity lower bound is indeed the same order as k_γ, which controls also the upper bound in Cor. 3.5. However, this is an asymptotic analysis, which holds for a highly limited set of distributions. Moreover, since Theorem 4.10 holds asymptotically for each distribution separately, we cannot use it to deduce a uniform finite-sample lower bound for families of distributions.

For our analysis we require *finite-sample* bounds for the smallest eigenvalue of a random Gram-matrix. Rudelson and Vershynin [2009, 2008] provide such finite-sample lower bounds for distributions which are products of identically distributed sub-Gaussians. In Theorem 4.11 below we provide a new and more general result, which holds for any sub-Gaussian product distribution. The proof of Theorem 4.11 is provided in Section 4.4.3. Combining Theorem 4.11 with Theorem 4.6 above we prove the lower bound, stated in Theorem 4.12 below.

Theorem 4.11. *For any $\rho > 0$ and $\delta \in (0,1)$ there are $\beta > 0$ and $C > 0$ such that the following holds. For any $D_X \in \mathcal{D}_\rho^{sg}$ with covariance matrix $\Sigma \leq I$, and for any $m \leq \beta \cdot \mathrm{trace}(\Sigma) - C$, if $\mathbb{X}$ is the $m \times d$ matrix of a sample drawn from D_X^m, then $\mathbb{P}[\lambda_{\min}(\mathbb{X}\mathbb{X}^T) \geq m] \geq \delta$.*

Theorem 4.12 (Sample complexity lower bound for distributions in $\mathcal{D}^{\text{sg}}_\rho$). *For any $\rho > 0$ there are constants $\beta > 0, C \geq 0$ such that for any D with $D_X \in \mathcal{D}^{sg}_\rho$, for any $\gamma > 0$ and for any $\epsilon < \frac{1}{2} - \ell^*_\gamma(D)$, $m(\epsilon, \gamma, D, 1/4) \geq \beta k_\gamma(D_X) - C$.*

Proof. Assume w.l.o.g. that the orthonormal basis $a_1, \dots, a_d$ of independent sub-Gaussian directions of D_X, defined in Def. 4.9, is the natural basis $e_1, \dots, e_d$. Define $\lambda_i = \mathbb{E}_{X \sim D_X}[X[i]^2]$, and assume w.l.o.g. $\lambda_1 \geq \dots \geq \lambda_d > 0$. Let $\mathbb{X}$ be the $m \times d$ matrix of a sample drawn from D^m_X. Fix $\delta \in (0, 1)$, and let β and C be the constants for ρ and δ in Theorem 4.11. Throughout this proof we abbreviate $k_\gamma \triangleq k_\gamma(D_X)$. Let $m \leq \beta(k_\gamma - 1) - C$. We would like to use Theorem 4.11 to bound $\lambda_{\text{min}}(\mathbb{X}\mathbb{X}^T)$ with high probability, so that Theorem 4.6 can be applied to get the desired lower bound. However, Theorem 4.11 holds only if $\Sigma \leq I$. Thus we split to two cases—one in which the dimensionality controls the lower bound, and one in which the norm controls it. The split is based on the value of λ_{k_γ}.

Case I Assume $\lambda_{k_\gamma} \geq \gamma^2$. Then $\forall i \in [k_\gamma], \lambda_i \geq \gamma^2$. By our assumptions on D_X, for all $i \in [d]$ the random variable $X[i]$ is sub-Gaussian with relative moment ρ. Consider the random variables $Z[i] = X[i]/\sqrt{\lambda_i}$ for $i \in [k_\gamma]$. $Z[i]$ is also sub-Gaussian with relative moment ρ, and $\mathbb{E}[Z[i]^2] = 1$. Consider the product distribution of $Z[1], \dots, Z[k_\gamma]$, and let Σ' be its covariance matrix. We have $\Sigma' = I_{k_\gamma}$, and $\text{trace}(\Sigma') = k_\gamma$. Let $\mathbb{Z}$ be the matrix of a sample of size m drawn from this distribution. By Theorem 4.11, $\mathbb{P}[\lambda_{\text{min}}(\mathbb{Z}\mathbb{Z}^T) \geq m] \geq \delta$, which is equivalent to

$$\mathbb{P}[\lambda_{\text{min}}(\mathbb{X} \cdot \text{diag}(1/\lambda_1, \dots, 1/\lambda_{k_\gamma}, 0, \dots, 0) \cdot \mathbb{X}^T) \geq m] \geq \delta.$$

Since $\forall i \in [k_\gamma], \lambda_i \geq \gamma^2$, we have $\mathbb{P}[\lambda_{\text{min}}(\mathbb{X}\mathbb{X}^T) \geq m\gamma^2] \geq \delta$.

Case II Assume $\lambda_{k_\gamma} < \gamma^2$. Then $\lambda_i < \gamma^2$ for all $i \in \{k_\gamma, \dots, d\}$. Consider the random variables $Z[i] = X[i]/\gamma$ for $i \in \{k_\gamma, \dots, d\}$. $Z[i]$ is sub-Gaussian with relative moment ρ and $\mathbb{E}[Z[i]^2] \leq 1$. Consider the product distribution of $Z[k_\gamma], \dots, Z[d]$, and let Σ' be its covariance matrix. We have $\Sigma' < I_{d-k_\gamma+1}$. By the minimality in Eq. (2.1) we also have $\text{trace}(\Sigma') = \frac{1}{\gamma^2} \sum_{i=k_\gamma}^{d} \lambda_i \geq k_\gamma - 1$. Let $\mathbb{Z}$ be the matrix of a sample of size m drawn from this product distribution. By Theorem 4.11, $\mathbb{P}[\lambda_{\text{min}}(\mathbb{Z}\mathbb{Z}^T) \geq m] \geq \delta$. Equivalently,

$$\mathbb{P}[\lambda_{\text{min}}(\mathbb{X} \cdot \text{diag}(0, \dots, 0, 1/\gamma^2, \dots, 1/\gamma^2) \cdot \mathbb{X}^T) \geq m] \geq \delta,$$

therefore $\mathbb{P}[\lambda_{\text{min}}(\mathbb{X}\mathbb{X}^T) \geq m\gamma^2] \geq \delta$.

In both cases $\mathbb{P}[\lambda_{\text{min}}(\mathbb{X}\mathbb{X}^T) \geq m\gamma^2] \geq \delta$. This holds for any $m \leq \beta(k_\gamma - 1) - C$, thus by Theorem 4.6 $m(\epsilon, \gamma, D, \delta/2) \geq \lfloor(\beta(k_\gamma - 1) - C)/2\rfloor$ for $\epsilon < 1/2 - \ell^*_\gamma(D)$. We finalize the proof by setting $\delta = \frac{1}{2}$ and adjusting β and C. □

4.4 Proofs

4.4.1 Proof of Lemma 4.3

To prove Lemma 4.3, we first prove the following lemma. Denote by conv(A) the convex hull of a set A.

Lemma 4.13. *Let $\gamma > 0$. For each $y \in \{\pm 1\}^m$, select $r_y \in \mathbb{R}^d$ such that for all $i \in [m]$, $r_y[i]y[i] \geq \gamma$. Let $R = \{r_y \in \mathbb{R}^m \mid y \in \{\pm 1\}^m\}$. Then $\{\pm\gamma\}^m \subseteq$ conv(R).*

Proof. We will prove the claim by induction on the dimension m.

Base case For $m = 1$, we have $R = \{a, b\} \subseteq \mathbb{R}$ where $a \leq -\gamma$ and $b \geq \gamma$. Clearly, conv$(R) = [a, b]$, and $\pm\gamma \in [a, b]$.

Inductive step Assume the lemma holds for $m-1$. For a vector $t \in \mathbb{R}^m$, denote by $\bar{t}$ its projection $(t[1], \ldots, t[m-1])$ on $\mathbb{R}^{m-1}$. Similarly, for a set of vectors $S \subseteq \mathbb{R}^m$, let $\bar{S} = \{\bar{s} \mid s \in S\} \subseteq \mathbb{R}^{m-1}$. Define $Y_+ = \{\pm 1\}^{m-1} \times \{+1\}$ and $Y_- = \{\pm 1\}^{m-1} \times \{-1\}$. Let $R_+ = \{r_y \mid y \in Y_+\}$, and similarly for R_-. Then the induction hypothesis holds for $\bar{R}_+$ and $\bar{R}_-$ with dimension $m - 1$. Let $z \in \{\pm\gamma\}^m$. We wish to prove $z \in$ conv(R). From the induction hypothesis we have $\bar{z} \in$ conv$(\bar{R}_+)$ and $\bar{z} \in$ conv$(\bar{R}_-)$. Thus, for all $y \in \{\pm 1\}$ there exist $\alpha_y, \beta_y \geq 0$ such that $\sum_{y \in Y_+} \alpha_y = \sum_{y \in Y_-} \beta_y = 1$, and

$$\bar{z} = \sum_{y \in Y_+} \alpha_y \bar{r}_y = \sum_{y \in Y_-} \beta_y \bar{r}_y.$$

Let $z_a = \sum_{y \in Y_+} \alpha_y r_y$ and $z_b = \sum_{y \in Y_-} \beta_y r_y$ We have that $\forall y \in Y_+, r_y[m] \geq \gamma$, and $\forall y \in Y_-, r_y[m] \leq -\gamma$. Therefore, $z_b[m] \leq -\gamma \leq z[m] \leq \gamma \leq z_a[m]$. In addition, $\bar{z}_a = \bar{z}_b = \bar{z}$. Select $\lambda \in [0, 1]$ such that $z[m] = \lambda z_a[m] + (1 - \lambda) z_b[m]$, then $z = \lambda z_a + (1 - \lambda) z_b$. Since $z_a, z_b \in$ conv(R), we have $z \in$ conv(R). □

Proof of Lemma 4.3. Denote by $f(S)$ the vector $(f(x_1), \ldots, f(x_m))$. Recall that $r \in \mathbb{R}^m$ is the witness for the shattering of S, and let

$$L = \{f(S) - r \mid f \in \mathcal{F}\} \subseteq \mathbb{R}^m.$$

Since S is shattered, for any $y \in \{\pm 1\}^m$ there is an $r_y \in L$ such that $\forall i \in [m], r_y[i]y[i] \geq \gamma$. By Lemma 4.13, $\{\pm\gamma\}^m \subseteq$ conv(L). Since $\mathcal{F}$ is convex, L is also convex. Therefore $\{\pm\gamma\}^m \subseteq L$. □

4.4.2 Proof of Lemma 4.8

Proof of Lemma 4.8. It suffices to consider diagonal moment matrices: If $\mathbb{B}$ is not diagonal, let $\mathbb{V} \in \mathbb{R}^{d \times d}$ be an orthogonal matrix such that $\mathbb{V}\mathbb{B}\mathbb{V}^T$ is diagonal, and let $Y = \mathbb{V}X$. We have $\mathbb{E}[\exp(t\|Y\|^2)] = \mathbb{E}[\exp(t\|X\|^2)]$ and $\text{trace}(\mathbb{V}\mathbb{B}\mathbb{V}^T) = \text{trace}(\mathbb{B})$. In addition, for all $u \in \mathbb{R}^d$,

$$\mathbb{E}[\exp(\langle u, Y\rangle)] = \mathbb{E}[\exp(\langle \mathbb{V}^T u, X\rangle)] \leq \\ \exp(\frac{1}{2}\langle \mathbb{B}\mathbb{V}^T u, \mathbb{V}^T u\rangle) = \exp(\frac{1}{2}\langle \mathbb{V}\mathbb{B}\mathbb{V}^T u, u\rangle).$$

Therefore Y is sub-Gaussian with the diagonal moment matrix $\mathbb{V}\mathbb{B}\mathbb{V}^T$. Thus assume w.l.o.g. that $\mathbb{B} = \text{diag}(\lambda_1, \ldots, \lambda_d)$ where $\lambda_1 \geq \ldots \geq \lambda_d \geq 0$.

We have $\exp(t\|X\|^2) = \prod_{i \in [d]} \exp(tX[i]^2)$. In addition, for any $t > 0$ and $x \in \mathbb{R}$, $2\sqrt{\Pi t} \cdot \exp(tx^2) = \int_{-\infty}^{\infty} \exp(sx - \frac{s^2}{4t})ds$. Therefore, for any $u \in \mathbb{R}^d$,

$$\begin{aligned}
(2\sqrt{\Pi t})^d \cdot \mathbb{E}[\exp(t\|X\|^2)] &= \mathbb{E}\left[\prod_{i\in[d]} \int_{-\infty}^{\infty} \exp(u[i]X[i] - \frac{u[i]^2}{4t})du[i]\right] \\
&= \mathbb{E}\left[\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \prod_{i\in[d]} \exp(u[i]X[i] - \frac{u[i]^2}{4t})du[i]\right] \\
&= \mathbb{E}\left[\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \exp(\langle u, X\rangle - \frac{\|u\|^2}{4t}) \prod_{i\in[d]} du[i]\right] \\
&= \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \mathbb{E}[\exp(\langle u, X\rangle)] \exp(-\frac{\|u\|^2}{4t}) \prod_{i\in[d]} du[i]
\end{aligned}$$

By the sub-Gaussianity of X, the last expression is bounded by

$$\begin{aligned}
&\leq \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \exp(\frac{1}{2}\langle \mathbb{B}u, u\rangle - \frac{\|u\|^2}{4t}) \prod_{i\in[d]} du[i] \\
&= \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \prod_{i\in[d]} \exp(\frac{\lambda_i u[i]^2}{2} - \frac{u[i]^2}{4t})du[i] \\
&= \prod_{i\in[d]} \int_{-\infty}^{\infty} \exp(u[i]^2(\frac{\lambda_i}{2} - \frac{1}{4t}))du[i] = \Pi^{d/2}(\prod_{i\in[d]}(\frac{1}{4t} - \frac{\lambda_i}{2}))^{-\frac{1}{2}}.
\end{aligned}$$

The last equality follows from the fact that for any $a > 0$, $\int_{-\infty}^{\infty} \exp(-a \cdot s^2)ds = \sqrt{\Pi/a}$, and from the assumption $t \leq \frac{1}{4\lambda_1}$. We conclude that

$$\mathbb{E}[\exp(t\|X\|^2)] \leq (\prod_{i\in[d]}(1-2\lambda_i t))^{-\frac{1}{2}} \leq \exp(2t\cdot\sum_{i=1}^{d}\lambda_i) = \exp(2t\cdot\mathrm{trace}(\mathbb{B})),$$

where the second inequality holds since $\forall x \in [0,1]$, $(1-x/2)^{-1} \leq \exp(x)$. □

4.4.3 Proof of Theorem 4.11

In the proof of Theorem 4.11 we use the fact $\lambda_{\min}(\mathbb{X}\mathbb{X}^T) = \inf_{\|x\|_2=1}\|\mathbb{X}^T x\|^2$ and bound the right-hand side via an ϵ-net of the unit sphere in $\mathbb{R}^m$, denoted by $S^{m-1} \triangleq \{x \in \mathbb{R}^m \mid \|x\|_2 = 1\}$. An ϵ-net of the unit sphere is a set $C \subseteq S^{m-1}$ such that $\forall x \in S^{m-1}, \exists x' \in C, \|x - x'\| \leq \epsilon$. Denote the minimal size of an ϵ-net for S^{m-1} by $\mathcal{N}_m(\epsilon)$, and by $\mathcal{C}_m(\epsilon)$ a minimal ϵ-net of S^{m-1}, so that $\mathcal{C}_m(\epsilon) \subseteq S^{m-1}$ and $|\mathcal{C}_m(\epsilon)| = \mathcal{N}_m(\epsilon)$. The proof of Theorem 4.11 requires several lemmas. First we prove a concentration result for the norm of a matrix defined by sub-Gaussian variables. Then we bound the probability that the squared norm of a vector is small.

Lemma 4.14. *Let $\mathbb{Y}$ be a $d \times m$ matrix with $m \leq d$, such that $\mathbb{Y}_{ij}$ are independent sub-Gaussian variables with moment B. Let Σ be a diagonal $d \times d$ PSD matrix such that $\Sigma \leq I$. Then for all $t \geq 0$ and $\epsilon \in (0,1)$,*

$$\mathbb{P}[\|\sqrt{\Sigma}\mathbb{Y}\| \geq t] \leq \mathcal{N}_m(\epsilon)\exp(\frac{\mathrm{trace}(\Sigma)}{2} - \frac{t^2(1-\epsilon)^2}{4B^2}).$$

Proof. We have $\|\sqrt{\Sigma}\mathbb{Y}\| \leq \max_{x\in\mathcal{C}_m(\epsilon)}\|\sqrt{\Sigma}\mathbb{Y}x\|/(1-\epsilon)$, see for instance in Bennett et al. [1975]. Therefore,

$$\mathbb{P}[\|\sqrt{\Sigma}\mathbb{Y}\| \geq t] \leq \sum_{x\in\mathcal{C}_m(\epsilon)}\mathbb{P}[\|\sqrt{\Sigma}\mathbb{Y}x\| \geq (1-\epsilon)t]. \tag{4.2}$$

Fix $x \in \mathcal{C}_m(\epsilon)$. Let $V = \sqrt{\Sigma}\mathbb{Y}x$, and assume $\Sigma = \mathrm{diag}(\lambda_1,\ldots,\lambda_d)$. For $u \in \mathbb{R}^d$,

$$\begin{aligned}\mathbb{E}[\exp(\langle u, V\rangle)] &= \mathbb{E}[\exp(\sum_{i\in[d]}u_i\sqrt{\lambda_i}\sum_{j\in[m]}\mathbb{Y}_{ij}x_j)] = \prod_{j,i}\mathbb{E}[\exp(u_i\sqrt{\lambda_i}\mathbb{Y}_{ij}x_j)]\\ &\leq \prod_{j,i}\exp(u_i^2\lambda_i B^2 x_j^2/2) = \exp(\frac{B^2}{2}\sum_{i\in[d]}u_i^2\lambda_i\sum_{j\in[m]}x_j^2)\\ &= \exp(\frac{B^2}{2}\sum_{i\in[d]}u_i^2\lambda_i) = \exp(\langle B^2\Sigma u, u\rangle/2).\end{aligned}$$

Thus V is a sub-Gaussian vector with moment matrix $B^2\Sigma$. Let $s = 1/(4B^2)$. Since $\Sigma \leq I$, we have $s \leq 1/(4B^2 \max_{i\in[d]} \lambda_i)$. Therefore, by Lemma 4.8,

$$\mathbb{E}[\exp(s\|V\|^2)] \leq \exp(2sB^2 \operatorname{trace}(\Sigma)).$$

By Chernoff's method, $\mathbb{P}[\|V\|^2 \geq z^2] \leq \mathbb{E}[\exp(s\|V\|^2)]/\exp(sz^2)$. Thus

$$\mathbb{P}[\|V\|^2 \geq z^2] \leq \exp(2sB^2 \operatorname{trace}(\Sigma) - sz^2) = \exp(\frac{\operatorname{trace}(\Sigma)}{2} - \frac{z^2}{4B^2}).$$

Set $z = t(1-\epsilon)$. Then for all $x \in S^{m-1}$

$$\mathbb{P}[\|\sqrt{\Sigma}\mathbb{Y}x\| \geq t(1-\epsilon)] = \mathbb{P}[\|V\| \geq t(1-\epsilon)] \leq \exp(\frac{\operatorname{trace}(\Sigma)}{2} - \frac{t^2(1-\epsilon)^2}{4B^2}).$$

Therefore, by Eq. (4.2),

$$\mathbb{P}[\|\sqrt{\Sigma}\mathbb{Y}\| \geq t] \leq \mathcal{N}_m(\epsilon)\exp(\frac{\operatorname{trace}(\Sigma)}{2} - \frac{t^2(1-\epsilon)^2}{4B^2}).$$

□

Lemma 4.15. *Let $\mathbb{Y}$ be a $d \times m$ matrix with $m \leq d$, such that $\mathbb{Y}_{ij}$ are independent centered random variables with variance 1 and fourth moments at most B. Let Σ be a diagonal $d \times d$ PSD matrix such that $\Sigma \leq I$. There exist $\alpha > 0$ and $\eta \in (0,1)$ that depend only on B such that for any $x \in S^{m-1}$*

$$\mathbb{P}[\|\sqrt{\Sigma}\mathbb{Y}x\|^2 \leq \alpha \cdot (\operatorname{trace}(\Sigma) - 1)] \leq \eta^{\operatorname{trace}(\Sigma)}.$$

To prove Lemma 4.15 we require Lemma 4.16 [Rudelson and Vershynin, 2008, Lemma 2.2] and Lemma 4.17, which extends Lemma 2.6 in the same work.

Lemma 4.16. *Let $T_1, \ldots, T_n$ be independent non-negative random variables. Assume that there are $\theta > 0$ and $\mu \in (0,1)$ such that for any i, $\mathbb{P}[T_i \leq \theta] \leq \mu$. There are $\alpha > 0$ and $\eta \in (0,1)$ that depend only on θ and μ such that $\mathbb{P}[\sum_{i=1}^n T_i < \alpha n] \leq \eta^n$.*

Lemma 4.17. *Let $\mathbb{Y}$ be a $d \times m$ matrix with $m \leq d$, such that the columns of $\mathbb{Y}$ are i.i.d. random vectors. Assume further that $\mathbb{Y}_{ij}$ are centered, and have a variance of 1 and a fourth moment at most B. Let Σ be a diagonal $d \times d$ PSD matrix. Then for all $x \in S^{m-1}$, $\mathbb{P}[\|\sqrt{\Sigma}\mathbb{Y}x\| \leq \sqrt{\operatorname{trace}(\Sigma)/2}] \leq 1 - 1/(196B)$.*

Proof. Let $x \in S^{m-1}$, and $T_i = (\sum_{j=1}^m \mathbb{Y}_{ij}x_j)^2$. Let $\lambda_1, \ldots, \lambda_d$ be the values on the diagonal of Σ, and let $T_\Sigma = \|\sqrt{\Sigma}\mathbb{Y}x\|^2 = \sum_{i=1}^d \lambda_i T_i$. First, since $\mathbb{E}[\mathbb{Y}_{ij}] = 0$ and $\mathbb{E}[\mathbb{Y}_{ij}] = 1$ for all i, j, we have

$$\mathbb{E}[T_i] = \sum_{i \in [m]} x_j^2 \mathbb{E}[\mathbb{Y}_{ij}^2] = \|x\|^2 = 1.$$

Therefore $\mathbb{E}[T_\Sigma] = \mathrm{trace}(\Sigma)$. Second, since $\mathbb{Y}_{i1}, \ldots, \mathbb{Y}_{im}$ are independent and centered, we have [Ledoux and Talagrand, 1991, Lemma 6.3]

$$\mathbb{E}[T_i^2] = \mathbb{E}[(\sum_{j \in [m]} \mathbb{Y}_{ij} x_j)^4] \le 16 \mathbb{E}_\sigma[(\sum_{j \in [m]} \sigma_j \mathbb{Y}_{ij} x_j)^4],$$

where $\sigma_1, \ldots, \sigma_m$ are independent uniform $\{\pm 1\}$ variables. Now, by Khinchine's inequality [Nazarov and Podkorytov, 2000],

$$\mathbb{E}_\sigma[(\sum_{j \in [m]} \sigma_j \mathbb{Y}_{ij} x_j)^4] \le 3\mathbb{E}[(\sum_{j \in [m]} \mathbb{Y}_{ij}^2 x_j^2)^2] = 3 \sum_{j,k \in [m]} x_j^2 x_k^2 \mathbb{E}[\mathbb{Y}_{ij}^2]\mathbb{E}[\mathbb{Y}_{ik}^2].$$

Now $\mathbb{E}[\mathbb{Y}_{ij}^2]\mathbb{E}[\mathbb{Y}_{ik}^2] \le \sqrt{\mathbb{E}[\mathbb{Y}_{ij}^4]\mathbb{E}[\mathbb{Y}_{ik}^4]} \le B$. Thus $\mathbb{E}[T_i^2] \le 48B \sum_{j,k \in [m]} x_j^2 x_k^2 = 48B\|x\|^4 = 48B$. Thus,

$$\mathbb{E}[T_\Sigma^2] = \mathbb{E}[(\sum_{i=1}^{d} \lambda_i T_i)^2] = \sum_{i,j=1}^{d} \lambda_i \lambda_j \mathbb{E}[T_i T_j]$$
$$\le \sum_{i,j=1}^{d} \lambda_i \lambda_j \sqrt{\mathbb{E}[T_i^2]\mathbb{E}[T_j^2]} \le 48B(\sum_{i=1}^{d} \lambda_i)^2 = 48B \cdot \mathrm{trace}(\Sigma)^2.$$

By the Paley-Zigmund inequality [Paley and Zygmund, 1932], for $\theta \in [0,1]$

$$\mathbb{P}[T_\Sigma \ge \theta \mathbb{E}[T_\Sigma]] \ge (1-\theta)^2 \frac{\mathbb{E}[T_\Sigma]^2}{\mathbb{E}[T_\Sigma^2]} \ge \frac{(1-\theta)^2}{48B}.$$

Therefore, setting $\theta = 1/2$, we get $\mathbb{P}[T_\Sigma \le \mathrm{trace}(\Sigma)/2] \le 1 - 1/(196B)$. □

Proof of Lemma 4.15. Let $\lambda_1, \ldots, \lambda_d \in [0,1]$ be the values on the diagonal of Σ. Consider a partition $Z_1, \ldots, Z_k$ of $[d]$, and denote $L_j = \sum_{i \in Z_j} \lambda_i$. There exists such a partition such that for all $j \in [k]$, $L_j \le 1$, and for all $j \in [k-1]$, $L_j > \frac{1}{2}$. Let $\Sigma[j]$ be the sub-matrix of Σ that includes the rows and columns whose indexes are in Z_j. Let $\mathbb{Y}[j]$ be the sub-matrix of $\mathbb{Y}$ that includes the rows in Z_j. Denote $T_j = \|\sqrt{\Sigma[j]}\mathbb{Y}[j]x\|^2$. Then

$$\|\sqrt{\Sigma}\mathbb{Y}x\|^2 = \sum_{j \in [k]} \sum_{i \in Z_j} \lambda_i (\sum_{j=1}^{m} \mathbb{Y}_{ij} x_j)^2 = \sum_{j \in [k]} T_j.$$

We have $\mathrm{trace}(\Sigma) = \sum_{i=1}^d \lambda_i \geq \sum_{j\in[k-1]} L_j \geq \frac{1}{2}(k-1)$. In addition, $L_j \leq 1$ for all $j \in [k]$. Thus $\mathrm{trace}(\Sigma) \leq k \leq 2\,\mathrm{trace}(\Sigma)+1$. For all $j \in [k-1]$, $L_j \geq \frac{1}{2}$, thus by Lemma 4.17, $\mathbb{P}[T_j \leq 1/4] \leq 1 - 1/(196B)$. Therefore, by Lemma 4.16 there are $\alpha > 0$ and $\eta \in (0,1)$ that depend only on B such that

$$\begin{aligned}&\mathbb{P}[\|\sqrt{\Sigma}\mathbb{Y}x\|^2 < \alpha\cdot(\mathrm{trace}(\Sigma)-1)] \leq \mathbb{P}[\|\sqrt{\Sigma}\mathbb{Y}x\|^2 < \alpha(k-1)]\\ &\quad= \mathbb{P}[\sum_{j\in[k]} T_j < \alpha(k-1)] \leq \mathbb{P}[\sum_{j\in[k-1]} T_j < \alpha(k-1)] \leq \eta^{k-1} \leq \eta^{2\,\mathrm{trace}(\Sigma)}.\end{aligned}$$

The lemma follows by substituting η for η^2. □

Proof of Theorem 4.11. We have

$$\sqrt{\lambda_{\min}(\mathbb{X}\mathbb{X}^T)} = \inf_{x\in S^{m-1}} \|\mathbb{X}^T x\| \geq \min_{x\in\mathcal{C}_m(\epsilon)} \|\mathbb{X}^T x\| - \epsilon\|\mathbb{X}^T\|. \tag{4.3}$$

For brevity, denote $L = \mathrm{trace}(\Sigma)$. Assume $L \geq 2$. Let $m \leq L\cdot\min(1,(c-K\epsilon)^2)$ where c, K, ϵ are constants that will be set later such that $c - K\epsilon > 0$. By Eq. (4.3)

$$\begin{aligned}&\mathbb{P}[\lambda_{\min}(\mathbb{X}\mathbb{X}^T) \leq m] \leq \mathbb{P}[\lambda_{\min}(\mathbb{X}\mathbb{X}^T) \leq (c-K\epsilon)^2 L]\\ &\quad\leq \mathbb{P}[\min_{x\in\mathcal{C}_m(\epsilon)} \|\mathbb{X}^T x\| - \epsilon\|\mathbb{X}^T\| \leq (c-K\epsilon)\sqrt{L}] \qquad (4.4)\\ &\quad\leq \mathbb{P}[\|\mathbb{X}^T\| \geq K\sqrt{L}] + \mathbb{P}[\min_{x\in\mathcal{C}_m(\epsilon)} \|\mathbb{X}^T x\| \leq c\sqrt{L}]. \qquad (4.5)\end{aligned}$$

The last inequality holds since the inequality in line (4.4) implies at least one of the inequalities in line (4.5). We will now upper-bound each of the terms in line (4.5). We assume w.l.o.g. that Σ is not singular (since zero rows and columns can be removed from $\mathbb{X}$ without changing $\lambda_{\min}(\mathbb{X}\mathbb{X}^T)$). Define $\mathbb{Y} \triangleq \sqrt{\Sigma^{-1}}\mathbb{X}^T$. Note that $\mathbb{Y}_{ij}$ are independent sub-Gaussian variables with (absolute) moment ρ. To bound the first term in line (4.5), note that by Lemma 4.14, for any $K > 0$,

$$\mathbb{P}[\|\mathbb{X}^T\| \geq K\sqrt{L}] = \mathbb{P}[\|\sqrt{\Sigma}\mathbb{Y}\| \geq K\sqrt{L}] \leq \mathcal{N}_m(\frac{1}{2})\exp(L(\frac{1}{2} - \frac{K^2}{16\rho^2})).$$

By Rudelson and Vershynin [2009], Proposition 2.1, for all $\epsilon \in [0,1]$, $\mathcal{N}_m(\epsilon) \leq 2m(1+\frac{2}{\epsilon})^{m-1}$. Therefore

$$\mathbb{P}[\|\mathbb{X}^T\| \geq K\sqrt{L}] \leq 2m5^{m-1}\exp(L(\frac{1}{2} - \frac{K^2}{16\rho^2})).$$

Let $K^2 = 16\rho^2(\frac{3}{2} + \ln(5) + \ln(2/\delta))$. Recall that by assumption $m \le L$, and $L \ge 2$. Therefore

$$\begin{aligned}\mathbb{P}[\|\mathbb{X}^T\| \ge K\sqrt{L}] &\le 2m5^{m-1}\exp(-L(1+\ln(5)+\ln(2/\delta)))\\ &\le 2L5^{L-1}\exp(-L(1+\ln(5)+\ln(2/\delta))).\end{aligned}$$

Since $L \ge 2$, we have $2L\exp(-L) \le 1$. Therefore

$$\mathbb{P}[\|\mathbb{X}^T\| \ge K\sqrt{L}] \le 2L\exp(-L-\ln(2/\delta)) \le \exp(-\ln(2/\delta)) = \frac{\delta}{2}. \tag{4.6}$$

To bound the second term in line (4.5), since $\mathbb{Y}_{ij}$ are sub-Gaussian with moment ρ, $\mathbb{E}[\mathbb{Y}_{ij}^4] \le 5\rho^4$ [Buldygin and Kozachenko, 1998, Lemma 1.4]. Thus, by Lemma 4.15, there are $\alpha > 0$ and $\eta \in (0,1)$ that depend only on ρ such that for all $x \in S^{m-1}$, $\mathbb{P}[\|\sqrt{\Sigma}\mathbb{Y}x\|^2 \le \alpha(L-1)] \le \eta^L$. Set $c = \sqrt{\alpha/2}$. Since $L \ge 2$, we have $c\sqrt{L} \le \sqrt{\alpha(L-1)}$. Thus

$$\begin{aligned}\mathbb{P}[\min_{x\in\mathcal{C}_m(\epsilon)} \|\mathbb{X}^T x\| \le c\sqrt{L}] &\le \sum_{x\in\mathcal{C}_m(\epsilon)} \mathbb{P}[\|\mathbb{X}^T x\| \le c\sqrt{L}]\\ &\le \sum_{x\in\mathcal{C}_m(\epsilon)} \mathbb{P}[\|\sqrt{\Sigma}\mathbb{Y}x\| \le \sqrt{\alpha(L-1)}] \le \mathcal{N}_m(\epsilon)\eta^L.\end{aligned}$$

Let $\epsilon = c/(2K)$, so that $c - K\epsilon > 0$. Let $\theta = \min(\frac{1}{2}, \frac{\ln(1/\eta)}{2\ln(1+2/\epsilon)})$. Set $L_\circ$ such that $\forall L \ge L_\circ$, $L \ge \frac{2\ln(2/\delta)+2\ln(L)}{\ln(1/\eta)}$. For $L \ge L_\circ$ and $m \le \theta L \le L/2$,

$$\begin{aligned}\mathcal{N}_m(\epsilon)\eta^L &\le 2m(1+2/\epsilon)^{m-1}\eta^L\\ &\le L\exp(L(\theta\ln(1+2/\epsilon) - \ln(1/\eta)))\\ &= \exp(\ln(L) + L(\theta\ln(1+2/\epsilon) - \ln(1/\eta)/2) - L\ln(1/\eta)/2)\\ &\le \exp(L(\theta\ln(1+2/\epsilon) - \ln(1/\eta)/2) + \ln(\delta/2)) \qquad (4.7)\\ &\le \exp(\ln(\delta/2)) = \frac{\delta}{2}. \qquad (4.8)\end{aligned}$$

Line (4.7) follows from $L \ge L_\circ$, and line (4.8) follows from $\theta\ln(1+2/\epsilon) - \ln(1/\eta)/2 \le 0$. Set $\beta = \min\{(c-K\epsilon)^2, 1, \theta\}$. Combining Eq. (4.5), Eq. (4.6) and Eq. (4.8) we have that if $L \ge \bar{L} \triangleq \max(L_\circ, 2)$, then $\mathbb{P}[\lambda_{\min}(\mathbb{X}\mathbb{X}^T) \le m] \le \delta$ for all $m \le \beta L$. Specifically, this holds for all $L \ge 0$ and for all $m \le \beta(L - \bar{L})$. Letting $C = \beta\bar{L}$ and substituting δ for $1-\delta$ we get the statement of the theorem. □

Chapter 5

Discussion (Part I)

Cor. 3.5 and Theorem 4.12 together provide a tight characterization of the sample complexity of any sub-Gaussian product distribution with a bounded relative moment. Formally, fix $\rho > 0$. For any D such that $D_X \in \mathcal{D}^{\text{sg}}_{\rho}$, and for any $\gamma > 0$ and $\epsilon \in (0, \frac{1}{4} - \ell^*_{\gamma}(D))$

$$\Omega(k_{\gamma}(D_X)) \leq m(\epsilon, \gamma, D) \leq \tilde{O}\left(\frac{k_{\gamma}(D_X)}{\epsilon^2}\right). \tag{5.1}$$

The upper bound holds uniformly for all distributions, and the constants in the lower bound depend only on ρ. This result shows that the true sample complexity of learning each of these distributions with MEM is characterized by the margin-adapted dimension. An interesting conclusion can be drawn as to the influence of the conditional distribution of labels $D_{Y|X}$: Since Eq. (5.1) holds for *any* $D_{Y|X}$, the effect of the direction of the best separator on the sample complexity is bounded, even for highly non-spherical distributions.

We note that the upper bound that we have proved involves logarithmic factors which might not be necessary. There are upper bounds that depend on the margin alone and on the dimension alone without logarithmic factors, as shown in Theorem 1.7 and Theorem 1.20. On the other hand, in our bound, which combines the two quantities, there is a logarithmic dependence which stems from the margin component of the bound. It might be possible to tighten the bound and remove the logarithmic dependence.

We can use Eq. (5.1) to easily characterize the sample complexity behavior for interesting distributions, and to compare L_2 margin minimization to other learning methods, as we henceforth demonstrate.

Gaps between L_1 and L_2 regularization in the presence of irrelevant features Ng [2004] considers learning a single relevant feature in the presence of many irrelevant features, and compares using L_1 regularization and L_2 regularization. When $\|X\|_\infty \leq 1$, upper bounds on learning with L_1 regularization guarantee a sample complexity of $O(\ln(d))$ for an L_1-based learning rule [Zhang, 2002]. In order to compare this with the sample complexity of L_2 regularized learning and establish a gap, one must use a *lower bound* on the L_2 sample complexity. The argument provided by Ng actually assumes scale-invariance of the learning rule, and is therefore valid only for *unregularized* linear learning. In contrast, using our results we can easily establish a lower bound of $\Omega(d)$ for many specific distributions with a bounded $\|X\|_\infty$ and $Y = \text{sign}(X[i])$ for some i. For instance, if each coordinate is a bounded independent sub-Gaussian random variable with a bounded relative moment, we have $k_1 = \lceil d/2 \rceil$ and Theorem 4.12 implies a lower bound of $\Omega(d)$ on the L_2 sample complexity.

Gaps between generative and discriminative learning for a Gaussian mixture Consider two classes, each drawn from a unit-variance spherical Gaussian in $\mathbb{R}^d$ with a large distance $2v >> 1$ between the class means, such that $d >> v^4$. Then $\mathbb{P}_D[X|Y = y] = \mathcal{N}(yv \cdot e_1, I_d)$, where e_1 is a unit vector in $\mathbb{R}^d$. For any v and d, we have $D_X \in \mathcal{D}_1^{\text{sg}}$. For large values of v, we have extremely low margin error at $\gamma = v/2$, and so we can hope to learn the classes by looking for a large-margin separator. Indeed, we can calculate $k_\gamma = \lceil d/(1 + \frac{v^2}{4}) \rceil$, and conclude that the required sample complexity is $\tilde{\Theta}(d/v^2)$. Now consider a generative approach: fitting a spherical Gaussian model for each class. This amounts to estimating each class center as the empirical average of the points in the class, and classifying based on the nearest estimated class center. It is possible to show that for any constant $\epsilon > 0$, and for large enough v and d, $O(d/v^4)$ samples are enough in order to ensure an error of ϵ. This establishes a rather large gap of $\Omega(v^2)$ between the sample complexity of the discriminative approach and that of the generative one.

5.1 On the limitations of the covariance matrix

We have shown matching upper and lower bounds for the sample complexity of learning with MEM, for any sub-Gaussian product distribution with a bounded relative moment. This shows that the margin-adapted dimension fully characterizes the sample complexity of learning with MEM for such distributions. What properties of a distribution play a role for general distributions? In the following theorem we show that these properties must include more than the covariance matrix of the distribution, even when assuming sub-Gaussian tails and bounded relative moments.

Theorem 5.1. *For any integer $d > 1$, there exist two distributions D and P over $\mathbb{R}^d \times \{\pm 1\}$ with identical covariance matrices, such that for any $\epsilon \in (0, \frac{1}{4})$, $m(\epsilon, 1, P, \frac{1}{4}) \geq \Omega(d)$ while $m(\epsilon, 1, D, \delta) \leq \lceil \log_2(1/\delta) \rceil$. Both D_X and P_X are sub-Gaussian random vectors, with a relative moment of $\sqrt{2}$ in all directions.*

Proof. Let D_a and D_b be distributions over $\mathbb{R}^d$ such that D_a is uniform over $\{\pm 1\}^d$ and D_b is uniform over $\{\pm 1\} \times \{0\}^{d-1}$. Let D_X be a balanced mixture of D_a and D_b. Let P_X be uniform over $\{\pm 1\} \times \{\frac{1}{\sqrt{2}}\}^{d-1}$. For both D and P, let $\mathbb{P}[Y = \langle e_1, X\rangle] = 1$. The covariance matrix of D_X and P_X is diag$(1, \frac{1}{2}, \dots, \frac{1}{2})$, thus $k_1(D_X) = k_1(P_X) \geq \Omega(d)$.

By Eq. (4.1), P_X, D_a and D_b are all sub-Gaussian product distribution with relative moment 1, thus also with moment $\sqrt{2} > 1$. The projection of D_X along any direction $u \in \mathbb{R}^d$ is sub-Gaussian with relative moment $\sqrt{2}$ as well, since

$$\begin{aligned}
&\mathbb{E}_{X\sim D_X}[\exp(\langle u, X\rangle)] = \frac{1}{2}(\mathbb{E}_{X\sim D^a}[\exp(\langle u, X\rangle)] + \mathbb{E}_{X\sim D^b}[\exp(\langle u, X\rangle)]) \\
&= \frac{1}{2}(\prod_{i\in[d]}(\exp(u_i) + \exp(-u_i))/2 + (\exp(u_1) + \exp(-u_1))/2) \\
&\leq \frac{1}{2}(\prod_{i\in[d]}\exp(u_i^2/2) + \exp(u_1^2/2)) \leq \exp(\|u\|^2/2) \leq \exp((\|u\|^2 + u_1^2)/2) \\
&= \exp(\mathbb{E}_{X\sim D_X}[\langle u, X\rangle^2]).
\end{aligned}$$

For P we have by Theorem 4.12 that for any $\epsilon \leq \frac{1}{4}$, $m(\epsilon, 1, P, \frac{1}{4}) \geq \Omega(k_1(P_X)) \geq \Omega(d)$. In contrast, any MEM algorithm $\mathcal{A}_1$ will output the correct separator for D whenever the sample has at least one point drawn from D_b. This is because the separator e_1 is the only $w \in \mathbb{B}_1^d$ that classifies this point with zero 1-margin errors. Such a point exists in a sample of size m with probability $1 - 2^{-m}$. Therefore $\ell_{0/1}(\mathcal{A}_1, D, m, 1/2^m) = 0$. It follows that for all $\epsilon > 0$, $m(\epsilon, 1, D, \delta) \leq \lceil \log_2(1/\delta)\rceil$. □

5.2 Summary

We have shown that the true sample complexity of large-margin learning of each of a rich family of distributions is characterized by the margin-adapted dimension. Characterizing the true sample complexity allows a better comparison between this learning approach and other algorithms, and has many potential applications, such as semi-supervised learning and feature construction. The challenge of characterizing the true sample complexity extends to any distribution and any learning approach. Theorem 5.1 shows that other properties but the covariance matrix must be taken into account for general distributions. We believe that obtaining answers to these questions is of great importance, both to learning theory and to learning applications.

Part II

Multiple-Instance Learning

Chapter 6

Introduction (Part II)

In this part of the thesis we consider the learning problem termed Multiple-Instance Learning (MIL), first introduced in Dietterich et al. [1997]. MIL is a special type of a supervised classification problem. As in classical supervised classification, in MIL the learner receives a sample of labeled examples drawn i.i.d. from an arbitrary and unknown distribution, and its objective is to discover a classification rule with a small expected error over the same distribution. In MIL additional structure is assumed, whereby the examples are received as *bags* of *instances*, such that each bag is composed of several instances. It is assumed that each instance has a true label, however the learner only observes the labels of the bags. In classical MIL the label of a bag is the Boolean OR of the labels of the instances the bag contains. Various generalizations to MIL have been proposed [see e.g. Raedt, 1998, Weidmann et al., 2003]. Here we consider both classical MIL and the more general setting, where a function other than Boolean OR determines bag labels based on instance labels. This function is known to the learner a-priori. We term the more general setting *generalized MIL*.

It is possible, in principle, to view MIL as a regular supervised classification task, where a bag is a single example, and the instances in a bag are merely part of its internal representation. Such a view, however, means that one must analyze each specific MIL problem separately, and that results and methods that apply to one MIL problem are not transferable to other MIL problems. We propose instead a generic approach to the analysis of MIL, in which the properties of a MIL problem are analyzed as a function of the properties of the matching non-MIL problem. As we show here, the connections between the MIL and the non-MIL properties are strong and useful. The generic approach has the advantage that it automatically extends all knowledge and methods that apply to non-MIL problems into knowledge and methods that apply to MIL, without requiring specialized analysis for each specific MIL problem. Our results are thus applicable to diverse hypothesis classes, relationships between bag labels and instance labels, and target losses. Moreover, the generic approach allows a better theoretical understanding of the relationship, in general, between regular learning and Multi-Instance Learning with the same hypothesis class.

The generic approach can also be helpful for the design of algorithms, since it allows deriving generic methods and approaches that hold across different settings. For instance, as we show below, a generic PAC-learning algorithm can be derived for a large class of MIL problems with different hypothesis classes. Another application is a generic bag-construction mechanism which we describe in Chapter 9, and learning when bags have a manifold structure [Babenko et al., 2011]. As generic analysis goes, it might be possible to improve upon it in some specific cases. Identifying these cases and providing tighter analysis for them is an important topic for future work. We do show that in some important cases—most notably that of learning separating hyperplanes with classical MIL—our analysis is tight up to constants.

MIL has been used in numerous applications. In Dietterich et al. [1997] the drug design application motivates this setting. In this application, the goal is to predict which molecules would bind to a specific binding site. Each molecule has several possible conformations (shapes) it can take. If at least one of the conformations binds to the binding site, then the molecule is labeled positive. However, it is not possible to experimentally identify which conformation was the successful one. Thus, a molecule can be thought of as a bag of conformations, where each conformation is an instance in the bag representing the molecule. This application employs the hypothesis class of Axis Parallel Rectangles (APRs), and has made APRs the hypothesis class of choice in several theoretical works that we mention below. There are many other applications for MIL, including image classification [Maron and Ratan, 1998], web index page recommendation [Zhou et al., 2005] and text categorization [Andrews, 2007].

Previous theoretical analysis of the computational aspects of MIL has been done in two main settings. In the first setting, analyzed for instance in Auer et al. [1998], Blum and Kalai [1998], Long and Tan [1998], it is assumed that all the instances are drawn i.i.d. from a single distribution over instances, so that the instances in each bag are statistically independent. Under this independence assumption, learning from an i.i.d. sample of bags is as easy as learning from an i.i.d. sample of instances with one-sided label noise. This is stated in the following theorem.

Theorem 6.1 (Blum and Kalai, 1998)**.** *If a hypothesis class is PAC-learnable in polynomial time from one-sided random classification noise, then the same hypothesis class is PAC-learnable in polynomial time in MIL under the independence assumption. The computational complexity of learning is polynomial in the bag size and in the sample size.*

The assumption of statistical independence of the instances in each bag is, however, very limiting, as it is irrelevant to many applications.

In the second setting one assumes that bags are drawn from an arbitrary distribution *over bags*, so that the instances within a bag may be statistically dependent. This is clearly much more useful in practice, since bags usually describe a complex object with internal structure, thus it is implausible to assume even approximate independence of instances in a bag. For the hypothesis class of APRs and an arbitrary distribution over bags, it is shown in Auer et al. [1998] that if there exists a PAC-learning algorithm for MIL with APRs, and this algorithm is polynomial in both the size of the bag and the dimension of the Euclidean space, then it is possible to polynomially PAC-learn DNF formulas, a problem which is solvable only if $\mathcal{RP} = \mathcal{NP}$ [Pitt and Valiant, 1986]. In addition, if it is possible to improperly learn MIL with APRs (that is, to learn a classifier which is not itself an APR), then it is possible to improperly learn DNF formulas, a problem which has not been solved to this date for general distributions. This result implies that it is not possible to PAC-learn MIL on APRs using an algorithm which is efficient in both the bag size and the problem's dimensionality. It does not, however, preclude the possibility of performing MIL efficiently in other cases.

In practice, numerous algorithms have been proposed for MIL, each focusing on a different specialization of this problem. Almost none of these algorithms assume statistical independence of instances in a bag. Moreover, some of the algorithms explicitly exploit presumed dependences between instances in a bag. Dietterich et al. [1997] propose several heuristic algorithms for finding an APR that predicts the label of an instance and of a bag. Diverse Density [Maron and Lozano-Pérez, 1998] and EM-DD [Zhang and Goldman, 2001] employ assumptions on the structure of the bags of instances. DPBoost [Andrews and Hofmann, 2003], mi-SVM and MI-SVM [Andrews et al., 2002], and Multi-Instance Kernels [Gärtner et al., 2002] are approaches for learning MIL using margin-based objectives. Some of these methods work quite well in practice. However, no generalization guarantees have been provided for any of them.

In Chapters 7 and 8 we analyze MIL and generalized MIL in a general framework, independent of a specific application, and provide results that hold for any underlying hypothesis class. We assume a fixed hypothesis class defined over instances. We then investigate the relationship between learning with respect to this hypothesis class in the classical supervised learning setting with no bags, and learning with respect to the same hypothesis class in MIL. We address sample complexity in Chapter 7 and computational feasibility in Chapter 8.

Our sample complexity analysis shows that for binary hypotheses and thresholded real-valued hypotheses, the distribution-free sample complexity for generalized MIL grows only logarithmically with the maximal bag size. We also provide poly-logarithmic sample complexity bounds for the case of margin learning. It should be noted that many real-life applications admit large bag sizes, rendering the dependence on the bag size of practical importance. For instance, in image classification applications, the instances commonly correspond to small image patches. Thus a single bag (an image) can contain hundreds of instances or more.

We further provide distribution-dependent sample complexity bounds for more general loss functions. These bound are useful when only the average bag size is bounded. The results imply generalization bounds for previously proposed algorithms for MIL. Addressing the computational feasibility of MIL, we provide a new learning algorithm with provable guarantees for a class of bag-labeling functions that includes the Boolean OR, used in classical MIL, as a special case. Given a non-MIL learning algorithm for the desired hypothesis class, which can handle one-sided errors, we improperly learn MIL with the same hypothesis class. The construction is simple to implement, and provides a computationally efficient PAC-learning of MIL, with only a polynomial dependence of the run time on the bag size. A preliminary version of the results in these chapters has been published in Sabato and Tishby [2009].

The analysis above considers the problem of learning to classify bags using a labeled sample of bags, and do not attempt to learn to classify single instances using a labeled sample of bags. We point out that it is not generally possible to find a low-error classification rule for instances based on a bag sample. As a simple counter example, assume that the label of a bag is the Boolean OR of the labels of its instances, and that every bag includes both a positive instance and a negative instance. In this case all bags are labeled as positive, and it is not possible to distinguish the two types of instances by observing only bag labels.

In Chapter 9 we show a setting in which MIL can be used to improve the sample complexity of non-MIL learning, by constructing the artificial bags. We show how this paradigm can be implemented effectively. The results in this chapter were first published in Sabato et al. [2010b].

6.1 Notations and Definitions

Let $\mathcal{X}$ be the input space, also called the domain of instances. A bag is a finite ordered set of instances from $\mathcal{X}$. Denote the set of allowed sizes for bags in a specific MIL problem by $R \subseteq \mathbb{N}$. For any set A we denote $A^{(R)} \triangleq \cup_{n \in R} A^n$. Thus the domain of bags with a size in R and instances from $\mathcal{X}$ is $\mathcal{X}^{(R)}$. A bag of size n is denoted by $\bar{\mathbf{x}} = (x[1], \ldots, x[n])$ where each $x[j] \in \mathcal{X}$ is an instance in the bag. We denote the number of instances in $\bar{\mathbf{x}}$ by $|\bar{\mathbf{x}}|$. For an unlabeled set of bags $S = \{\bar{\mathbf{x}}_i\}_{i \in [m]}$, we denote the set of instances in the bags of S by $S^{\cup} \triangleq \{x_i[j] \mid i \in [m], j \in [|\bar{\mathbf{x}}_i|]\}$. Since this is a multi-set, any instance which repeats in several bags in S is represented the same amount of time in $S^{\cup}$. For any univariate function $f : A \to B$, we may also use its extension to a multivariate function from sequences of elements in A to sequences of elements in B, defined by $f(a[1], \ldots, a[k]) = (f(a[1]), \ldots, f(a[k]))$.

Let $I \subseteq \mathbb{R}$ be the range of hypotheses over instances or bags. $\mathcal{H} \subseteq I^{\mathcal{X}}$ is a hypothesis class for instances. Every MIL problem is defined by a fixed bag-labeling function $\psi : I^{(R)} \to I$ that determines the bag labels given the instance labels. Formally, every instance hypothesis $h : \mathcal{X} \to I$

defines a bag hypothesis, denoted by $\overline{h} : \mathcal{X}^{(R)} \to I$ and defined by

$$\forall \bar{\mathbf{x}} \in \mathcal{X}^{(R)}, \quad \overline{h}(\bar{\mathbf{x}}) \triangleq \psi(h(x[1]), \ldots, h(x[r])).$$

The hypothesis class for bags given $\mathcal{H}$ and ψ is denoted $\overline{\mathcal{H}} \triangleq \{\overline{h} \mid h \in \mathcal{H}\}$. Importantly, the identity of ψ is known to the learner a-priori, thus each ψ defines a different generalized MIL problem. For instance, in classical MIL, $I = \{\pm 1\}$ and ψ is the Boolean OR.

We assume the labeled bags are drawn from a fixed distribution D over $\mathcal{X}^{(R)} \times \{\pm 1\}$, where each pair drawn from D constitutes a bag and its binary label. The MIL learner receives a labeled sample of bags $\{(\bar{\mathbf{x}}_1, y_1), \ldots, (\bar{\mathbf{x}}_m, y_m)\} \subseteq \mathcal{X}^{(R)} \times \{\pm 1\}$ drawn from D^m, and returns a classifier $\hat{h} : \mathcal{X}^{(R)} \to I$. Its goal is to achieve a low loss $\ell(\hat{h}, D)$.

Classes of Real-Valued bag-functions

In classical MIL the bag function is the Boolean OR over binary labels, that is $I = \{\pm 1\}$ and $\psi = \mathrm{OR} : \{\pm 1\}^{(R)} \to \{\pm 1\}$. A natural extension of the Boolean OR to a function over reals is the $\max$ function. We further consider two classes of bag functions over reals, each representing a different generalization of the $\max$ function, which conserves a different subset of its properties.

The first class we consider is the class of bag-functions that extend monotone Boolean functions. Monotone Boolean functions map Boolean vectors to $\{\pm 1\}$, such that the map is monotone-increasing in each of the inputs. The set of monotone Boolean functions is exactly the set of functions that can be represented by some composition of AND and OR functions, thus it includes the Boolean OR. The natural extension of monotone Boolean functions to real functions over real vectors is achieved by replacing OR with $\max$ and AND with $\min$. Formally, we define extensions of monotone Boolean functions as follows.

Definition 6.2. *A function from $\mathbb{R}^n$ into $\mathbb{R}$ is* an extension of an n-ary monotone Boolean function *if it belongs to the set $\mathcal{M}_n$ defined inductively as follows, where the input to a function is $\mathbf{z} \in \mathbb{R}^n$:*

$$\begin{aligned}
&(1)\ \forall j \in [n], \quad \mathbf{z} \mapsto z[j] \in \mathcal{M}_n;\\
&(2)\ \forall k \in \mathbb{N}^+, \quad f_1, \ldots, f_k \in \mathcal{M}_n \Longrightarrow \mathbf{z} \mapsto \max_{j\in[k]}\{f_j(\mathbf{z})\} \in \mathcal{M}_n;\\
&(3)\ \forall k \in \mathbb{N}^+, \quad f_1, \ldots, f_k \in \mathcal{M}_n \Longrightarrow \mathbf{z} \mapsto \min_{j\in[k]}\{f_j(\mathbf{z})\} \in \mathcal{M}_n.
\end{aligned}$$

We say that a bag-function $\psi : \mathbb{R}^{(R)} \to \mathbb{R}$ extends monotone Boolean functions if for all $n \in R$, $\psi_{|\mathbb{R}^n} \in \mathcal{M}_n$.

The class of extensions to Boolean functions thus generalizes the $\max$ function in a natural way.

The second class of bag functions we consider generalizes the $\max$ function by noting that for bounded inputs, the $\max$ function can be seen as a variant of the infinity-norm $\|\mathbf{z}\|_\infty = \max |z[i]|$.

Another natural bag-function over reals is the average function, defined as $\psi(\mathbf{z}) = \frac{1}{n}\sum_{i\in[n]} z_i$, which can be seen as a variant of the 1-norm $\|\mathbf{z}\|_1 = \sum_{i\in[n]} |z[i]|$. More generally, we treat the case where the hypotheses map into $I = [-1, 1]$, and consider the class of bag functions inspired by a p-norm, defined as follows.

Definition 6.3. *For $p \in [1, \infty)$, the* p-norm bag function $\psi_p : [-1, +1]^{(R)} \to [-1, +1]$ *is defined by:*

$$\forall \mathbf{z} \in \mathbb{R}^n, \quad \psi_p(\mathbf{z}) \triangleq \left(\frac{1}{n}\sum_{i=1}^{n}(z[i]+1)^p\right)^{1/p} - 1.$$

For $p = \infty$, Define $\psi_\infty \equiv \lim_{p\to\infty} \psi_p$.

Since the inputs of ψ_p are in $[-1, +1]$, we have $\psi_p(\mathbf{z}) \equiv n^{-1/p} \cdot \|\mathbf{z} + \mathbf{1}\|_p - 1$ where n is the length of $\mathbf{z}$. Note that the average function is simply ψ_1, and $\psi_\infty \equiv \|\mathbf{z} + \mathbf{1}\|_\infty - 1 \equiv \max$. Other values of p fall between these two extremes: Due to the p-norm inequality, which states that for all $p \in [1, \infty)$ and $\mathbf{x} \in \mathbb{R}^n$, $\frac{1}{n}\|\mathbf{x}\|_1 \leq n^{-1/p}\|\mathbf{x}\|_p \leq \|\mathbf{x}\|_\infty$, we have that for all $\mathbf{z} \in [-1, +1]^n$

$$\text{average} \equiv \psi_1(\mathbf{z}) \leq \psi_p(\mathbf{z}) \leq \psi_\infty(\mathbf{z}) \equiv \max.$$

Many of our results hold when the scale of the output of the bag-function is related to the scale of its inputs. Formally, we consider cases where the output of the bag-function does not change by much unless its inputs change by much. This is formalized in the following definition of a Lipschitz bag function.

Definition 6.4. *A bag function $\psi : \mathbb{R}^{(R)} \to \mathbb{R}$ is c-*Lipschitz with respect to the infinity norm *for $c > 0$ if*

$$\forall n \in R, \forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^n, \quad |\psi(\mathbf{a}) - \psi(\mathbf{b})| \leq c\|\mathbf{a} - \mathbf{b}\|_\infty.$$

The average bag-function and the $\max$ bag functions are 1-Lipschitz. Moreover, all extensions of monotone Boolean functions are 1-Lipschitz with respect to the infinity norm—this is easy to verify by induction on Def. 6.2. All p-norm bag functions are also 1-Lipschitz, as the following derivation shows:

$$|\psi_p(\mathbf{a}) - \psi_p(\mathbf{b})| = n^{-1/p} \cdot | \, \|\mathbf{a} + 1\|_p - \|\mathbf{b} + 1\|_p| \leq n^{-1/p} \cdot \|\mathbf{a} - \mathbf{b}\|_p \leq \|\mathbf{a} - \mathbf{b}\|_\infty.$$

Thus, our results for Lipschitz bag-functions hold in particular for the two bag-function classes we have defined here, and in specifically for the $\max$ function.

Chapter 7

MIL with any Hypothesis Class

In this chapter we consider the complexity properties of hypothesis classes for MIL. In Section 7.1 the sample complexity of generalized MIL for binary hypotheses is analyzed. We provide a useful lemma bounding covering numbers for MIL in Section 7.2. In Section 7.3 we analyze the sample complexity of generalized MIL with real-valued functions for large-margin learning. Distribution-dependent results for binary learning and real-valued learning based on the average bag size are presented in Section 7.4.

7.1 Binary MIL

In this section we consider binary MIL. In binary MIL we let $I = \{\pm 1\}$, thus we have a binary instance hypothesis class $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$. We further let our loss be the zero-one loss, defined by $\ell_{0/1}(y, \hat{y}) = \mathbb{I}[y \neq \hat{y}]$. The distribution-free sample complexity of learning relative to a binary hypothesis class with the zero-one loss is governed by the VC-dimension of the hypothesis class [Vapnik and Chervonenkis, 1971]. Thus we bound the VC-dimension of $\overline{\mathcal{H}}$ as a function of the maximal possible bag size $r = \max R$, and of the VC-dimension of $\mathcal{H}$. We show that the VC-dimension of $\overline{\mathcal{H}}$ is at most logarithmic in r, and at most linear in the VC-dimension of $\mathcal{H}$, for any bag-labeling function $\psi : \{\pm 1\}^{(R)} \rightarrow \{\pm 1\}$. It follows that the sample complexity of MIL grows only logarithmically with the size of the bag. Thus MIL is feasible even for quite large bags. In fact, based on the results we show henceforth, Sabato et al. [2010a] have shown that MIL can sometimes be used to accelerate even single-instance learning. We further provide lower bounds that show that the dependence of the upper bound on r and on the VC-dimension of $\mathcal{H}$ is imperative, for a large class of Boolean bag-labeling functions. We also show a matching lower bound for the VC-dimension of classical MIL with separating hyperplanes.

7.1.1 VC-Dimension Upper Bound

Our first theorem establishes a VC-Dimension upper bound for generalized MIL. To prove the theorem we require the following useful lemma.

Lemma 7.1. *For any $R \subseteq \mathbb{N}$ and any bag function $\psi : \{\pm 1\}^{(R)} \rightarrow \{\pm 1\}$, and for any hypothesis class $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ and a finite set of bags $S \subseteq \mathcal{X}^{(R)}$,*

$$|\overline{\mathcal{H}}_{|S}| \leq |\mathcal{H}_{|S^{\cup}}|.$$

Proof. Let $h_1, h_2 \in \overline{\mathcal{H}}$ be bag hypotheses. There exist instance hypotheses $g_1, g_2 \in \mathcal{H}$ such that $\overline{g}_i = h_i$ for $i = 1, 2$. Assume that $h_{1|S} \neq h_{2|S}$. We show that $g_{1|S^{\cup}} \neq g_{2|S^{\cup}}$, thus proving the lemma.

From the assumption it follows that $\overline{g}_{1|S} \neq \overline{g}_{2|S}$. Thus there exists at least one bag $\mathbf{x} \in S$ such that $\overline{g}_2(\mathbf{x}) \neq \overline{g}_2(\mathbf{x})$. Denote its size by n. We have $\psi(g_1(x[1]), \ldots, g_1(x[n])) \neq \psi(g_2(x[1]), \ldots, g_2(x[n]))$. Hence there exists a $j \in [n]$ such that $g_1(x[j]) \neq g_2(x[j])$. By the definition of $S^{\cup}$, $x[j] \in S^{\cup}$. Therefore $g_{1|S^{\cup}} \neq g_{2|S^{\cup}}$. □

Theorem 7.2. *Assume that $\mathcal{H}$ is a hypothesis class with a finite VC-dimension d. Let $r \in \mathbb{N}$ and assume that $R \subseteq [r]$. Let the bag-labeling function $\psi : \{\pm 1\}^{(R)} \rightarrow \{\pm 1\}$ be some Boolean function. Denote the VC-dimension of $\overline{\mathcal{H}}$ by d_r. We have*

$$d_r \leq \max\{16, 2d \log(2er)\}.$$

Proof. For a set of hypotheses $\mathcal{J}$, denote by $\mathcal{J}_{|A}$ the restriction of each of its members to A, so that $\mathcal{J}_A \triangleq \{h_{|A} \mid h \in \mathcal{J}\}$. Since d_r is the VC-dimension of $\overline{\mathcal{H}}$, there exists a set of bags $S \subseteq \mathcal{X}^{(R)}$ of size d_r that is shattered by $\overline{\mathcal{H}}$, so that $|\overline{\mathcal{H}}_{|S}| = 2^{d_r}$. By Lemma 7.1 $|\overline{\mathcal{H}}_{|S}| \leq |\mathcal{H}_{|S^{\cup}}|$, therefore $2^{d_r} \leq |\mathcal{H}_{|S^{\cup}}|$. In addition, $R \subseteq [r]$ implies $|S^{\cup}| \leq r d_r$. By applying Sauer's lemma to $\mathcal{H}$ we get

$$2^{d_r} \leq |\mathcal{H}_{|S^{\cup}}| \leq \left(\frac{e|S^{\cup}|}{d}\right)^d \leq \left(\frac{erd_r}{d}\right)^d, \tag{7.1}$$

Where e is the base of the natural logarithm. It follows that $d_r \leq d(\log(er) - \log d) + d \log d_r$. To provide an explicit bound for d_r, we bound $d \log d_r$ by dividing to cases:

1. Either $d \log d_r \leq \frac{1}{2} d_r$, thus $d_r \leq 2d(\log(er) - \log d) \leq 2d \log(er)$,

2. or $\frac{1}{2} d_r < d \log d_r$. In this case,

   (a) either $d_r \leq 16$,

(b) or $d_r > 16$. In this case $\sqrt{d_r} < d_r / \log d_r < 2d$, thus $d \log d_r = 2d \log \sqrt{d_r} \leq 2d \log 2d$. Substituting in the implicit bound we get $d_r \leq d(\log(er) - \log d) + 2d \log 2d \leq 2d \log(2er)$.

Combining the cases we have $d_r \leq \max\{16, 2d\log(2er)\}$. □

7.1.2 VC-Dimension Lower Bounds

In this section we show lower bounds for the VC-dimension of binary MIL, indicating that the dependence on d and r in Theorem 7.2 is tight in two important settings.

We say that a bag-function $\psi : \{\pm 1\}^{(R)} \to \{\pm 1\}$ is *r-sensitive* if there exists a number $n \in R$ and a vector $\mathbf{c} \in \{\pm 1\}^n$ such that for at least r different numbers $j_1, \ldots, j_r \in [n]$, $\psi(c[1], \ldots, c[j_i], \ldots, c[n]) \neq \psi(c[1], \ldots, -c[j_i], \ldots, c[n])$. Many commonly used Boolean functions, such as OR, AND, Parity, and all their variants that stem from negating some of the inputs, are r-sensitive for every $r \in R$. Our first lower bound shows if ψ is r-sensitive, the bound in Theorem 7.2 cannot be improved without restricting the set of considered instance hypothesis classes.

Theorem 7.3. *Assume that the bag function $\psi : \{\pm 1\}^{(R)} \to \{\pm 1\}$ is r-sensitive for some $r \in \mathbb{N}$. For any natural d and any instance domain $\mathcal{X}$ with $|\mathcal{X}| \geq rd\lfloor \log(r) \rfloor$, there exists a hypothesis class $\mathcal{H}$ with a VC-dimension at most d, such that the VC dimension of $\overline{\mathcal{H}}$ is at least $d\lfloor \log(r) \rfloor$.*

Proof. Since ψ is r-sensitive, there are a vector $\mathbf{c} \in \{\pm 1\}^n$ and a set $J \subseteq n$ such that $|J| = r$ and $\forall j \in J, \psi(c[1], \ldots, c[n]) \neq \psi(c[1], \ldots, -c[j], \ldots, c[n])$. Since ψ maps all inputs to $\{\pm 1\}$, it follows that $\forall j \in J, \psi(c[1], \ldots, -c[j], \ldots, c[n]) = -\psi(c[1], \ldots, c[n])$. Denote $a = \psi(c[1], \ldots, c[n])$. Then we have

$$\forall j \in J, y \in \{\pm 1\}, \quad \psi(c[1], \ldots, c[j] \cdot y, \ldots, c[n]) = a \cdot y. \tag{7.2}$$

For simplicity of notation, we henceforth assume w.l.o.g. that $n = r$ and $J = [r]$.

Let $S \subseteq \mathcal{X}^r$ be a set of $d\lfloor \log(r) \rfloor$ bags of size r, such that all the instances in all the bags are distinct elements of $\mathcal{X}$. Divide S into d mutually exclusive subsets, each with $\lfloor \log(r) \rfloor$ bags. Denote bag p in subset t by $\bar{\mathbf{x}}_{(p,t)}$. We define the hypothesis class

$$\mathcal{H} \triangleq \{h[k_1, \ldots, k_d] \mid \forall i \in [d], k_i \in [2^{\lfloor \log(r) \rfloor}]\},$$

where $h[k_1, \ldots, k_d]$ is defined as follows (see illustration in Table 7.1): For $x \in \mathcal{X}$ which is not an instance of any bag in S, $h[k_1, \ldots, k_d] = -1$. For $x = x_{(p,t)}[j]$, let $b_{(p,n)}$ be bit p in the binary representation of the number n, and define

$$h[k_1,\ldots,k_d](x_{(p,t)}[j]) = \begin{cases} c[j]\cdot a(2b_{(p,j-1)}-1) & j = k_t, \\ c[j] & j \neq k_t. \end{cases}$$

t	p	Instance label $h(x_{(p,t)}[r])$								Bag label $\overline{h}(\bar{\mathbf{x}}_i)$
1	1	$-$	$-$	$-$	$+$	$-$	$-$	$-$	$-$	$+$
	2	$-$	$-$	$-$	$+$	$-$	$-$	$-$	$-$	$+$
	3	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$
2	1	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$+$	$+$
	2	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$+$	$+$
	3	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$+$	$+$
3	1	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$
	2	$-$	$+$	$-$	$-$	$-$	$-$	$-$	$-$	$+$
	3	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$

Table 7.1: An example of the hypotheses $h = h[4,8,3]$, with ψ = OR (so that $\mathbf{c}$ is the all -1 vector), $r = 8$, and $d = 3$. Each line represents a bag in S, each column represents an instance in the bag.

We now show that S is shattered by $\overline{\mathcal{H}}$, indicating that the VC-dimension of $\overline{\mathcal{H}}$ is at least $|S| = d\lfloor\log(r)\rfloor$. To complete the proof, we further show that the VC-dimension of $\mathcal{H}$ is no more than d.

S is shattered by $\overline{\mathcal{H}}$: Let $\{y_{(p,t)}\}_{p\in\lfloor\log(r)\rfloor, t\in[d]}$ be some labeling over $\{\pm1\}$ for the bags in S. For each $t \in [d]$ let

$$k_t \triangleq 1 + \sum_{p=1}^{\lfloor\log(r)\rfloor} \frac{y_{(p,t)}+1}{2}\cdot 2^{p-1}.$$

Then by Eq. (7.2), for all $p \in [\lfloor\log(r)\rfloor]$ and $t \in [d]$,

$$\begin{aligned}\overline{h}[k_1,\ldots,k_d](\bar{\mathbf{x}}_{(p,t)}) &= \psi(c[1],\ldots,c[k_t]\cdot a(2b_{(p,k_t-1)}-1),\ldots,c[r]) \\ &= a^2(2b_{(p,k_t-1)}-1) = 2b_{(p,k_t-1)}-1 = y_{(p,t)}.\end{aligned}$$

Thus $h[k_1,\ldots,k_d]$ labels S according to $\{y_{(p,t)}\}$.

The VC-dimension of $\mathcal{H}$ is no more than d: Let $A \subseteq \mathcal{X}$ of size $d+1$. If there is an element in A which is not an instance in S then this element is labeled -1 by all $h \in \mathcal{H}$, therefore A is not shattered. Otherwise, all elements in A are instances in bags in S. Since there are d subsets of S, there exist two elements in A which are instances of bags in the same subset t. Denote these instances by $x(p_1, t)[j_1]$ and $x(p_2, t)[j_2]$. Consider all the possible labelings of the two elements by hypotheses in $\mathcal{H}$. If A is shattered, there must be four possible labelings for these elements. However, by the definition of $h[k_1, \ldots, k_d]$ it is easy to see that if $j_1 = j_2 = j$ then there are at most two possible labelings by hypotheses in $\mathcal{H}$, and if $j_1 \neq j_2$ then there are at most three possible labelings. Thus A is not shattered by $\mathcal{H}$, hence the VC-dimension of $\mathcal{H}$ is no more than d. □

Theorem 7.6 below provides a lower bound for the VC-dimension of MIL for the important case where the bag-function is the Boolean OR and the hypothesis class is the class of separating hyperplanes in $\mathbb{R}^n$. It suffices to consider the class $\mathcal{W}$ of separators with a bounded norm, since scaling does not change the labeling. Let $r \in \mathbb{N}$. We denote the VC-dimension of $\overline{\mathcal{W}(\mathbb{R}^n)}$ for $R = \{r\}$ and $\psi = \text{OR}$ by $d_{r,n}$. We prove a lower bound for $d_{r,n}$ using two lemmas: Lemma 7.4 provides a lower bound for $d_{r,3}$, and Lemma 7.5 links $d_{r,n}$ for small n with $d_{r,n}$ for large n. The resulting general lower bound, which holds for $r = \max R$, is then stated in Theorem 7.6.

Lemma 7.4. *Let $d_{r,n}$ be the VC-dimension of $\overline{\mathcal{W}(\mathbb{R}^n)}$ as defined above. Then $d_{r,3} \geq \lfloor \log(2r) \rfloor$.*

Proof. Denote $L \triangleq \lfloor \log(2r) \rfloor$. We will construct a set S of L bags of size r that is shattered by $\mathcal{W}(\mathbb{R}^3)$. The construction is illustrated in Figure 7.1.

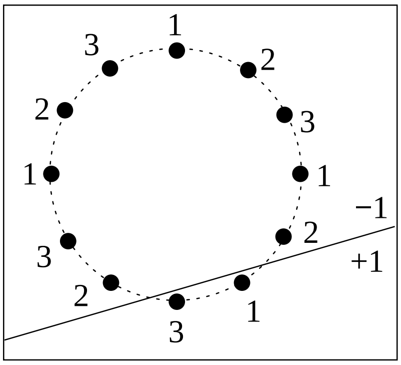

Figure 7.1: An illustration of the constructed shattered set, with $r = 4$ and $L = \log 4 + 1 = 3$. Each dot corresponds to an instance. The numbers next to the instances denote the bag to which an instance belongs, and match the sequence N defined in the proof. In this illustration bags 1 and 3 are labeled as positive by the bag-hypothesis represented by the solid line.

Let $\mathbf{n} = (n_1, \ldots, n_K)$ be a sequence of indices from $[L]$, created by concatenating all the subsets of $[L]$ in some arbitrary order, so that $K = L2^{L-1}$, and every index appears $2^{L-1} \leq r$ times in $\mathbf{n}$. Define a set $A = \{\mathbf{a}_k \mid k \in [K]\} \subseteq \mathbb{R}^3$ where $\mathbf{a}_k \triangleq (\cos(2\pi k/K), \sin(2\pi k/K), 1) \in \mathbb{R}^3$, so that $\mathbf{a}_1, \ldots, \mathbf{a}_K$ are equidistant on a unit circle on a plane embedded in $\mathbb{R}^3$. Define the set of bags $S = \{\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_L\}$ such that $\bar{\mathbf{x}}_i = (x_i[1], \ldots, x_i[r])$ where $\{x_i[j] \mid j \in [r]\} = \{a_k \mid n_k = i\}$.

We now show that S is shattered by $\mathcal{W}(\mathbb{R}^3)$: Let $(y_1, \ldots, y_L)$ be some binary labeling of L bags, and let $Y = \{i \mid y_i = +1\}$. By the definition of $\mathbf{n}$, there exist j_1, j_2 such that $Y = \{n_k \mid j_1 \leq k \leq j_2\}$. Clearly, there exists a hyperplane $\mathbf{w} \in \mathbb{R}^3$ that separates the vectors $\{\mathbf{a}_k \mid j_1 \leq k \leq j_2\}$ from the rest of the vectors in A. Thus $\text{sign}(\langle \mathbf{w} \rangle \mathbf{a}_k) = +1$ if and only if $j_1 \leq k \leq j_2$. It follows that $\overline{h}_{\mathbf{w}}(\bar{\mathbf{x}}_i) = +1$ if and only if there is a $k \in \{j_1, \ldots, j_2\}$ such that $\mathbf{a}_k$ is an instance in $\bar{\mathbf{x}}_i$, that is such that $n_k = i$. This condition holds if and only if $i \in Y$, hence $\overline{h}_{\mathbf{w}}$ classifies S according to the given labeling. It follows that S is shattered by $\mathcal{W}(\mathbb{R}^3)$, therefore $d_{r,3} \geq |S| = \lfloor \log(2r) \rfloor$. □

Lemma 7.5. *Let k, n, r be natural number such that $k \leq n$. Then $d_{r,n} \geq \lfloor n/k \rfloor d_{r,k}$.*

Proof. For a vector $\mathbf{x} \in \mathbb{R}^k$ and a number $t \in \{0, \ldots, \lfloor n/k \rfloor\}$ define the vector $s(\mathbf{x}, t) \triangleq (0, \ldots, 0, x[1], \ldots, x[k], 0, \ldots, 0) \in \mathbb{R}^n$, where $x[1]$ is at coordinate $kt + 1$. Similarly, for a bag $\bar{\mathbf{x}}_i = (\mathbf{x}_i[1], \ldots, \mathbf{x}_i[r]) \in (\mathbb{R}^k)^r$, define the bag $s(\bar{\mathbf{x}}_i, t) \triangleq (s(\mathbf{x}_i[1], t), \ldots, s(\mathbf{x}_i[r], t)) \in (\mathbb{R}^n)^r$.

Let $S_k = \{\bar{\mathbf{x}}_i\}_{i \in [d_{r,k}]} \subseteq (\mathbb{R}^k)^r$ be a set of bags with instances in $\mathbb{R}^k$ that is shattered by $\overline{\mathcal{W}(\mathbb{R}^k)}$. Define S_n, a set of bags with instances in $\mathbb{R}^n$: $S_n \triangleq \{s(\bar{\mathbf{x}}_i, t)]\}_{i \in [d_{r,k}], t \in [\lfloor n/k \rfloor]} \subseteq (\mathbb{R}^n)^r$. Then S_n is shattered by $\mathcal{W}(\mathbb{R}^n)$: Let $\{y_{(i,t)}\}_{i \in [d_{r,k}], t \in [\lfloor n/k \rfloor]}$ be some labeling for S_n. S_k is shattered by $\mathcal{W}(\mathbb{R}^k)$, hence there are separators $\mathbf{w}_1, \ldots, \mathbf{w}_{\lfloor n/k \rfloor} \in \mathbb{R}^k$ such that $\forall i \in [d_{r,k}], t \in \lfloor n/k \rfloor$, $\overline{h}_{\mathbf{w}_t}(\bar{\mathbf{x}}_i) = y_{(i,t)}$.

Set $\mathbf{w} \triangleq \sum_{t=0}^{\lfloor n/k \rfloor} s(\mathbf{w}_t, t)$. Then $\langle \mathbf{w} \rangle s(\mathbf{x}, t) = \langle \mathbf{w}_t \rangle \mathbf{x}$. Therefore

$$\begin{aligned}\overline{h}_{\mathbf{w}}(s(\bar{\mathbf{x}}_i, t)) &= \text{OR}(\text{sign}(\langle \mathbf{w} \rangle s(\mathbf{x}_i[1], t)), \ldots, \text{sign}(\langle \mathbf{w} \rangle s(\mathbf{x}_i[r], t))) \\ &= \text{OR}(\text{sign}(\langle \mathbf{w}_t \rangle \mathbf{x}_i[1]), \ldots, \text{sign}(\langle \mathbf{w}_t \rangle \mathbf{x}_i[r])) = \overline{h}_{\mathbf{w}_t}(\bar{\mathbf{x}}_i) = y_{(i,t)}.\end{aligned}$$

S_n is thus shattered, hence $d_{r,n} \geq |S_n| = \lfloor n/k \rfloor d_{r,k}$. □

The desired theorem is an immediate consequence of the two lemmas above, by noting that whenever $r \in R$, the VC-dimension of $\overline{\mathcal{W}(\mathbb{R}^n)}$ is at least $d_{r,n}$.

Theorem 7.6. *Let $\mathcal{W}(\mathbb{R}^n)$ be the class of separating hyperplanes in $\mathbb{R}^n$ as defined above. Assume that the bag function is $\psi = OR$ and the set of allowed bag sizes is R. Let $r = \max R$. Then the VC-dimension of $\overline{\mathcal{W}(\mathbb{R}^n)}$ is at least $\lfloor n/3 \rfloor \lfloor \log 2r \rfloor$.*

7.1.3 Pseudo dimension for thresholded functions

In this section we consider binary hypothesis classes that are generated from real-valued functions using thresholds. Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ be a set of real valued functions. The binary hypothesis class of thresholded functions generated by $\mathcal{F}$ is $T_{\mathcal{F}} = \{(x, z) \mapsto \text{sign}(f(x) - z) \mid f \in \mathcal{F}\}$, where $x \in \mathcal{X}$ and $z \in \mathbb{R}$. The sample complexity of learning with $T_{\mathcal{F}}$ and the zero-one loss is governed by the pseudo-dimension of $\mathcal{F}$, which is equal to the VC-dimension of $T_{\mathcal{F}}$ [Pollard, 1984]. In this section we consider a bag-labeling function $\psi : \mathbb{R}^{(R)} \to \mathbb{R}$, and bound the pseudo-dimension of $\overline{\mathcal{F}}$, thus providing an upper bound on the sample complexity of binary MIL with $T_{\overline{\mathcal{F}}}$. The following bound holds for bag-labeling functions that extend monotone Boolean functions, defined in Def. 6.2.

Theorem 7.7. *Let $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ be a function class with pseudo-dimension d . Let $R \subseteq [r]$, and assume that $\psi : \mathbb{R}^{(R)} \to \mathbb{R}$ extends monotone Boolean functions. Let d_r be the pseudo-dimension of $\overline{\mathcal{F}}$. Then*

$$d_r \leq \max\{16, 2d\log(2er)\}.$$

Proof. First, by Def. 6.2, we have that for any ψ which extends monotone Boolean functions, any $n \in R$ and any $\mathbf{y} \in \mathbb{R}^n$,

$$\begin{aligned} \operatorname{sign}(\psi(y[1], \ldots, y[n]) - z) &= \operatorname{sign}(\psi(y[1] - z, \ldots, y[n] - z)) \\ &= \psi(\operatorname{sign}(y[1] - z, \ldots, y[n] - z)). \end{aligned} \tag{7.3}$$

This can be seen by noting that each of the equalities holds for each of the operations allowed by $\mathcal{M}_n$ for each n, thus by induction they hold for all functions in $\mathcal{M}_n$ and all combinations of them.

For a real-valued function f let $t_f : \mathcal{X} \times \mathbb{R} \to \{\pm 1\}$ be defined by $t_f(y, z) = \operatorname{sign}(f(y) - z)$. We have $T_{\mathcal{F}} = \{t_f \mid f \in \mathcal{F}\}$, and $T_{\overline{\mathcal{F}}} = \{t_{\overline{f}} \mid f \in \mathcal{F}\}$. In addition, for all $f \in \mathcal{F}$, $z \in \mathbb{R}$, $n \in R$ and $\bar{\mathbf{x}} \in \mathcal{X}^n$, we have

$$\begin{aligned} t_{\overline{f}}(\bar{\mathbf{x}}, z) &= \operatorname{sign}(\overline{f}(\bar{\mathbf{x}}) - z) = \operatorname{sign}(\psi(f(x[1]), \ldots, f(x[n])) - z) \\ &= \psi(\operatorname{sign}(f(x[1]) - z, \ldots, f(x[n]) - z)) \\ &= \psi(t_f(x[1], z), \ldots, t_f(x[n], z)) = \overline{t_f}(\bar{\mathbf{x}}, z), \end{aligned} \tag{7.4}$$

where the equality on line (7.4) follows from Eq. (7.3). Therefore

$$T_{\overline{\mathcal{F}}} = \{t_{\overline{f}} \mid f \in \mathcal{F}\} = \{\overline{t_f} \mid f \in \mathcal{F}\} = \{\overline{h} \mid h \in T_{\mathcal{F}}\} = \overline{T_{\mathcal{F}}}.$$

The VC-dimension of $T_{\mathcal{F}}$ is equal to the pseudo-dimension of $\mathcal{F}$, which is d. Thus, by Theorem 7.2 and the equality above, the VC-dimension of $T_{\overline{\mathcal{F}}}$ is bounded by $\max\{16, 2d\log(2er)\}$. The proof is completed by noting that d_r, the pseudo-dimension of $\overline{\mathcal{F}}$, is exactly the VC-dimension of $T_{\overline{\mathcal{F}}}$. □

This concludes our results for distribution-free sample complexity of Binary MIL. In Section 7.4 we provide sample complexity analysis for distribution-dependent binary MIL, as a function of the average bag size.

7.2 Covering Numbers bounds for MIL

Covering numbers are a useful measure of the complexity of a function class, since they allow bounding the sample complexity of a class in various settings, based on uniform convergence guarantees [see e.g. Anthony and Bartlett, 1999]. In this section we provide a lemma that relates the covering numbers of bag hypothesis classes with those of the underlying instance hypothesis class. We will use this lemma in subsequent sections to derive sample complexity upper bounds for additional settings of MIL. Let $\mathcal{F} \subseteq \mathbb{R}^A$ be a set of real-valued functions over some domain A. A γ-cover of $\mathcal{F}$ with respect to a norm $\|\cdot\|_\circ$ defined on functions is a set of functions $\mathcal{C} \subseteq \mathbb{R}^A$ such that for any $f \in \mathcal{F}$ there exists a $g \in \mathcal{C}$ such that $\|f - g\|_\circ \leq \gamma$. The *covering number* for given $\gamma > 0$, $\mathcal{F}$ and $\circ$, denoted by $\mathcal{N}(\gamma, \mathcal{F}, \circ)$, is the size of the smallest such γ-covering for $\mathcal{F}$.

Let $S \subseteq A$ be a finite set. We consider coverings with respect to the $L_p(S)$ norm for $p \geq 1$, defined by

$$\|f\|_{L_p(S)} \triangleq \left(\frac{1}{|S|}\sum_{s\in S}|f(s)|^p\right)^{1/p}.$$

For $p = \infty$, $L_\infty(S)$ is defined by $\|f\|_{L_\infty(S)} \triangleq \max_{s\in S}|f(S)|$. The covering number of $\mathcal{F}$ for a sample size m with respect to the L_p norm is

$$\mathcal{N}_m(\gamma, \mathcal{F}, p) \triangleq \sup_{S\subseteq A:|S|=m} \mathcal{N}(\gamma, \mathcal{F}, L_p(S)).$$

A small covering number for a function class implies faster uniform convergence rates, hence smaller sample complexity for learning. The following lemma bounds the covering number of bag hypothesis-classes whenever the bag function is Lipschitz with respect to the infinity norm (see Def. 6.4). Recall that all extensions of monotone Boolean functions (Def. 6.2) and all p-norm bag-functions (Def. 6.3) are 1-Lipschitz, thus the following lemma holds for them with $a = 1$.

Lemma 7.8. *Let $R \subseteq \mathbb{N}$ and suppose the bag function $\psi : \mathbb{R}^{(R)} \to \mathbb{R}$ is a-Lipschitz with respect to the infinity norm, for some $a > 0$. Let $S \subseteq \mathcal{X}^{(R)}$ be a finite set of bags, and let r be the average size of a bag in S. For any $\gamma > 0$, $p \in [1, \infty]$, and hypothesis class $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$,*

$$\mathcal{N}(\gamma, \overline{\mathcal{H}}, L_p(S)) \leq \mathcal{N}(\frac{\gamma}{ar^{1/p}}, \mathcal{H}, L_p(S^\cup)).$$

Proof. First, note that by the Lipschitz condition on ψ, for any bag $\bar{\mathbf{x}}$ of size n and hypotheses $h, g \in \mathcal{H}$,

$$|\overline{h}(\bar{\mathbf{x}}) - \overline{g}(\bar{\mathbf{x}})| = |\psi(h(x[1]), \ldots, h(x[n])) - \psi(g(x[1]), \ldots, g(x[n]))| \leq a \max_{x\in\bar{\mathbf{x}}} |h(x) - g(x)|. \tag{7.5}$$

Let $\mathcal{C}$ be a minimal γ-cover of $\mathcal{H}$ with respect to the norm defined by $L_p(S^\cup)$, so that $|\mathcal{C}| = \mathcal{N}(\gamma, \mathcal{H}, L_p(S^\cup))$. For every $h \in \mathcal{H}$ there exists a $g \in \mathcal{C}$ such that $\|h - g\|_{L_p(S^\cup)} \le \gamma$. Assume $p < \infty$. Then by Eq. (7.5)

$$\begin{aligned}
\|\overline{h} - \overline{g}\|_{L_p(S)} &= \left(\frac{1}{|S|}\sum_{\bar{\mathbf{x}} \in S} |\overline{h}(\bar{\mathbf{x}}) - \overline{g}(\bar{\mathbf{x}})|^p\right)^{1/p} \le \left(\frac{a^p}{|S|}\sum_{\bar{\mathbf{x}} \in S} \max_{x \in \bar{\mathbf{x}}} |h(x) - g(x)|^p\right)^{1/p} \\
&\le \left(\frac{a^p}{|S|}\sum_{\bar{\mathbf{x}} \in S}\sum_{x \in \bar{\mathbf{x}}} |h(x) - g(x)|^p\right)^{1/p} = \frac{a}{|S|^{1/p}}\left(\sum_{x \in S^\cup} |h(x) - g(x)|^p\right)^{1/p} \\
&= a\left(\frac{|S^\cup|}{|S|}\right)^{1/p}\left(\frac{1}{|S^\cup|}\sum_{x \in S^\cup} |h(x) - g(x)|^p\right)^{1/p} \\
&= ar^{1/p}\|h - g\|_{L_p(S^\cup)} \le ar^{1/p} \cdot \gamma.
\end{aligned}$$

It follows that $\overline{\mathcal{C}}$ is a $(ar^{1/p}\gamma)$-covering for $\overline{\mathcal{H}}$. For $p = \infty$ we have

$$\begin{aligned}
\|\overline{h} - \overline{g}\|_{L_\infty(S)} &= \max_{\bar{\mathbf{x}} \in S} |\overline{h}(\bar{\mathbf{x}}) - \overline{g}(\bar{\mathbf{x}})| \le a \max_{\bar{\mathbf{x}} \in S} \max_{x \in \bar{\mathbf{x}}} |h(x) - g(x)| \\
&= a \max_{x \in S^\cup} |h(x) - g(x)| = a\|h - g\|_{L_\infty(S^\cup)} \le a\gamma = a \cdot r^{1/p} \cdot \gamma.
\end{aligned}$$

Thus in both cases, $\overline{\mathcal{C}}$ is a $ar^{1/p}\gamma$-covering for $\overline{\mathcal{H}}$, and its size is $\mathcal{N}(\gamma, \mathcal{H}, L_p(S^\cup))$. Thus

$$\mathcal{N}(ar^{1/p}\gamma, \overline{\mathcal{H}}, L_p(S^\cup)) \le \mathcal{N}(\gamma, \overline{\mathcal{H}}, L_p(S^\cup)).$$

We get the statement of the lemma by substituting γ with $\frac{\gamma}{ar^{1/p}}$. □

As an immediate corollary, we have the following bound for covering numbers of a given sample size.

Corollary 7.9. *Let $r \in \mathbb{N}$, and let $R \subseteq [r]$. Suppose the bag function $\psi : \mathbb{R}^{(R)} \to \mathbb{R}$ is a-Lipschitz with respect to the infinity norm for some $a > 0$. Let $\gamma > 0, p \in [1, \infty]$, and $\mathcal{H} \in \mathbb{R}^{\mathcal{X}}$. For any $m \ge 0$,*

$$\mathcal{N}_m(\gamma, \overline{\mathcal{H}}, p) \le \mathcal{N}_{rm}(\frac{\gamma}{a \cdot r^{1/p}}, \mathcal{H}, p).$$

7.3 Margin Learning for MIL

Large-margin classification is a popular supervised learning approach, which has received attention also as a method for MIL. For instance, MI-SVM [Andrews et al., 2002] attempts to optimize an adaptation of the soft-margin SVM objective [Cortes and Vapnik, 1995] to MIL, in which the margin of a bag is the maximal margin achieved by any of its instances. It has not been shown, however, whether minimizing the objective function of MI-SVM, or other margin formulations for MIL, allows learning with a reasonable sample size. We fill in this gap in Theorem 7.10 below, which bounds the γ-fat-shattering dimension (see e.g. Anthony and Bartlett 1999) of MIL. The objective of MI-SVM amounts to replacing the hypothesis class $\mathcal{H}$ of separating hyperplanes with the class of bag-hypotheses $\overline{\mathcal{H}}$ where the bag function is $\psi = \max$. Since $\max$ is the real-valued extension of OR, this objective function is natural in our MIL formulation. The distribution-free sample complexity of large-margin learning with the zero-one loss is proportional to the fat-shattering dimension [Alon et al., 1997]. Thus, we provide an upper bound on the fat-shattering dimension of MIL as a function of the fat-shattering dimension of the underlying hypothesis class, and of the maximal allowed bag size. The bound holds for any Lipschitz bag-function. Let $\gamma > 0$ be the desired margin. For a hypothesis class H, denote its γ-fat-shattering dimension by $\mathrm{Fat}(\gamma, H)$

Theorem 7.10. *Let $r \in \mathbb{N}$ and assume $R \subseteq [r]$. Let $B, a > 0$. Let $\mathcal{H} \subseteq [0, B]^{\mathcal{X}}$ be a real-valued hypothesis class and assume that the bag function $\psi : [0, B]^{(R)} \to [0, aB]$ is a-lipschitz with respect to the infinity norm. Then for all $\gamma \in (0, aB]$*

$$\mathrm{Fat}(\gamma, \overline{\mathcal{H}}) \leq \max\left\{33,\ 24\mathrm{Fat}(\frac{\gamma}{64a}, \mathcal{H}) \log^2\left(\frac{6 \cdot 2048 \cdot B^2 a^2}{\gamma^2} \cdot \mathrm{Fat}(\frac{\gamma}{64a}, \mathcal{H}) \cdot r\right)\right\}. \tag{7.6}$$

This theorem shows that for margin learning as well, the dependence of the bag size on the sample complexity is poly-logarithmic. In the proof of the theorem we use Theorem 1.11 and Theorem 1.12, which link the covering number and the fat-shattering number. Theorem 1.12 deals with the case $m \geq \mathrm{Fat}(\frac{\gamma}{4}, \mathcal{H})$. Here we consider all $m \geq 1$, thus we slightly weaken the statement of the theorem and use the following inequality, in which the fraction in the exponent is not divided by $\mathrm{Fat}(\frac{\gamma}{4}, \mathcal{H})$:

$$\mathcal{N}_m(\gamma, F, \infty) < 2\left(\frac{4B^2 m}{\gamma^2}\right)^{\mathrm{Fat}(\frac{\gamma}{4}, F)\log(4eBm/\gamma)}. \tag{7.7}$$

It is easy to check that this inequality follows from Theorem 1.12 for all $m \geq \mathrm{Fat}(\frac{\gamma}{4}, \mathcal{H}) \geq 1$, and from the trivial upper bound $\mathcal{N}_m(\gamma, \mathcal{H}, \infty) \leq (B/\gamma)^m \leq (B/\gamma)^{\mathrm{Fat}(\frac{\gamma}{4}, \mathcal{H})}$ for all $m \leq \mathrm{Fat}(\frac{\gamma}{4}, \mathcal{H})$.

Proof of Theorem 7.10. From Theorem 1.11 and Lemma 7.8 it follows that for $m \geq \mathrm{Fat}(16\gamma, \overline{\mathcal{H}})$,

$$\mathrm{Fat}(16\gamma, \overline{\mathcal{H}}) \leq \frac{8}{\log e} \log \mathcal{N}_m(\gamma, \overline{\mathcal{H}}, \infty) \leq 6 \log \mathcal{N}_{rm}(\gamma/a, \mathcal{H}, \infty). \tag{7.8}$$

By Eq. (7.7), for all $m \geq 1$, if $\mathrm{Fat}(\gamma/4) \geq 1$ then

$$\begin{aligned}
\forall \gamma \leq \frac{B}{2e}, \quad \log \mathcal{N}_m(\gamma, \mathcal{H}, \infty) &\leq 1 + \mathrm{Fat}(\frac{\gamma}{4}, \mathcal{H}) \log(\frac{4eBm}{\gamma}) \log\left(\frac{4B^2 m}{\gamma^2}\right) \\
&\leq \mathrm{Fat}(\frac{\gamma}{4}, \mathcal{H}) \log(\frac{8eBm}{\gamma}) \log\left(\frac{4B^2 m}{\gamma^2}\right) \qquad (7.9) \\
&\leq \mathrm{Fat}(\frac{\gamma}{4}, \mathcal{H}) \log^2(\frac{4B^2 m}{\gamma^2}). \qquad (7.10)
\end{aligned}$$

The inequality in line (7.9) holds since we have added 1 to the second factor, and the value of the other factors is at least 1. The last inequality follows since if $\gamma \leq \frac{B}{2e}$, we have $8eB/\gamma \leq 4B^2/\gamma^2$. Eq. (7.10) also holds if $\mathrm{Fat}(\gamma/4) < 1$, since this implies $\mathrm{Fat}(\gamma/4) = 0$ and $\mathcal{N}_m(\gamma, \mathcal{H}, \infty) = 1$. Combining Eq. (7.8) and Eq. (7.10), we get that if $m \geq \mathrm{Fat}(16\gamma, \overline{\mathcal{H}})$ then

$$\forall \gamma \leq \frac{aB}{2e}, \quad \mathrm{Fat}(16\gamma, \overline{\mathcal{H}}) \leq 6\mathrm{Fat}(\frac{\gamma}{4a}, \mathcal{H}) \log^2(\frac{4B^2 a^2 rm}{\gamma^2}). \qquad (7.11)$$

Set $m = \lceil \mathrm{Fat}(16\gamma, \overline{\mathcal{H}}) \rceil \leq \mathrm{Fat}(16\gamma, \overline{\mathcal{H}}) + 1$. If $\mathrm{Fat}(16\gamma, \overline{\mathcal{H}}) \geq 1$, we have that $m \geq \mathrm{Fat}(16\gamma, \overline{\mathcal{H}})$ and also $m \leq 2\mathrm{Fat}(16\gamma, \overline{\mathcal{H}})$. Thus Eq. (7.11) holds, and

$$\begin{aligned}
\forall \gamma \leq \frac{aB}{2e}, \quad \mathrm{Fat}(16\gamma, \overline{\mathcal{H}}) &\leq 6\mathrm{Fat}(\frac{\gamma}{4a}, \mathcal{H}) \log^2(\frac{4B^2 a^2}{\gamma^2} \cdot r \cdot (\mathrm{Fat}(16\gamma, \overline{\mathcal{H}}) + 1)) \\
&\leq 6\mathrm{Fat}(\frac{\gamma}{4a}, \mathcal{H}) \log^2(\frac{8B^2 a^2}{\gamma^2} \cdot r \cdot \mathrm{Fat}(16\gamma, \overline{\mathcal{H}})).
\end{aligned}$$

Now, it is easy to see that if $\mathrm{Fat}(16\gamma, \overline{\mathcal{H}}) < 1$, this inequality also holds. Therefore it holds in general. Substituting γ with $\gamma/16$, we have that

$$\forall \gamma \leq \frac{8aB}{e}, \quad \mathrm{Fat}(\gamma, \overline{\mathcal{H}}) \leq 6\mathrm{Fat}(\frac{\gamma}{64a}, \mathcal{H}) \log^2(\frac{2048B^2 a^2}{\gamma^2} \cdot r \cdot \mathrm{Fat}(\gamma, \overline{\mathcal{H}})). \qquad (7.12)$$

Note that the condition on γ holds, in particular, for all $\gamma \leq aB$.

To derive the desired Eq. (7.6) from Eq. (7.12), let $\beta = 6\mathrm{Fat}(\gamma/64a, \mathcal{H})$ and $\eta = 2048B^2a^2/\gamma^2$. Denote $F = \mathrm{Fat}(\gamma, \overline{\mathcal{H}})$. Then Eq. (7.12) can be restated as $F \leq \beta \log^2(\eta r F)$. It follows that $\sqrt{F}/\log(\eta r F) \leq \sqrt{\beta}$, Thus

$$\frac{\sqrt{F}}{\log(\eta r F)} \log\left(\frac{\sqrt{\eta r F}}{\log(\eta r F)}\right) \leq \sqrt{\beta} \log(\sqrt{\beta \eta r}).$$

Therefore

$$\frac{\sqrt{F}}{\log(\eta r F)}(\log(\eta r F)/2 - \log(\log(\eta r F))) \leq \sqrt{\beta}\log(\beta\eta r)/2,$$

hence

$$(1 - \frac{2\log(\log(\eta r F))}{\log(\eta r F)})\sqrt{F} \leq \sqrt{\beta}\log(\beta\eta r).$$

Now, it is easy to verify that $\log(\log(x))/\log(x) \leq \frac{1}{4}$ for all $x \geq 33 \cdot 2048$. Assume $F \geq 33$ and $\gamma \leq aB$. Then

$$\eta r F = 2048 B^2 a^2 r F/\gamma^2 \geq 2048 F \geq 33 \cdot 2048.$$

Therefore $\log(\log(\eta r F))/\log(\eta r F) \leq \frac{1}{4}$, which implies $\frac{1}{2}\sqrt{F} \leq \sqrt{\beta}\log(\beta\eta r)$. Thus $F \leq 4\beta\log^2(\beta\eta r)$. Substituting the parameters with their values, we get the desired bound, stated in Eq. (7.6). □

7.4 Sample Complexity by Average Bag Size

The upper bounds we have shown so far provide distribution-free sample complexity bounds, which depend only on the maximal possible bag size. In this section we show that even if the bag size is unbounded, we can still have a sample complexity guarantee, if the *average* bag size for the input distribution is bounded.

7.4.1 Binary MIL

Our first result complements the distribution-free sample complexity bounds that were provided for binary MIL in Section 7.1. The average (or expected) bag size under a distribution D over $\mathcal{X}^{(R)} \times \{\pm 1\}$ is $\mathbb{E}_{(\bar{\mathbf{X}},Y)\sim D}[|\bar{\mathbf{X}}|]$. Our sample complexity bound for binary MIL depends on the average bag size and the VC dimension of the instance hypothesis class. Recall that the zero-one loss is defined by $\ell_{0/1}(y, \hat{y}) = \mathbb{I}[y \neq \hat{y}]$. For a sample of labeled examples $S = \{(x_i, y_i)\}_{i\in[m]}$, we use S_X to denote the examples of S, that is $S_X = \{x_i\}_{i\in[m]}$.

Theorem 7.11. *Let $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ be a binary hypothesis class with VC-dimension d. Let $R \subseteq \mathbb{N}$ and assume a bag function $\psi : \{\pm 1\}^{(R)} \to \{\pm 1\}$. Let r be the average bag size under distribution D over labeled bags. Then*

$$\mathcal{R}(\overline{\mathcal{H}}_{\ell_{0/1}}, D) \leq 17\sqrt{\frac{d\ln(4er)}{m}}.$$

Proof. Let S be a labeled bag-sample of size m. Dudley's entropy integral [Dudley, 1967], presented in Section 1.5.1 states that

$$\mathcal{R}(\overline{\mathcal{H}}_{\ell_{0/1}}, S) \leq \frac{12}{\sqrt{m}} \int_0^1 \sqrt{\ln \mathcal{N}(\gamma, \overline{\mathcal{H}}_{\ell_{0/1}}, L_2(S))}\, d\gamma.$$

If $\mathcal{C}$ is a γ-cover for $\overline{\mathcal{H}}$ with respect to the norm $L_2(S_X)$, then $\mathcal{C}_{\ell_{0/1}}$ is a $\gamma/2$-cover for $\overline{\mathcal{H}}_{\ell_{0/1}}$ with respect to the norm $L_2(S)$. This can be seen as follows: Let $h_{\ell_{0/1}} \in \overline{\mathcal{H}}_{\ell_{0/1}}$ for some $h \in \overline{\mathcal{H}}$. Let $f \in \mathcal{C}$ such that $\|f - h\|_{L_2(S_X)} \leq \gamma$. We have

$$\begin{aligned}
\|f_{\ell_{0/1}} - h_{\ell_{0/1}}\|_{L_2(S)} &= \left(\frac{1}{m}\sum_{(x,y)\in S} |f_{\ell_{0/1}}(x,y) - h_{\ell_{0/1}}(x,y)|^2\right)^{1/2} \\
&= \left(\frac{1}{m}\sum_{(x,y)\in S} |\ell_{0/1}(y, f(x)) - \ell_{0/1}(y, h(x))|^2\right)^{1/2} \\
&= \left(\frac{1}{m}\sum_{x\in S_X} (\frac{1}{2}|f(x) - h(x)|)^2\right)^{1/2} = \frac{1}{2}\|f - h\|_{L_2(S_X)} \leq \gamma/2.
\end{aligned}$$

Therefore $\mathcal{C}_{\ell_{0/1}}$ is a $\gamma/2$-cover for $L_2(S)$. It follows that we can bound the γ-covering number of $\overline{\mathcal{H}}_{\ell_{0/1}}$ by:

$$\mathcal{N}(\gamma, \overline{\mathcal{H}}_{\ell_{0/1}}, L_2(S)) \leq \mathcal{N}(2\gamma, \overline{\mathcal{H}}, L_2(S_X)). \tag{7.13}$$

Let $r(S)$ be the average bag size in the sample S, that is $r(S) = |S^{\cup}|/|S|$. By Lemma 7.8,

$$\mathcal{N}(\gamma, \overline{\mathcal{H}}, L_2(S_X)) \leq \mathcal{N}(\gamma/\sqrt{r(S)}, \mathcal{H}, L_2(S_X^{\cup})). \tag{7.14}$$

From Eq. (1.8), Eq. (7.13) and Eq. (7.14) we conclude that

$$\mathcal{R}(\overline{\mathcal{H}}_{\ell_{0/1}}, S) \leq \frac{12}{\sqrt{m}} \int_0^1 \sqrt{\ln \mathcal{N}(2\gamma/\sqrt{r(S)}, \mathcal{H}, L_2(S_X^{\cup}))}\, d\gamma.$$

As presented in Eq. (1.1), It was shown in Dudley [1978] that for any $\mathcal{H}$ with VC-dimension d, and any $\gamma > 0$,

$$\ln \mathcal{N}(\gamma, \mathcal{H}, L_2(S_X^{\cup})) \leq 2d \ln\left(\frac{4e}{\gamma^2}\right).$$

Therefore

$$\begin{aligned}\mathcal{R}(\overline{\mathcal{H}}_{\ell_{0/1}}, S) &\leq \frac{12}{\sqrt{m}} \int_0^1 \sqrt{2d \ln\left(\frac{er(S)}{\gamma^2}\right)}\, d\gamma \\ &\leq 17\sqrt{\frac{d}{m}} \left(\int_0^1 \sqrt{\ln(er(S))}\, d\gamma + \int_0^1 \sqrt{\ln(1/\gamma^2)}\, d\gamma\right) \\ &= 17\sqrt{\frac{d(\ln(er(S)) + \sqrt{\pi/2})}{m}} \leq 17\sqrt{\frac{d\ln(4er(S))}{m}}.\end{aligned}$$

The function $\sqrt{\ln(x)}$ is concave for $x \geq 1$. Therefore we may take the expectation of both sides of this inequality and apply Jensen's inequality, to get

$$\begin{aligned}\mathcal{R}_m(\overline{\mathcal{H}}_{\ell_{0/1}}, D) = \mathbb{E}_{S\sim D^m}[\mathcal{R}(\overline{\mathcal{H}}_{\ell_{0/1}}, S)] &\leq \mathbb{E}_{S\sim D^m}\left[17\sqrt{\frac{d\ln(4er(S))}{m}}\right] \\ &\leq 17\sqrt{\frac{d\ln(4e\cdot \mathbb{E}_{S\sim D^m}[r(S)])}{m}} = 17\sqrt{\frac{d\ln(4er)}{m}}.\end{aligned}$$

□

We conclude that even when the bag size is not bounded, the sample complexity of binary MIL with a specific distribution depends only logarithmically on the average bag size in this distribution, and linearly on the VC-dimension of the underlying instance hypothesis class.

7.4.2 Real-Valued Hypothesis Classes

In our second result we wish to bound the sample complexity of MIL when using other loss functions that accept real valued predictions. This bound will depend on the average bag size, and on the Rademacher complexity of the instance hypothesis class.

We consider the case where both the bag function and the loss function are Lipschitz. For the bag function, recall that all extensions of monotone Boolean functions are Lipschitz with respect to the infinity norm. For the loss function $\ell : \{\pm 1\} \times \mathbb{R} \to \mathbb{R}$, we require that it is Lipschitz in its second argument, i.e. that there is a constant $a > 0$ such that for all $y \in \{\pm 1\}$ and $y_1, y_2 \in \mathbb{R}$, $|\ell(y, y_1) - \ell(y, y_2)| \leq a|y_1 - y_2|$. This property is satisfied by many popular losses. For instance, consider the hinge-loss $\ell_{\mathrm{hl}(\gamma)}$, defined in Section 1.3. This loss is $1/\gamma$-Lipschitz in its second argument.

The following lemma provides a bound on the empirical Rademacher complexity of MIL, as a function of the average bag size in the sample and of the behavior of the worst-case Rademacher complexity over instances. We will subsequently use this bound to bound the average Rademacher complexity of MIL with respect to a distribution. We consider losses with the range $[0,1]$. To avoid degenerate cases, we consider only losses such that there exists at least one labeled bag $(\bar{\mathbf{x}}, y) \subseteq \mathcal{X}^{(R)} \times \{\pm 1\}$ and hypotheses $h, g \in \mathcal{H}$ such that $h_\ell(\bar{\mathbf{x}}, y) = 0$ and $g_\ell(\bar{\mathbf{x}}, y) = 1$. We say that such a loss has a *full range*.

Lemma 7.12. *Let $\mathcal{H} \subseteq [0,B]^{\mathcal{X}}$ be a hypothesis class. Let $R \subseteq \mathbb{N}$, and let the bag function $\psi : \mathbb{R}^{(R)} \to \mathbb{R}$ be a_1-Lipschitz with respect to the infinity norm. Assume a loss function $\ell : \{\pm 1\} \times \mathbb{R} \to [0,1]$, which is a_2-Lipschitz in its second argument. Further assume that ℓ has a full range. Suppose there is a continuous decreasing function $f : (0,1] \to \mathbb{R}$ such that*

$$\forall \gamma \in (0,1], \quad f(\gamma) \in \mathbb{N} \implies \mathcal{R}^{\sup}_{f(\gamma)}(\mathcal{H}) \leq \gamma.$$

Let S be a labeled bag-sample of size m, with an average bag size r. Then for all $\epsilon \in (0,1]$,

$$\mathcal{R}(\overline{\mathcal{H}}_\ell, S) \leq 4\epsilon + \frac{10}{\sqrt{m}} \log\left(\frac{4ea_1^2 a_2^2 B^2 rm}{\epsilon^2}\right)\left(1 + \int_\epsilon^1 \sqrt{f(\frac{\gamma}{4a_1 a_2})}\, d\gamma\right).$$

Proof. The refinement of Dudley's entropy integral [Srebro et al., 2010, Lemma A.3], presented in Section 1.5.1, states that for all $\epsilon \in (0,1]$, for all real function classes $\mathcal{F}$ with range $[0,1]$ and for all sets S,

$$\mathcal{R}(\mathcal{F}, S) \leq 4\epsilon + \frac{10}{\sqrt{m}} \int_\epsilon^1 \sqrt{\ln \mathcal{N}(\gamma, \mathcal{F}, L_2(S))}\, d\gamma.$$

Since the range of ℓ is $[0,1]$, this holds for $\mathcal{F} = \overline{\mathcal{H}}_\ell$. In addition, for any set S, the $L_2(S)$ norm is bounded from above by the $L_\infty(S)$ norm. Therefore $\mathcal{N}(\gamma, \mathcal{F}, L_2(S)) \leq \mathcal{N}(\gamma, \mathcal{F}, L_\infty(S))$. Thus, by Eq. (1.9) we have

$$\mathcal{R}(\overline{\mathcal{H}}_\ell, S) \leq 4\epsilon + \frac{10}{\sqrt{m}} \int_\epsilon^1 \sqrt{\ln \mathcal{N}(\gamma, \overline{\mathcal{H}}_\ell, L_\infty(S))}\, d\gamma. \tag{7.15}$$

Now, let $h, g \in \mathcal{H}$ and consider $\overline{h}_\ell, \overline{g}_\ell \in \overline{\mathcal{H}}_\ell$. Since ℓ is a_2-Lipschitz, we have

$$\begin{aligned}
\|\overline{h}_\ell - \overline{g}_\ell\|_{L_\infty(S)} &= \max_{i \in [m]} |\overline{h}_\ell(\bar{\mathbf{x}}_i, y_i) - \overline{g}_\ell(\bar{\mathbf{x}}_i, y_i)| = \max_{i \in [m]} |\ell(y_i, \overline{h}(\bar{\mathbf{x}}_i)) - \ell(y_i, \overline{g}(\bar{\mathbf{x}}_i))| \\
&\leq a_2 \max_{i \in [m]} |\overline{h}(\bar{\mathbf{x}}_i) - \overline{g}(\bar{\mathbf{x}}_i)| = a_2 \|\overline{h} - \overline{g}\|_{L_\infty(S_X)}.
\end{aligned}$$

It follows that if $\mathcal{C} \subseteq \overline{\mathcal{H}}$ is a γ/a_2-cover for $\overline{\mathcal{H}}$ then $\mathcal{C}_\ell \subseteq \overline{\mathcal{H}}_\ell$ is a γ-cover for $\overline{\mathcal{H}}_\ell$. Therefore $\mathcal{N}(\gamma, \overline{\mathcal{H}}_\ell, L_\infty(S)) \leq \mathcal{N}(\gamma/a_2, \overline{\mathcal{H}}, L_\infty(S_X))$. By Lemma 7.8,

$$\mathcal{N}(\gamma/a_2, \overline{\mathcal{H}}, L_\infty(S_X)) \leq \mathcal{N}(\gamma/a_1 a_2, \mathcal{H}, L_\infty(S_X^\cup)) \leq \mathcal{N}_{rm}(\gamma/a_1 a_2, \mathcal{H}, \infty).$$

Combining this with Eq. (7.15) it follows that

$$\mathcal{R}(\overline{\mathcal{H}}_\ell, S) \le 4\epsilon + \frac{10}{\sqrt{m}} \int_\epsilon^1 \sqrt{\mathcal{N}_{rm}(\gamma/a_1 a_2, \mathcal{H}, \infty)}\, d\gamma. \tag{7.16}$$

Now, let $\gamma \in (0,1]$, and let $\gamma_\circ = \sup\{\gamma_\circ \le \gamma \mid f(\gamma_\circ) \in \mathbb{N}\}$. Since $\mathcal{R}^{\sup}_{f(\gamma_\circ)}(\mathcal{H}) \le \gamma_\circ$, by Theorem 1.16 the $\gamma_\circ$-fat-shattering dimension of $\mathcal{H}$ is at most $f(\gamma_\circ)$. It follows that

$$\mathrm{Fat}(\gamma, \mathcal{H}) \le \mathrm{Fat}(\gamma_\circ, \mathcal{H}) \le f(\gamma_\circ) \le 1 + f(\gamma).$$

The last inequality follows from the definition of $\gamma_\circ$, since f is continuous and decreasing. Therefore, by Theorem 1.12,

$$\begin{aligned}
\forall \gamma \le B, \quad \log \mathcal{N}_m(\gamma, \mathcal{H}, \infty) &\le 1 + (f(\frac{\gamma}{4}) + 1)\log(\frac{4eBm}{\gamma}) \log\left(\frac{4B^2 m}{\gamma^2}\right) \\
&\le (f(\frac{\gamma}{4}) + 1)\log(\frac{4eBm}{\gamma}) \log\left(\frac{4eB^2 m}{\gamma^2}\right) && (7.17)\\
&\le (f(\frac{\gamma}{4}) + 1)\log^2(\frac{4eB^2 m}{\gamma^2}). && (7.18)
\end{aligned}$$

The inequality in line (7.17) holds since we have added $\log(e) \ge 1$ to the third factor, and the value of the other factors is at least 1. The last inequality follows since $\gamma \le B$.

We now show that the assumption $\gamma \le B$ does not restrict us: By the assumptions on ℓ, there are $h, g \in \mathcal{H}$ and a labeled bag $(\bar{\mathbf{x}}, y)$ such that $\overline{h}_\ell(\bar{\mathbf{x}}, y) = 1$ and $\overline{g}_\ell(\bar{\mathbf{x}}, y) = 0$. Let $n = |\bar{\mathbf{x}}|$. By the Lipschitz assumptions we have

$$\begin{aligned}
1 &= |\overline{h}_\ell(\bar{\mathbf{x}}, y) - \overline{g}_\ell(\bar{\mathbf{x}}, y)| = |\ell(y, \overline{h}(\bar{\mathbf{x}})) - \ell(y, \overline{g}(\bar{\mathbf{x}}))| \le a_2 |\overline{h}(\bar{\mathbf{x}}) - \overline{g}(\bar{\mathbf{x}})| \\
&= a_2 |\psi(h(x[1]), \ldots, h(x[n])) - \psi(g(x[1]), \ldots, g(x[n]))| \\
&\le a_2 a_1 \max_{j \in [n]} |h(x[j]) - g(x[j])| \le a_1 a_2 B.
\end{aligned}$$

Thus $1 \le a_1 a_2 B$. It follows that for all $\gamma \in (0,1]$, $\gamma / a_1 a_2 \le B$. Thus Eq. (7.18) can be combined with Eq. (7.16) to get that for all $\epsilon \in (0, 1]$,

$$\begin{aligned}
\mathcal{R}(\overline{\mathcal{H}}_\ell, S) &\leq 4\epsilon + \frac{10}{\sqrt{m}} \int_\epsilon^1 \sqrt{\left(f(\frac{\gamma}{4a_1a_2}) + 1\right) \log^2\left(\frac{4ea_1^2a_2^2B^2rm}{\gamma^2}\right)}\, d\gamma \\
&\leq 4\epsilon + \frac{10}{\sqrt{m}} \log\left(\frac{4ea_1^2a_2^2B^2rm}{\epsilon^2}\right) \int_\epsilon^1 \sqrt{f(\frac{\gamma}{4a_1a_2}) + 1}\, d\gamma \\
&\leq 4\epsilon + \frac{10}{\sqrt{m}} \log\left(\frac{4ea_1^2a_2^2B^2rm}{\epsilon^2}\right) \left(1 + \int_\epsilon^1 \sqrt{f(\frac{\gamma}{4a_1a_2})}\, d\gamma\right).
\end{aligned}$$

The last inequality follows from the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for non-negative a and b, and from $\int_\epsilon^1 1 \leq 1$. □

Based on Lemma 7.12, we will now bound the average Rademacher complexity of MIL, as a function of the worst-case Rademacher complexity over instances, and the expected bag size. Since the number of instances in a bag sample of a certain size is not fixed, but depends on the bag sizes in the specific sample, we will need to consider the behavior of $\mathcal{R}^{\mathrm{sup}}_m(\mathcal{H})$ for different values of m. For many learnable function classes, the Rademacher complexity is proportional to $\frac{1}{\sqrt{m}}$, or to $\frac{ln^\beta(m)}{\sqrt{m}}$ for some non-negative β. The following theorem bounds the average Rademacher complexity of MIL in all these cases. The resulting bound indicates that here too there is a poly-logarithmic dependence of the sample complexity on the average bag size. Following the proof we show an application of the bound to a specific function class.

Theorem 7.13. *Let $\mathcal{H} \subseteq [0, B]^{\mathcal{X}}$ be a hypothesis class. Let $R \subseteq \mathbb{N}$, and let the bag function $\psi : \mathbb{R}^{(R)} \rightarrow \mathbb{R}$ be a_1-Lipschitz with respect to the infinity norm. Assume a loss function $\ell : \{\pm 1\} \times \mathbb{R} \rightarrow [0, 1]$, which is a_2-Lipschitz in its second argument. Further assume that ℓ has a full range. Suppose that there are $C, \beta, K \geq 0$ such that for all $m \geq K$,*

$$\mathcal{R}^{\mathrm{sup}}_m(\mathcal{H}) \leq \frac{C \ln^\beta(m)}{\sqrt{m}}.$$

Then there exists a number $N \geq 0$ that depends only on C, β and K such that for any distribution D with average bag size r, and for all $m \geq 1$,

$$\mathcal{R}_m(\overline{\mathcal{H}}_\ell, D) \leq \frac{4 + 10\log(4ea_1^2a_2^2B^2rm^2)\left(N + \frac{a_1a_2}{\beta+1} C \ln^{\beta+1}(16a_1^2a_2^2m)\right)}{\sqrt{m}}.$$

Proof. Let S be a labeled bag sample of size m, and let $\tilde{r}$ be its average bag size. Denote $T(x) = C\ln^\beta(x)$, and define $f(\gamma) = \frac{4T^2(1/\gamma^2)}{\gamma^2}$. We will show that $\mathcal{R}^{\mathrm{sup}}_{f(\gamma)}(\mathcal{H}) \leq \gamma$, thus allowing the use of Lemma 7.12. We have $\mathcal{R}_m \leq T(m)/\sqrt{m}$, thus it suffices to show that

$T(f(\gamma))/\sqrt{f(\gamma)} \leq \gamma$.

Let $z(\gamma) = \sqrt{f(\gamma)}/T(f(\gamma))$. We will now show that $z(\gamma)T(z^2(\gamma)) \geq \frac{1}{\gamma}T(1/\gamma^2)$. Since the function $xT(x^2) = Cx\ln^\beta(x^2)$ is monotonic increasing for $x \geq 1$, we will conclude that $z(\gamma) \geq 1/\gamma$ for all $\gamma \leq 1$.

It is easy to see that for all values of $\beta, C \geq 0$, there is a number $n \geq 0$ such that for all $x \geq n$,

$$C^2\ln^{2\beta}(x) \leq x^{1-2^{-1/\beta}}.$$

For such x we have

$$\begin{aligned} T(x/T^2(x)) &= C\ln^\beta(\frac{x}{C^2\ln^{2\beta}(x)}) = C(\ln(x) - \ln(C^2\ln^{2\beta}(x)))^\beta \\ &\geq C(\ln(x) - (1-2^{-1/\beta})\ln(x)))^\beta = C\ln^\beta(x)/2 = T(x)/2. \end{aligned} \tag{7.19}$$

Let $\gamma_\circ > 0$ such that $f(\gamma_\circ) = k = \max\{n, K\}$. Since $f(\gamma)$ is monotonic decreasing with γ, for all $\gamma \leq \gamma_\circ$, $f(\gamma) \geq k$. Therefore, for $\gamma \leq \gamma_\circ$,

$$z(\gamma)T(z^2(\gamma)) = \frac{\sqrt{f(\gamma)}}{T(f(\gamma))}T(\frac{f(\gamma)}{T^2(f(\gamma))}) \geq \frac{1}{2}\frac{\sqrt{f(\gamma)}}{T(f(\gamma))}T(f(\gamma)) = \frac{1}{2}\sqrt{f(\gamma)} = T(1/\gamma^2)/\gamma.$$

The middle inequality follows from Eq. (7.19), and the last equality follows from the definition of $f(\gamma)$. We conclude that $z(\gamma) \geq \frac{1}{\gamma}$. Therefore, for all $\gamma \leq \gamma_\circ$,

$$\mathcal{R}^{\sup}_{f(\gamma)}(\mathcal{H}) \leq \frac{T(f(\gamma))}{\sqrt{f(\gamma)}} = 1/z(\gamma) \leq \gamma.$$

Define $\tilde{f}$ as follows:

$$\tilde{f}(\gamma) = \begin{cases} f(\gamma) & \gamma \leq \gamma_\circ \\ k & \gamma > \gamma_\circ. \end{cases}$$

For $\gamma \leq \gamma_\circ$, clearly $\mathcal{R}^{\sup}_{\tilde{f}(\gamma)}(\mathcal{H}) \leq \gamma$, and for $\gamma > \gamma_\circ$,

$$\mathcal{R}^{\sup}_{\tilde{f}(\gamma)}(\mathcal{H}) = \mathcal{R}^{\sup}_{k}(\mathcal{H}) = \mathcal{R}^{\sup}_{f(\gamma_\circ)}(\mathcal{H}) \leq \gamma_\circ \leq \gamma.$$

Therefore for all $\gamma \in (0,1]$, $\mathcal{R}^{\text{sup}}_{\tilde{f}(\gamma)}(\mathcal{H}) \leq \gamma$. By Lemma 7.12, for all $\epsilon \in (0,1]$,

$$\begin{aligned}
\mathcal{R}(\overline{\mathcal{H}}_\ell, S) &\leq 4\epsilon + \frac{10}{\sqrt{m}} \log\left(\frac{4ea_1^2a_2^2B^2\tilde{r}m}{\epsilon^2}\right)\left(1 + \int_\epsilon^1 \sqrt{\tilde{f}(\frac{\gamma}{4a_1a_2})}\, d\gamma\right) \\
&= 4\epsilon + \frac{10}{\sqrt{m}} \log\left(\frac{4ea_1^2a_2^2B^2\tilde{r}m}{\epsilon^2}\right)\left(1 + \int_{4a_1a_2\gamma_\circ}^1 \sqrt{k}\, d\gamma + \int_\epsilon^{4a_1a_2\gamma_\circ} \sqrt{f(\frac{\gamma}{4a_1a_2})}\, d\gamma\right) \\
&\leq 4\epsilon + \frac{10}{\sqrt{m}} \log\left(\frac{4ea_1^2a_2^2B^2\tilde{r}m}{\epsilon^2}\right)\left(1 + \sqrt{k} + \int_\epsilon^{4a_1a_2\gamma_\circ} \sqrt{f(\frac{\gamma}{4a_1a_2})}\, d\gamma\right). \qquad (7.20)
\end{aligned}$$

Denote $N = 1 + \sqrt{k}$. Now, if $\beta > 0$ we have

$$\begin{aligned}
&\int_\epsilon^{4a_1a_2\gamma_\circ} \sqrt{f(\frac{\gamma}{4a_1a_2})}\, d\gamma \leq \int_\epsilon^\infty \sqrt{f(\frac{\gamma}{4a_1a_2})}\, d\gamma = 2a_1a_2 \int_\epsilon^\infty \frac{T(16a_1^2a_2^2/\gamma^2)}{\gamma}\, d\gamma \\
&= 2a_1a_2C \int_\epsilon^\infty \frac{\ln^\beta(16a_1^2a_2^2/\gamma^2)}{\gamma}\, d\gamma = 2a_1a_2C \left[-\ln^{\beta+1}(\frac{16a_1^2a_2^2}{\gamma^2})/(2(\beta+1))\right]_\epsilon^\infty \\
&= \frac{a_1a_2C}{\beta+1}\left(\ln^{\beta+1}(\frac{16a_1^2a_2^2}{\epsilon^2})\right).
\end{aligned}$$

The same inequality holds also for $\beta = 0$, since in that case

$$\begin{aligned}
&\int_\epsilon^{4a_1a_2\gamma_\circ} \sqrt{f(\frac{\gamma}{4a_1a_2})}\, d\gamma = 2a_1a_2 \int_\epsilon^{4a_1a_2\gamma_\circ} \frac{T(16a_1^2a_2^2/\gamma^2)}{\gamma}\, d\gamma \\
&\quad = 2a_1a_2C \int_\epsilon^{4a_1a_2\gamma_\circ} \frac{1}{\gamma} = 2a_1a_2C\left[\ln(\gamma)\right]_\epsilon^{4a_1a_2\gamma_\circ} = 2a_1a_2C\ln(\frac{4a_1a_2\gamma_\circ}{\epsilon}) \\
&\leq 2a_1a_2C\ln(\frac{4a_1a_2}{\epsilon}) = \frac{a_1a_2C}{\beta+1}\left(\ln^{\beta+1}(\frac{16a_1^2a_2^2}{\epsilon^2})\right).
\end{aligned}$$

Therefore we can further bound Eq. (7.20) to get

$$\mathcal{R}(\overline{\mathcal{H}}_\ell, S) \leq 4\epsilon + \frac{10}{\sqrt{m}} \log\left(\frac{4ea_1^2a_2^2B^2\tilde{r}m}{\epsilon^2}\right)\left(N + \frac{a_1a_2C}{\beta+1}\ln^{\beta+1}(\frac{16a_1^2a_2^2}{\epsilon^2})\right).$$

Setting $\epsilon = 1/\sqrt{m}$ we get

$$\mathcal{R}(\overline{\mathcal{H}}_\ell, S) \leq \frac{4 + 10\log(4ea_1^2a_2^2B^2\tilde{r}m^2)\left(N + \frac{a_1a_2C}{\beta+1}\ln^{\beta+1}(16a_1^2a_2^2m)\right)}{\sqrt{m}}.$$

Now, for a given sample S denote its average bag size by $\tilde{r}(S)$. We have

$$\begin{aligned}\mathcal{R}_m(\overline{\mathcal{H}}, D) &= \mathbb{E}_{S\sim D^m}[\mathcal{R}(\overline{\mathcal{H}}_\ell, S)] \\ &\leq \mathbb{E}\left[\frac{4 + 10\log(4ea_1^2a_2^2B^2\tilde{r}(S)m^2)\left(N + \frac{a_1a_2C}{\beta+1}\ln^{\beta+1}(16a_1^2a_2^2m)\right)}{\sqrt{m}}\right] \\ &\leq \frac{4 + 10\log(4ea_1^2a_2^2B^2rm^2)\left(N + \frac{a_1a_2C}{\beta+1}\ln^{\beta+1}(16a_1^2a_2^2m)\right)}{\sqrt{m}}.\end{aligned}$$

In the last inequality we used Jensen's inequality and the fact that $\mathbb{E}_{S\sim D^m}[\tilde{r}(S)] = r$. This is the desired bound, hence the theorem is proven. □

To demonstrate the implications of this theorem, consider the case of MIL with soft-margin kernel SVM. Kernel SVM can operate in a general Hilbert space $\mathcal{S}$, which we denote by $\mathcal{T}$. The domain of instances is $\mathcal{X} = \{x \in \mathcal{T} \mid \|x\| \leq 1\}$, and the function class is the class of linear separators with a bounded norm $\mathcal{W}(\mathcal{S})$. The loss is the hinge-loss $\ell_{\mathrm{hl}(\gamma)}$ defined in Section 1.3, which is $1/\gamma$-Lipschitz in the second argument. We have [Bartlett and Mendelson, 2002]

$$\mathcal{R}_m^{\sup}(\mathcal{W}_{\ell_{\mathrm{hl}(\gamma)}}) \leq \frac{1}{\gamma\sqrt{m}} = \frac{\ln^0(m)}{\gamma\sqrt{m}}.$$

Thus we can apply Theorem 7.13 with $\beta = 0$. Note that $\mathcal{W} \subseteq [-1,1]^{\mathcal{X}}$, thus we can apply the theorem with $B = 2$ by simply shifting the output of each h_w by 1 and adjusting the loss function accordingly. By Theorem 7.13 there exists a number N such that for any 1-Lipschitz bag-function ψ (such as $\max$) and for any distribution D over labeled bags with an average bag size of r, we have

$$\mathcal{R}_m(\overline{\mathcal{H}}_\ell, D) \leq \frac{4 + 10\log(16erm^2/\gamma^2)\left(N + (1/\gamma)\cdot\ln(16m/\gamma)\right)}{\sqrt{m}}.$$

We can use this result and apply Eq. (1.3) to get an upper bound on the loss of MIL with soft-margin SVM.

Chapter 8

PAC-Learning for MIL

In the previous chapter we addressed the sample complexity of generalized MIL, showing that it grows only logarithmically with the bag size. We now turn to consider the computational aspect of MIL, and specifically the relationship between computational feasibility of MIL and computational feasibility of the learning problem for the underlying instance hypothesis.

We consider real-valued hypothesis classes $\mathcal{H} \in [-1, +1]^{\mathcal{X}}$, and provide a MIL algorithm which uses a learning algorithm that operates on single instances as an oracle. We show that if the oracle can minimize error with respect to $\mathcal{H}$, and the bag-function satisfies certain boundedness conditions, then the MIL algorithm is guaranteed to PAC-learn $\overline{\mathcal{H}}$. In particular, the guarantees hold if the bag-function is Boolean OR or $\max$, as in classical MIL and its extension to real-valued hypotheses.

Given an algorithm $\mathcal{A}$ that learns $\mathcal{H}$ from single instances, we provide an algorithm called `MILearn` that uses $\mathcal{A}$ to implement a *weak learner* for bags with respect to $\overline{\mathcal{H}}$. That is, for any weighted sample of bags, `MILearn` returns a hypothesis from $\overline{\mathcal{H}}$ that has some success in labeling the bag-sample correctly. This will allow the use of `MILearn` as the building block in a Boosting algorithm [Freund and Schapire, 1997], which will find a convex combination of hypotheses from $\overline{\mathcal{H}}$ that classifies unseen bags with high accuracy. Furthermore, if $\mathcal{A}$ is efficient then the resulting Boosting algorithm is also efficient, with a polynomial dependence on the maximal bag size.

We open with background on Boosting in Section 8.1. We then describe the weak learner in and analyze its properties in Section 8.2. In Section 8.3 we provide guarantees on a Boosting algorithm that uses our weak leaner, and conclude that the computational complexity of PAC-learning for MIL can be bounded by the computational complexity of agnostic PAC-learning for single instances.

8.1 Background: Boosting with Margin Guarantees

In this section we give some background on Boosting algorithms, which we will use to derive an efficient learning algorithm for MIL. Boosting methods [Freund and Schapire, 1997] are techniques that allow enhancing the power of a *weak learner*—a learning algorithm that achieves error slightly better than chance—to derive a classification rule that has low error on an input sample. The idea is to iteratively execute the weak learner on weighted versions of the input sample, and then to return a convex combination of the classifiers that were emitted by the weak learner in each round.

Let A be a domain of objects to classify, and let $H : [-1, +1]^A$ be the hypothesis class used by the weak learner. A Boosting algorithm receives as input a labeled sample $S = \{(x_i, y_i)\}_{i=1}^m \subseteq A \times \{\pm 1\}$, and iteratively feeds to the weak learner a reweighed version of S. Denote the m-dimensional simplex by $\Delta_m = \{\mathbf{w} \in \mathbb{R}^m \mid \sum_{i\in[m]} w_i = 1, \forall i \in [m], w[i] \geq 0\}$. For a vector $\mathbf{w} \in \Delta_m$, $S_{\mathbf{w}} = \{(w[i], x_i, y_i)\}_{i=1}^m$ is the sample S reweighed by $\mathbf{w}$. The Boosting algorithm runs in k rounds. On round t it sets a weight vector $\mathbf{w}_t \in \Delta_m$, calls the weak learner with input $S_{\mathbf{w}_t}$, and receives a hypothesis $h_t \in H$ as output from the weak learner. After k rounds, the Boosting algorithm returns a classifier $f_\circ : A \to [-1, +1]$, which is a convex combination of the hypotheses received from the weak learner: $f_\circ = \sum_{t\in[k]} \alpha_t h_t$, where $\alpha_1, \ldots, \alpha_k \geq 0$, and $\sum_{i\in[k]} \alpha_i = 1$.

The literature offers plenty of Boosting algorithms with desirable properties. For concreteness, we use the algorithm `AdaBoost`* [Rätsch and Warmuth, 2005], since it provides suitable guarantees on the *margin* of its output classifier. For a labeled example (x, y), the quantity $yf_\circ(x)$ is the margin of $f_\circ$ when classifying x. If the margin is positive, then $\mathrm{sign} \circ f_\circ$ classifies x correctly. The margin of any function f on a labeled sample $S = \{(x_i, y_i)\}_{i=1}^m$ is defined as

$$M(f, S) = \min_{i\in[m]} y_i f(x_i).$$

If $M(f, S)$ is positive, then the entire sample is classified correctly by $\mathrm{sign} \circ f$.

If S is an i.i.d. sample drawn from a distribution on $A \times \{\pm 1\}$, then classification error of $f_\circ$ on the distribution can be bounded based on $M(f_\circ, S)$ and the pseudo-dimension d of the hypothesis class H. The following bound [Schapire and Singer, 1999, Theorem 8] holds with probability $1 - \delta$ over the training samples, for any $m \geq d$:

$$\mathbb{P}[Y \cdot f_\circ(X) \leq 0] \leq O\left(\sqrt{\frac{d \ln^2(m/d)/M^2(f_\circ, S) + \ln(1/\delta)}{m}}\right). \tag{8.1}$$

In fact, inspection of the proof of this bound in Schapire and Singer [1999] reveals that the only property of the hypothesis class H that is used to achieve this result is the following bound, due to

Haussler and Long [1995], on the covering number of a hypothesis class H with pseudo-dimension d:

$$\forall \gamma \in (0,1], \quad \mathcal{N}_m(\gamma, \mathcal{H}, \infty) \leq \left(\frac{em}{\gamma d}\right)^d. \tag{8.2}$$

Thus, Eq. (8.1) holds whenever this covering bound holds—a fact that will be useful to us.

For `AdaBoost`*, a guarantee on the size of the margin of $f_\circ$ can be achieved if one can provide a guarantee on the *edge* of the hypotheses returned by the weak learner. The edge of a hypothesis measures of how successful it is in classifying labeled examples. Let $h : A \to [-1,+1]$ be a hypothesis and let D be a distribution over $A \times \{\pm 1\}$. The edge of h with respect to D is

$$\Gamma(h, D) \triangleq \mathbb{E}_{(X,Y)\sim D}[Y \cdot h(X)].$$

For a weighted and labeled sample $S = \{(w_i, x_i, y_i)\}_{i\in[m]} \subseteq \mathbb{R}_+ \times A \times \{\pm 1\}$,

$$\Gamma(h, S) \triangleq \sum_{i\in[m]} w_i y_i h(x_i).$$

Note that if $h(x)$ is interpreted as the probability of h to emit 1 for input x, then $\frac{1-\Gamma(h,D)}{2}$ is the expected misclassification error of h on D. Thus, a positive edge implies a labeling success of more than chance. For `AdaBoost`*, a positive edge on each of the weighted samples fed to the weak learner suffices to guarantee a positive margin of its output classifier $f_\circ$.

Theorem 8.1 (Rätsch and Warmuth 2005)**.** *Assume `AdaBoost`*receives a labeled sample S of size m as input. Suppose that `AdaBoost`* runs for k rounds and returns the classifier $f_\circ$. If for every round $t \in [k]$, $\Gamma(h_t, S_{\mathbf{w}_t}) \geq \rho$, then $M(f_\circ, S) \geq \rho - \sqrt{2\ln m/k}$.*

We present a simple corollary, which we will use when analyzing Boosting for MIL. This corollary shows that `AdaBoost`* can be used to transform a weak learner that approximates the best edge of a weighted sample to a Boosting algorithm that approximates the best margin of a labeled sample. The proof of the corollary employs the following well known result, originally by von Neumann [1928] and later extended [see e.g. Nash and Sofer, 1996]. For a hypothesis class H, denote by $\mathrm{co}(H)$ the set of all convex combinations of hypotheses in H. We say that $H \subseteq [-1,+1]^A$ is compact with respect to a sample $S = \{(x_i, y_i)\}_{i\in[m]} \subseteq A \times \{\pm 1\}$ if the set of vectors $\{(h(x_1), \ldots, h(x_m)) \mid h \in H\}$ is compact.

Theorem 8.2 (The Strong Min-Max theorem)**.** *If H is compact with respect to S, then*

$$\min_{\mathbf{w}\in\Delta_m} \sup_{h\in H} \Gamma(h, S_{\mathbf{w}}) = \sup_{f\in\mathrm{co}(H)} M(f, S).$$

Corollary 8.3. *Suppose that* `AdaBoost`* *is executed with an input sample S, and assume that H is compact with respect to S. Assumpe the weak learner used by* `AdaBoost`* *has the following guarantee: For any $\mathbf{w} \in \Delta_m$, if the weak learner receives $S_{\mathbf{w}}$ as input, then with probability at least $1-\delta$ it returns a hypothesis $h_\circ$ such that*

$$\Gamma(h_\circ, S_{\mathbf{w}}) \geq g(\sup_{h \in H} \Gamma(h, S_{\mathbf{w}})),$$

where $g : [-1,+1] \to [-1,+1]$ is some fixed non-decreasing function. Then for any input sample S, if `AdaBoost`* *runs k rounds, it returns a convex combination of hypotheses $f_\circ = \sum_{t \in [k]} \alpha_t h_t$, such that with probability at least $1 - k\delta$*

$$M(f_\circ, S) \geq g(\sup_{f \in \mathrm{co}(H)} M(f, S)) - \sqrt{2 \ln m / k}.$$

Proof. By Theorem 8.2, $\min_{\mathbf{w} \in \Delta_m} \sup_{h \in H} \Gamma(h, S_{\mathbf{w}}) = \sup_{f \in \mathrm{co}(H)} M(f, S)$. Thus, for any vector of weights $\mathbf{w}$ in the simplex, $\sup_{h \in H} \Gamma(h, S_{\mathbf{w}}) \geq \sup_{f \in \mathrm{co}(H)} M(f, S)$. It follows that in each round, the weak learner that receives $S_{\mathbf{w}_t}$ as input returns a hypothesis h_t such that $\Gamma(h_t, S_{\mathbf{w}_t}) \geq g(\sup_{h \in H} \Gamma(h, S_{\mathbf{w}_t})) \geq g(\sup_{f \in \mathrm{co}(H)} M(f, S))$. By Theorem 8.1, it follows that $M(f_\circ, S) \geq g(\sup_{f \in \mathrm{co}(H)} M(f, S)) - \sqrt{2 \ln m / k}$. □

8.2 The Weak Learner

In this section we will present our weak learner for MIL and provide guarantees for the edge it achieves. Our guarantees depend on boundedness properties of the bag-function ψ, which we define below. To motivate our definition of boundedness, consider the p-norm bag functions (see Def. 6.3), defined by $\psi_p(\mathbf{z}) \triangleq \left(\frac{1}{n} \sum_{i=1}^{n} (z[i] + 1)^p\right)^{1/p} - 1$. Recall that this class of functions includes the max function (ψ_∞) and the average function (ψ_1) as two extremes. Assume $R \subseteq [r]$ for some $r \in \mathbb{N}$. It is easy to verify that for any natural n, any sequence $z_1, \ldots, z_n \in [-1,+1]$, and all $p \in [1, \infty]$,

$$\frac{1}{n} \sum_{i \in [n]} z_i \leq \psi_p(z_1, \ldots, z_n) \leq \sum_{i \in [n]} z_i + n - 1.$$

Since $R \subseteq [r]$, it follows that for all $(z_1, \ldots, z_n) \in [-1,+1]^{(R)}$,

$$\frac{1}{r} \sum_{i \in [n]} z_i \leq \psi_p(z_1, \ldots, z_n) \leq \sum_{i \in [n]} z_i + r - 1. \tag{8.3}$$

We will show that in cases where the bag function is linearly bounded in the sum of its arguments, as in Eq. (8.3), a single-instance learning algorithm can be used to learn MIL. Our weak learner will be parameterized by the boundedness parameters of the bag-function, defined formally as follows.

Definition 8.4. *A function* $\psi : [-1,+1]^{(R)} \to [-1,+1]$ *is* (a,b,c,d)-bounded *if for all* $(z_1,\dots,z_n) \in [-1,+1]^{(R)}$,

$$a \sum_{i\in[n]} z_i + b \le \psi(z_1,\dots,z_n) \le c \sum_{i\in[n]} z_i + d.$$

Thus, for all $p \in [1,\infty)$, ψ_p over bags of size at most r is $(\frac{1}{r},0,1,r-1)$-bounded.

Before listing the weak learner `MILearn`, we introduce some notations. $\mathbf{h}_{\text{pos}}$ denotes a special bag-hypothesis that labels all bags as $+1$: $\forall x \in \mathcal{X}^{(R)}, \quad \mathbf{h}_{\text{pos}}(x) = 1$. We denote $\overline{\mathcal{H}}_+ \triangleq \overline{\mathcal{H}} \cup \{\mathbf{h}_{\text{pos}}\}$. Let $\mathcal{A}$ be an algorithm that receives a labeled and weighted instance sample as input, and returns a hypothesis $h \in \mathcal{H}$. The result of running $\mathcal{A}$ with input S is denoted $\mathcal{A}(S) \in \mathcal{H}$.

The algorithm `MILearn`, listed as Alg. 1 below, accepts as input a bag sample $\overline{S}$ and a bounded bag-function ψ. It also has access to the algorithm $\mathcal{A}$. We sometimes emphasize that `MILearn` uses a specific algorithm $\mathcal{A}$ as an oracle by writing `MILearn`$^{\mathcal{A}}$. `MILearn` constructs a sample of instances S_I from the instances that make up the bags in $\overline{S}$, labeling each instance in S_I with the label of the bag it came from. The weights of the instances depend on whether the bag they came from was positive or negative, and on the boundedness properties of ψ. Having constructed S_I, `MILearn` calls $\mathcal{A}$ with S_I. It then decides whether to return the bag-hypothesis induced by applying ψ to $\mathcal{A}(S_I)$, or to simply return $\mathbf{h}_{\text{pos}}$.

It is easy to see that the time complexity of `MILearn` is bounded by $O(f(N)+N)$, where N is the total number of instances in the bags of $\overline{S}$, and $f(n)$ is an upper bound on the time complexity of $\mathcal{A}$ when running on a sample of size n. As we presently show, the output of `MILearn` is a bag-hypothesis in $\overline{\mathcal{H}}_+$ whose edge on $\overline{S}$ depends on the best achievable edge for $\overline{S}$.

The guarantees for `MILearn`$^{\mathcal{A}}$ depend on the properties of $\mathcal{A}$. We define two properties that we consider for $\mathcal{A}$. The first property is that the edge of the hypothesis $\mathcal{A}$ returns is close to the best possible one on the input sample.

Definition 8.5 (ϵ-optimal)**.** *An algorithm $\mathcal{A}$ that accepts a weighted and labeled sample of instances in $\mathcal{X}$ and returns a hypothesis in $\mathcal{H}$ is* ϵ-optimal *if for all weighted samples $S \subseteq \mathbb{R}_+ \times \mathcal{X} \times \{\pm 1\}$ with total weight W,*

$$\Gamma(\mathcal{A}(S),S) \ge \sup_{h\in\mathcal{H}} \Gamma(h,S) - \epsilon W.$$

The second property is that the edge of the hypothesis that $\mathcal{A}$ returns is close to the best possible

Algorithm 1: `MILearn`$^{\mathcal{A}}$

```
Assumptions:
    • $\mathcal{H} \in [-1,+1]^{\mathcal{X}}$
    • Algorithm $\mathcal{A}$ receives a weighted instance sample and returns a hypothesis in $\mathcal{H}$.
Input:
    • $\overline{S} \triangleq \{(w_i, \bar{\mathbf{x}}_i, y_i)\}_{i\in[m]}$ — a labeled and weighted sample of bags,
    • $\psi$ — an $(a,b,c,d)$-bounded bag-function.
Output: $h_\circ \in \overline{\mathcal{H}}_+$.
$\alpha_{(+1)} \leftarrow a$, $\alpha_{(-1)} \leftarrow c$.
$S_I \leftarrow \{(\alpha_{y_i} \cdot w_i, x_i[j], y_i)\}_{i\in[m], j\in[r]}$.
$h_I \leftarrow \mathcal{A}(S_I)$.
if $\Gamma(\overline{h}_I, \overline{S}) \geq \Gamma(\mathbf{h}_{\text{pos}}, \overline{S})$ then
    $h_\circ \leftarrow \overline{h}_I$,
else
    $h_\circ \leftarrow \mathbf{h}_{\text{pos}}$.
Return $h_\circ$.
```

one on the input sample, but only compared to the edges that can be achieved by hypotheses that label all the negative instances of S with -1. For a hypothesis class $\mathcal{H}$ and a distribution D over labeled examples, we denote the set of hypotheses in $\mathcal{H}$ that label all negative examples in D with -1, by

$$\Omega(\mathcal{H}, D) = \{h \in \mathcal{H} \mid \mathbb{P}_{(X,Y)\sim D}[h(X) = -1 \mid Y = -1] = 1\}.$$

For a labeled sample S, $\Omega(\mathcal{H}, S) \triangleq \Omega(\mathcal{H}, U_S)$ where U_S is the uniform distribution over the examples in S.

Definition 8.6 (one-sided-ϵ-optimal)**.** *An algorithm $\mathcal{A}$ that accepts a weighted and labeled sample of instances in $\mathcal{X}$ and returns a hypothesis in $\mathcal{H}$ is* one-sided-ϵ-optimal *if for all weighted samples $S \subseteq \mathbb{R}_+ \times \mathcal{X} \times \{\pm 1\}$ with total weight W,*

$$\Gamma(\mathcal{A}(S), S) \geq \sup_{h\in\Omega(\mathcal{H},S)} \Gamma(h, S) - \epsilon W.$$

Clearly, any algorithm which is ϵ-optimal is also one-sided-ϵ-optimal, thus the first requirement from $\mathcal{A}$ is stronger. In our results below we compare the edge achieved using `MILearn` to the best possible edge for the sample $\overline{S}$. Denote the best edge achievable for $\overline{S}$ by a hypothesis in $\mathcal{H}$ by

$$\gamma^* \triangleq \sup_{h \in \overline{\mathcal{H}}} \Gamma(h, \overline{S}).$$

We denote by γ^*_+ the best edge that can be achieved by a hypothesis in $\Omega(\overline{\mathcal{H}}, \overline{S})$. Formally,

$$\gamma^*_+ \triangleq \sup_{h \in \Omega(\overline{\mathcal{H}}, \overline{S})} \Gamma(h, \overline{S}).$$

Denote the weight of the positive bags in the input sample $\overline{S}$ by $W_+ = \sum_{i:y_i=+1} w_i$ and the weight of the negative bags by $W_- = \sum_{i:y_i=-1} w_i$. We will henceforth assume without loss of generality that the total weight of all bags in the input sample is 1, that is $W_+ + W_- = 1$.

Note that for any (a, b, c, d)-bounded ψ, if there exists any sequence $z_1, \ldots, z_n$ such that $\psi(z_1, \ldots, z_n) = -1$, then

$$a \sum_{i \in [n]} z_i + b \leq -1 \leq c \sum_{i \in [n]} z_i + d. \tag{8.4}$$

This implies

$$\frac{-1 - d}{c} \leq \sum_{i \in [n]} z_i \leq \frac{-1 - b}{a}.$$

Rearranging, we get $d - \frac{c}{a} b - \frac{c}{a} + 1 \geq 0$, with equality if Eq. (8.4) holds with equalities. The next theorem provides a guarantee for `MILearn` that depends on the tightness of this inequality for the given bag function. As evident from Theorem 8.1, to guarantee a positive margin for the output of `AdaBoost`* when used with `MILearn` as the weak learner, we need to guarantee that the edge of the hypothesis returned by `MILearn` is always positive. Since the best edge cannot be more than 1, we emphasize in the theorem below that the edge achieved by `MILearn` is positive at least when the best edge is 1 (and possibly also for smaller edges, depending on the parameters). We subsequently show how these general guarantees translate to a specific result for the $\max$ function, and other bag functions with the same boundedness properties.

Theorem 8.7. *Let $r \in \mathbb{N}$ and $R \subseteq [r]$. Let $\psi : [-1, +1]^{(R)} \to [-1, +1]$ be an (a, b, c, d)-bounded bag-function such that $0 < a \leq c$. Let $\epsilon \in [0, \frac{1}{rc})$, and assume that $d - \frac{c}{a} b - \frac{c}{a} + 1 = \eta$. Denote $Z = \frac{c}{a}$. Consider running the algorithm `MILearn`$^{\mathcal{A}}$ with a weighted bag sample $\overline{S}$ of total weight 1, and let $h_\circ$ be the hypothesis returned by `MILearn`$^{\mathcal{A}}$. Then*

1. *If $\mathcal{A}$ is ϵ-optimal then*

$$\Gamma(h_\circ, \overline{S}) \geq \frac{Z\gamma^* - Z + \frac{1}{Z} - \frac{\eta}{2}(1 + \frac{1}{Z}) - rc\epsilon}{1 + (1 - \frac{\eta}{2})(1 - \frac{1}{Z})}.$$

   Thus, $\Gamma(h_\circ, \overline{S}) > 0$ whenever

$$\gamma^* > 1 - \frac{1}{Z^2} + \frac{\eta}{2}(\frac{1}{Z} + \frac{1}{Z^2}) + \frac{rc\epsilon}{Z}.$$

   In particular, if $\eta \leq 2(1 - rc\epsilon)/(Z+1)$ and $\gamma^* = 1$ then $\Gamma(h_\circ, \overline{S}) > 0$.

2. *If $\mathcal{A}$ is one-sided-ϵ-optimal, and $\psi(z_1, \ldots, z_n) = -1$ only if $z_1 = \ldots = z_n = -1$, then*

$$\Gamma(h_\circ, \overline{S}) \geq \frac{\gamma_+^* - \frac{\eta}{2}(Z+1) - rc\epsilon Z}{2Z - 1 - \frac{\eta}{2}(Z-1)}.$$

   Thus, $\Gamma(h_\circ, \overline{S}) > 0$ whenever

$$\gamma_+^* > \frac{\eta}{2}(Z+1) + rc\epsilon Z.$$

   In particular, if $\eta \leq 2(1 - rc\epsilon Z)/(Z+1)$ and $\gamma_+^* = 1$ then $\Gamma(h_\circ, \overline{S}) > 0$.

The first step in the proof of the theorem, is to provide a guarantee for the edge achieved on the bag sample by the hypothesis returned by $\mathcal{A}$ in step (3) of the algorithm. This is done in the following lemma.

Lemma 8.8. *Assume $\psi : [-1, +1]^{(R)} \to [-1, +1]$ is an (a, b, c, d)-bounded bag function with $0 < a \leq c$, and denote $Z = \frac{c}{a}$. Consider running the algorithm* `MILearn` *with a weighted bag sample $\overline{S}$ of total weight 1. Let h_I be the hypothesis returned by the oracle $\mathcal{A}$ in step (3) of* `MILearn`*. Let W be the total weight of the sample S_I created in* `MILearn`*, step (2). Then*

1. *If $\mathcal{A}$ is ϵ-optimal,*

$$\Gamma(\overline{h}_I, \overline{S}) \geq Z\gamma^* + (\frac{1}{Z} - Z + (1 - \frac{1}{Z})(d - Zb))W_+ + Zb - d - \epsilon W.$$

2. *If $\mathcal{A}$ is one-sided-ϵ-optimal, and $\psi(z_1, \ldots, z_n) = -1$ only if $z_1 = \ldots = z_n = -1$, then*

$$\Gamma(\overline{h}_I, \overline{S}) \geq \frac{1}{Z}\gamma_+^* + (\frac{1}{Z} - Z + (1 - \frac{1}{Z})(d - Zb))W_+ + Zb - d + Z - \frac{1}{Z} - \epsilon W.$$

Proof. For all $h \in \mathcal{H}$, and for all $\bar{\mathbf{x}} = (x_1, \ldots, x_n) \in \mathcal{X}^{(R)}$ we have $\overline{h}(\bar{\mathbf{x}}) = \psi(h(x_1), \ldots, h(x_n))$.

Since ψ is (a, b, c, d)-bounded, it follows that

$$a \sum_{x \in \bar{\mathbf{x}}} h(x) + b \leq \overline{h}(\bar{\mathbf{x}}) \leq c \sum_{x \in \bar{\mathbf{x}}} h(x) + d. \tag{8.5}$$

In addition, since a and c are positive we also have

$$(\overline{h}(\bar{\mathbf{x}}) - d)/c \leq \sum_{x \in \bar{\mathbf{x}}} h(x) \leq (\overline{h}(\bar{\mathbf{x}}) - b)/a. \tag{8.6}$$

Assume the input bag sample is $\overline{S} = \{(w_i, \bar{\mathbf{x}}_i, y_i)\}_{i \in [m]}$. Denote $I_+ = \{i \in [m] \mid y_i = +1\}$ and $I_- = \{i \in [m] \mid y_i = -1\}$. Let $h \in \mathcal{H}$ be a hypothesis. We have

$$\begin{aligned}\Gamma(\overline{h}, \overline{S}) &= \sum_{i \in I_+} w_i \overline{h}(\bar{\mathbf{x}}_i) - \sum_{i \in I_-} w_i \overline{h}(\bar{\mathbf{x}}_i) \\ &\geq \sum_{i \in I_+} w_i (a \sum_{x \in \bar{\mathbf{x}}_i} h(x) + b) - \sum_{i \in I_-} w_i (c \sum_{x \in \bar{\mathbf{x}}_i} h(x) + d) \qquad (8.7) \\ &= \sum_{i \in I_+} w_i a \sum_{x \in \bar{\mathbf{x}}_i} h(x) - \sum_{i \in I_-} w_i c \sum_{x \in \bar{\mathbf{x}}_i} h(x) + \sum_{i \in I_+} w_i b - \sum_{i \in I_-} w_i d. \qquad (8.8)\end{aligned}$$

line (8.7) follows from Eq. (8.5). As evident by steps (1,2) of `MILearn`, In the sample S_I all instances from positive bags have weight $\alpha(+1) = a$, and all instances from negative bags have weight $\alpha(-1) = c$. Therefore

$$\Gamma(h, S_I) = \sum_{i \in [m]} \sum_{x \in \bar{\mathbf{x}}_i} w_i y_i \alpha(y_i) h(x) = \sum_{i \in I_+} w_i a \sum_{x \in \bar{\mathbf{x}}_i} h(x) - \sum_{i \in I_-} w_i c \sum_{x \in \bar{\mathbf{x}}_i} h(x).$$

Combining this equality with Eq. (8.8) we get

$$\Gamma(\overline{h}, \overline{S}) \geq \Gamma(h, S_I) + \sum_{i \in I_+} w_i b - \sum_{i \in I_-} w_i d.$$

Since $\sum_{i \in I_+} w_i = W_+$ and $\sum_{i \in I_-} w_i = W_- = 1 - W_+$, it follows that

$$\Gamma(\overline{h}, \overline{S}) \geq \Gamma(h, S_I) + bW_+ - dW_- = \Gamma(h, S_I) + (b + d)W_+ - d. \tag{8.9}$$

Now, for any hypothesis h we can conclude from Eq. (8.6) that

$$\begin{aligned}
\Gamma(h, S_I) &= \sum_{i \in I_+} a w_i \sum_{x \in \bar{\mathbf{x}}_i} h(x) - \sum_{i \in I_-} c w_i \sum_{x \in \bar{\mathbf{x}}_i} h(x) \\
&\geq \sum_{i \in I_+} a w_i (\overline{h}(\bar{\mathbf{x}}_i) - d)/c - \sum_{i \in I_-} c w_i (\overline{h}(\bar{\mathbf{x}}_i) - b)/a \\
&= \sum_{i \in I_+} \frac{a}{c} w_i \overline{h}(\bar{\mathbf{x}}_i) - \sum_{i \in I_-} \frac{c}{a} w_i \overline{h}(\bar{\mathbf{x}}_i) - \sum_{i \in I_+} a d w_i / c + \sum_{i \in I_-} c b w_i / a \\
&= \frac{c}{a} \Gamma(\overline{h}, \overline{S}) + (\frac{a}{c} - \frac{c}{a}) \sum_{i \in I_+} w_i \overline{h}(\bar{\mathbf{x}}_i) - \frac{ad}{c} W_+ + \frac{cb}{a} W_- \\
&= \frac{c}{a} \Gamma(\overline{h}, \overline{S}) + (\frac{a}{c} - \frac{c}{a}) \sum_{i \in I_+} w_i \overline{h}(\bar{\mathbf{x}}_i) - (\frac{ad}{c} + \frac{cb}{a}) W_+ + \frac{cb}{a}.
\end{aligned}$$

In the last equality we used the fact that $W_- = 1 - W_+$. Since $Z = \frac{c}{a}$, it follows that

$$\Gamma(h, S_I) \geq Z\Gamma(\overline{h}, \overline{S}) + (\frac{1}{Z} - Z) \sum_{i \in I_+} w_i \overline{h}(\bar{\mathbf{x}}_i) - (\frac{d}{Z} + Zb) W_+ + Zb. \tag{8.10}$$

We will now lower-bound the right-hand-side of Eq. (8.10). Note that $\frac{1}{Z} - Z \leq 0$ since $c \geq a$. Therefore we need an upper bound for $\sum_{i \in I_+} w_i \overline{h}(\bar{\mathbf{x}}_i)$. We consider each of the two cases in the statement of the lemma separately.

Case 1: $\mathcal{A}$ is ϵ-optimal We have $\sum_{i \in I_+} w_i \overline{h}(\bar{\mathbf{x}}_i) \leq \sum_{i \in I_+} w_i = W_+$. Therefore, by Eq. (8.10) for any $h \in \mathcal{H}$

$$\Gamma(h, S_I) \geq Z\Gamma(\overline{h}, \overline{S}) + (\frac{1}{Z} - Z - \frac{d}{Z} - Zb) W_+ + Zb. \tag{8.11}$$

For a natural n, set h^n_* such that $\Gamma(\overline{h}^n_*, \overline{S}) \geq \gamma^* - \frac{1}{n}$. We have (see explanations below)

$$\begin{aligned}
\Gamma(\overline{h}_I, \overline{S}) &\geq \Gamma(h_I, S_I) + (b + d) W_+ - d \qquad (8.12)\\
&\geq \Gamma(h^n_*, S_I) + (b + d) W_+ - d - \epsilon W \qquad (8.13)\\
&\geq Z\Gamma(\overline{h}^n_*, \overline{S}) + (\frac{1}{Z} - Z - \frac{d}{Z} - Zb) W_+ + Zb + (b + d) W_+ - d - \epsilon W \qquad (8.14)\\
&= Z\Gamma(\overline{h}^n_*, \overline{S}) + (\frac{1}{Z} - Z + (1 - \frac{1}{Z})(d - Zb)) W_+ + Zb - d - \epsilon W \\
&\geq Z(\gamma^* - \frac{1}{n}) + (\frac{1}{Z} - Z + (1 - \frac{1}{Z})(d - Zb)) W_+ + Zb - d - \epsilon W.
\end{aligned}$$

Eq. (8.12) is a restatement of Eq. (8.9). Eq. (8.13) follows from the ϵ-optimality of $\mathcal{A}$. Eq. (8.14) follows from Eq. (8.11). By taking $n \to \infty$, this inequality proves case (1) of the lemma.

Case 2: $\mathcal{A}$ is one-sided-ϵ-optimal We have $\sum_{i \in I_+} w_i \overline{h}(\bar{\mathbf{x}}_i) \leq \sum_{i \in I_+} w_i = W_+$. Let $\overline{h} \in \Omega(\overline{\mathcal{H}}, \overline{S})$. Then for all $i \in I_-$, $\overline{h}(\bar{\mathbf{x}}_i) = -1$. Therefore

$$\begin{aligned}
\Gamma(\overline{h}, \overline{S}) &= \sum_{i \in I_+} w_i \overline{h}(\bar{\mathbf{x}}_i) - \sum_{i \in I_-} w_i \overline{h}(\bar{\mathbf{x}}_i) \\
&= \sum_{i \in I_+} w_i \overline{h}(\bar{\mathbf{x}}_i) + \sum_{i \in I_-} w_i \\
&= \sum_{i \in I_+} w_i \overline{h}(\bar{\mathbf{x}}_i) + W_-.
\end{aligned}$$

Therefore $\sum_{i \in I_+} w_i \overline{h}(\bar{\mathbf{x}}_i) = \Gamma(\overline{h}, \overline{S}) - W_- = \Gamma(\overline{h}, \overline{S}) + W_+ - 1$. Combining this with Eq. (8.10) we get

$$\begin{aligned}
\Gamma(h, S_I) &\geq Z\Gamma(\overline{h}, \overline{S}) + (\frac{1}{Z} - Z) \sum_{i \in I_+} w_i \overline{h}(\bar{\mathbf{x}}_i) - (\frac{d}{Z} + Zb)W_+ + Zb \\
&= Z\Gamma(\overline{h}, \overline{S}) + (\frac{1}{Z} - Z)(\Gamma(\overline{h}, \overline{S}) + W_+ - 1) - (\frac{d}{Z} + Zb)W_+ + Zb. \\
&= \frac{1}{Z}\Gamma(\overline{h}, \overline{S}) + (\frac{1}{Z} - Z - \frac{d}{Z} - Zb)W_+ + Zb - \frac{1}{Z} + Z. \qquad (8.15)
\end{aligned}$$

For a natural n, set $\overline{h}^n_+ \in \Omega(\overline{\mathcal{H}}, \overline{S})$ such that $\Gamma(\overline{h}^n_+, \overline{S}) \geq \gamma^*_+ - \frac{1}{n}$. For all bags $i \in I_-$, $\overline{h}^n_+(\bar{\mathbf{x}}_i) = -1$. Thus $\psi(h^n_+(x_i[1]), \ldots, h^n_+(x_i[|\bar{\mathbf{x}}_i|)) = -1$. By the assumption on ψ in case (2) of the lemma, this implies that for all $i \in I_-, j \in [|\bar{\mathbf{x}}_i|]$, $h^n_+(x_i[j]) = -1$. Therefore $h^n_+ \in \Omega(\mathcal{H}, S_I)$. We have (see explanations below)

$$\begin{aligned}
\Gamma(\overline{h}_I, \overline{S}) &\geq \Gamma(h_I, S_I) + (b + d)W_+ - d \qquad (8.16) \\
&\geq \Gamma(h^n_+, S_I) + (b + d)W_+ - d - \epsilon W \qquad (8.17) \\
&\geq \frac{1}{Z}\Gamma(\overline{h}^n_+, \overline{S}) + (\frac{1}{Z} - Z - \frac{d}{Z} - Zb)W_+ + Zb - \frac{1}{Z} + Z + (b + d)W_+ - d - \epsilon W \\
&\qquad (8.18) \\
&= \frac{1}{Z}\Gamma(\overline{h}^n_+, \overline{S}) + (\frac{1}{Z} - Z + (1 - \frac{1}{Z})(d - Zb))W_+ + Zb - d + Z - \frac{1}{Z} - \epsilon W \\
&\geq \frac{1}{Z}(\gamma^*_+ - \frac{1}{n}) + (\frac{1}{Z} - Z + (1 - \frac{1}{Z})(d - Zb))W_+ + Zb - d + Z - \frac{1}{Z} - \epsilon W.
\end{aligned}$$

Eq. (8.16) is a restatement of Eq. (8.9). Eq. (8.17) follows from the one-sided-ϵ-optimality of $\mathcal{A}$

and the fact that $h_+^n \in \Omega(\mathcal{H}, S_I)$. Eq. (8.18) follows from Eq. (8.15). By considering $n \to \infty$, this proves the second part of the lemma. □

Proof of Theorem 8.7. `MILearn` selects the hypothesis with the best edge on $\overline{S}$ between $\overline{h}_I$ and $\mathbf{h}_{\text{pos}}$. Therefore

$$\Gamma(h_\circ, \overline{S}) = \max(\Gamma(\mathbf{h}_{\text{pos}}, \overline{S}), \Gamma(\overline{h}_I, \overline{S})).$$

We have

$$\Gamma(\mathbf{h}_{\text{pos}}, \overline{S}) = \sum_{i \in [m]} w_i y_i \mathbf{h}_{\text{pos}}(\bar{\mathbf{x}}_i) = \sum_{i \in [m]} w_i y_i = W_+ - W_- = 2W_+ - 1.$$

Thus

$$\Gamma(h_\circ, \overline{S}) = \max(2W_+ - 1, \Gamma(\overline{h}_I, \overline{S})). \tag{8.19}$$

We now lower-bound $\Gamma(h_\circ, \overline{S})$ by bounding $\Gamma(\overline{h}_I, \overline{S})$ separately for the two cases of the theorem. Let W be the total weight of S_I. Since $R \subseteq [r]$, $a \leq c$, and $\sum_{i \in [m]} w_i = 1$, we have

$$W = \sum_{i: y_i = +1} \sum_{x \in \bar{\mathbf{x}}_i} a w_i + \sum_{i: y_i = -1} \sum_{x \in \bar{\mathbf{x}}_i} c w_i \leq rc \sum_{i \in [m]} w_i = rc \tag{8.20}$$

Case 1: $\mathcal{A}$ is ϵ-optimal From Lemma 8.8 and Eq. (8.20) we have

$$\begin{aligned}
\Gamma(\overline{h}_I, \overline{S}) &\geq Z\gamma^* + (\frac{1}{Z} - Z + (1 - \frac{1}{Z})(d - Zb))W_+ + Zb - d - rc\epsilon \\
&= Z\gamma^* + (\frac{1}{Z} - Z + (1 - \frac{1}{Z})(Z - 1 + \eta))W_+ - (Z - 1 + \eta) - rc\epsilon \\
&= Z\gamma^* + (\eta - 2)(1 - \frac{1}{Z})W_+ + 1 - \eta - Z - rc\epsilon.
\end{aligned}$$

The second line follows from the assumption $d - Zb - Z + 1 = \eta$. Combining this with Eq. (8.19) we get

$$\Gamma(h_\circ, \overline{S}) \geq \max\{2W_+ - 1,\ Z\gamma^* + (\eta - 2)(1 - \frac{1}{Z})W_+ + 1 - \eta - Z - rc\epsilon\}.$$

The right-hand-side is minimal when the two expressions in the maximum are equal. This occurs when

$$W_+ = W_\circ \triangleq \frac{Z\gamma^* + 2 - \eta - Z - rc\epsilon}{2 + (2 - \eta)(1 - \frac{1}{Z})}.$$

Therefore, for any value of W_+

$$\Gamma(h_\circ, \overline{S}) \geq 2W_\circ - 1 = \frac{Z\gamma^* - Z + \frac{1}{Z} - \frac{\eta}{2}(1 + \frac{1}{Z}) - rc\epsilon}{1 + (1 - \frac{\eta}{2})(1 - \frac{1}{Z})}.$$

Case 2: $\mathcal{A}$ is one-sided-ϵ-optimal From Lemma 8.8 and Eq. (8.20) we have

$$\begin{aligned}\Gamma(\overline{h}_I, \overline{S}) &\geq \frac{1}{Z}\gamma_+^* + (\frac{1}{Z} - Z + (1 - \frac{1}{Z})(d - Zb))W_+ + Zb - d + Z - \frac{1}{Z} - rc\epsilon\\ &= \frac{1}{Z}\gamma_+^* + (\frac{1}{Z} - Z + (1 - \frac{1}{Z})(Z - 1 + \eta))W_+ - (Z - 1 + \eta) + Z - \frac{1}{Z} - rc\epsilon\\ &= \frac{1}{Z}\gamma_+^* + (\eta - 2)(1 - \frac{1}{Z})W_+ + 1 - \eta - \frac{1}{Z} - rc\epsilon.\end{aligned}$$

The second line follows from the assumption $d - Zb = Z - 1 + \eta$. Combining this with Eq. (8.19) we get

$$\Gamma(h_\circ, \overline{S}) \geq \max\{2W_+ - 1, \ \frac{1}{Z}\gamma_+^* + (\eta - 2)(1 - \frac{1}{Z})W_+ + 1 - \eta - \frac{1}{Z} - rc\epsilon\}.$$

The right-hand-side is minimal when the two expressions in the maximum are equal. This occurs when

$$W_+ = W_\circ \triangleq \frac{\gamma_+^* - 1 + (2 - \eta - rc\epsilon)Z}{2Z + (2 - \eta)(Z - 1)}.$$

Substituting W_+ for $W_\circ$ in the lower bound, we get

$$\Gamma(h_\circ, \overline{S}) \geq 2W_\circ - 1 = \frac{\gamma_+^* - \frac{\eta}{2}(Z + 1) - rc\epsilon Z}{2Z - 1 - \frac{\eta}{2}(Z - 1)}.$$

□

Theorem 8.7 is stated in general terms, as it holds for any bounded ψ. In particular, if ψ is any function between an average and a $\max$, including any of the p-norm bag functions ψ_p defined in Def. 6.3, we can simplify the result, as captured by the following corollary.

Corollary 8.9. *Let $\mathcal{H} \subseteq [-1, +1]^{\mathcal{X}}$. Let $R \subseteq [r]$, and $\epsilon \in [0, \frac{1}{r})$. Assume a bag function $\psi : [-1, +1]^{(R)} \rightarrow [-1, +1]$ such that for any $z_1, \ldots, z_n \in [-1, +1]$,*

$$\frac{1}{n}\sum_{i \in [n]} z_i \leq \psi(z_1, \ldots, z_n) \leq \max_{i \in [n]} z_i.$$

Let $h_\circ$ be the hypothesis returned by $\texttt{MILearn}^{\mathcal{A}}$. *Then*

1. *If $\mathcal{A}$ is ϵ-optimal for some $\epsilon \in [0, 1/r]$, then*

$$\Gamma(h_\circ, \overline{S}) \geq \frac{r^2\gamma^* + 1 - r^2(1+\epsilon)}{2r-1}.$$

   Thus $\Gamma(h_\circ, \overline{S}) > 0$ whenever $\gamma^* \geq 1 - \frac{1}{r^2} + \frac{\epsilon}{r}$. In particular, if $\gamma^* = 1$ then $\Gamma(h_\circ, \overline{S}) > 0$.

2. *If $\mathcal{A}$ is one-sided-ϵ-optimal some $\epsilon \in [0, 1/r^2]$, then*

$$\Gamma(h_\circ, \overline{S}) \geq \frac{\gamma_+^* - r^2\epsilon}{2r-1}.$$

   Thus $\Gamma(h_\circ, \overline{S}) > 0$ whenever $\gamma_+^* > r^2\epsilon$. In particular, if $\gamma_+^* = 1$ then $\Gamma(h_\circ, \overline{S}) > 0$.

Proof. Let $z_1, \ldots, z_n \in [-1, +1]$. We have

$$\max_{i\in[n]} z_i \leq \sum_{i\in[n]} z_i - (n-1)\min(z_i) \leq \sum_{i\in[n]} z_i + n - 1.$$

Therefore, by the assumption on ψ, for any $n \in R$

$$\psi(z_1, \ldots, z_n) \leq \sum_{i\in[n]} z_i + n - 1 \leq \sum_{i\in[n]} z_i + r - 1.$$

In addition

$$\frac{1}{r}\sum_{i\in[n]} z_i \leq \frac{1}{n}\sum_{i\in[n]} z_i \leq \psi(z_1, \ldots, z_n).$$

Therefore ψ is $(\frac{1}{r}, 0, 1, r-1)$-bounded. It follows that $Z = r$ in this case, and $d - Zb - Z + 1 = 0$. Claim (1) follows by applying case (1) of Theorem 8.7 with $\eta = 0$.

For claim (2) we apply case (2) of Theorem 8.7. Thus we need to show that if $\psi(z_1, \ldots, z_n) = -1$ and $z_1, \ldots, z_n \in [-1, +1]$, then $z_1 = \ldots = z_n = -1$. We have that

$$-1 \leq \frac{1}{n}\sum_{i\in[n]} z_i \leq \psi(z_1, \ldots, z_n) \leq -1.$$

Therefore $\frac{1}{n}\sum_{i\in[n]} z_i = -1$. Since no z_i can be smaller than -1, $z_1 = \ldots = z_n = -1$. Thus case (2) of Theorem 8.7 holds. We get our claim (2) directly by subsituting the boundedness parameters of ψ in Theorem 8.7 case (2). □

8.3 From Single-Instance Learning to Multi-Instance Learning

In this section we combine the guarantees on `MILearn` with the guarantees on `AdaBoost`*, to show that efficient agnostic PAC-learning of the underlying instance hypothesis $\mathcal{H}$ implies efficient PAC-learning of MIL. For simplicity we formalize the results for the natural case where the bag function is $\psi = \max$. Results for other bounded bag functions can be derived in a similar fashion.

First, we formally define the notions of agnostic and one-sided PAC-learning algorithms. We then show that given an algorithm on instances that satisfies one of these definitions, we can construct an algorithm for MIL which approximately maximizes the margin on an input bag sample. Specifically, if the input bag sample is realizable by $\overline{\mathcal{H}}$, then the MIL algorithm we propose will find a convex combination of bag hypotheses that classifies the sample with zero error, and with a positive margin. Combining this with the margin-based generalization guarantees mentioned in Section 8.1, we conclude that we have an efficient PAC-learner for MIL.

Definition 8.10 (Agnostic PAC-learner and one-sided PAC-learner)**.** *Let $\mathcal{B}(\epsilon, \delta, S)$ be an algorithm that accepts as input $\delta, \epsilon \in (0, 1)$, and a labeled sample $S \in (\mathcal{X} \times \{\pm 1\})^m$, and emits as output a hypothesis $h \in \mathcal{H}$. $\mathcal{B}$ is an* agnostic PAC-learner *for $\mathcal{H}$ with complexity $c(\epsilon, \delta)$ if $\mathcal{B}$ runs for no more than $c(\epsilon, \delta)$ steps, and for any probability distribution D over $\mathcal{X} \times \{\pm 1\}$, if S is an i.i.d. sample from D of size $c(\epsilon, \delta)$, then with probability at least $1 - \delta$ over S and the randomization of $\mathcal{B}$,*

$$\Gamma(\mathcal{B}(\epsilon, \delta, S), D) \geq \sup_{h \in \mathcal{H}} \Gamma(h, D) - \epsilon.$$

$\mathcal{B}$ is a one-sided PAC-learner *if under the same conditions, with probability at least $1 - \delta$*

$$\Gamma(\mathcal{B}(\epsilon, \delta, S), D) \geq \sup_{h \in \Omega(\mathcal{H}, D)} \Gamma(h, D) - \epsilon.$$

Given an agnostic PAC-learner $\mathcal{B}$ for $\mathcal{H}$ and parameters $\epsilon, \delta \in (0, 1)$, the algorithm $\mathcal{O}^{\mathcal{B}}_{\epsilon,\delta}$, listed above as Alg. 2, is an ϵ-optimal algorithm with probability $1-\delta$. Similarly, if $\mathcal{B}$ is a one-sided PAC-learner, then $\mathcal{O}^{\mathcal{B}}_{\epsilon,\delta}$ is a one-sided-ϵ-optimal algorithm with probability $1 - \delta$. Our MIL algorithm is then simply `AdaBoost`* with `MILearn`$^{\mathcal{O}^{\mathcal{B}}_{\epsilon,\delta}}$ as the (high probability) weak learner. It is easy to see that this algorithm learns a convex combination of hypotheses from $\overline{\mathcal{H}}_+$. We also show below that under certain conditions this convex combination induces a positive margin on the input bag sample with high probability. Given this guaranteed margin, we bound the generalization error of the learning algorithm via Eq. (8.1).

The computational complexity of $\mathcal{O}^{\mathcal{B}}_{\epsilon,\delta}$ is polynomial in $c(\epsilon, \delta)$ and in the instance-sample size m. Therefore, the computational complexity of `MILearn`$^{\mathcal{O}^{\mathcal{B}}_{\epsilon,\delta}}$ is polynomial in $c(\epsilon, \delta)$ and in N, where N is the total number of instances in the input bag sample $\overline{S}$.

```
Algorithm 2: $\mathcal{O}^{\mathcal{B}}_{\epsilon,\delta}$
  Assumptions:
    • $\epsilon, \delta \in (0, 1)$.
    • $\mathcal{B}$ receives a labeled instance sample as input and returns a hypothesis in $\mathcal{H}$.
    • Algorithm $\mathcal{B}$ is a one-sided (or agnostic) PAC-learning algorithm with complexity $c(\epsilon, \delta)$.
  Input: A labeled and weighted instance sample $S = \{(w_i, x_i, y_i)\}_{i\in[m]} \subseteq \mathbb{R}_+ \times \mathcal{X} \times \{\pm 1\}$.
  Output: A hypothesis in $\mathcal{H}$
  For all $i \in [m]$, $p_i \leftarrow w_i / \sum_{i\in[m]} w_i$.
  For each $t \in [c(\epsilon, \delta)]$, independently draw a random $j_t$ such that $j_t = i$ with probability $p_i$.
  $\tilde{S} \leftarrow \{(x_{j_t}, y_{j_t})\}_{t\in[c(\epsilon,\delta)]}$.
  $h \leftarrow \mathcal{B}(\tilde{S})$
  Return $h$.
```

For 1-Lipschitz bag functions which have desired boundedness properties, both the sample complexity and the computational complexity of the proposed MIL algorithm are polynomial in the maximal bag size and linear in the complexity of the underlying instance hypothesis class. This is formally stated in the following theorem, for the case of a realizable distribution over labeled bags. Note that in particular, the theorem holds for all the p-norm bag-functions, since they are 1-Lipschitz and satisfy the boundedness conditions.

Theorem 8.11. *Let $\mathcal{H} \subseteq [-1, +1]^{\mathcal{X}}$ be a hypothesis class with pseudo-dimension d. Let $\mathcal{B}$ be a one-sided PAC-learner for $\mathcal{H}$ with complexity $c(\epsilon, \delta)$. Let $r \in \mathbb{N}$, and let $R \subseteq [r]$. Assume that the bag function $\psi : [-1, +1]^{(R)} \to [-1, +1]$ is 1-Lipschitz with respect to the infinity norm, and that for any $(z_1, \ldots, z_n) \in [-1, +1]^{(R)}$*

$$\frac{1}{n}\sum_{i\in[n]} z_i \leq \psi(z_1, \ldots, z_n) \leq \max_{i\in[n]} z_i.$$

Assume that $\overline{\mathcal{H}}$ is compact with respect to any sample of size m. Let D be a distribution over $\mathcal{X}^{(R)} \times \{\pm 1\}$ which is realizable by $\overline{\mathcal{H}}$, that is there exists an $h \in \mathcal{H}$ such that $\mathbb{P}_{(\bar{\mathbf{X}},Y)\sim D}[\overline{h}(\bar{\mathbf{X}}) = Y] = 1$. Assume $m \geq 10d\ln(er)$, and let $\epsilon = \frac{1}{2r^2}$ and $k = 32(2r-1)^2\ln(m)$.

For all $\delta \in (0, 1)$, if `AdaBoost`* is executed for k rounds on a random sample $S \sim D^m$, with `MILearn`$^{\mathcal{O}^{\mathcal{B}}_{\epsilon,\delta/2k}}$ as the weak learner, then with probability $1 - \delta$, the classifier $f_\circ$ returned by

`AdaBoost` satisfies*

$$\mathbb{P}_D[Yf(\bar{\mathbf{X}}) \leq 0] \leq O\left(\sqrt{\frac{dr^2 \ln(r)\ln^2(m) + \ln(2/\delta)}{m}}\right). \tag{8.21}$$

Proof. Since $\mathcal{B}$ is a one-sided PAC-learning algorithm, $\mathcal{O}^{\mathcal{B}}_{\epsilon,\delta/2k}$ is one-sided-ϵ-optimal with probability at least $1-\delta/2k$. Therefore, by case (2) of Cor. 8.9, if $\mathtt{MILearn}^{\mathcal{O}^{\mathcal{B}}_{\epsilon,\delta/k}}$ receives a weighted bag sample $S_{\mathbf{w}}$, with probability $1-\delta/2k$ it returns a bag hypothesis $h_\circ \in \overline{\mathcal{H}}_+$ such that

$$\Gamma(h_\circ, S_{\mathbf{w}}) \geq \frac{\sup_{h\in\Omega(\overline{\mathcal{H}},S)} \Gamma(h, S_{\mathbf{w}}) - r^2\epsilon}{2r-1}.$$

Thus, by Cor. 8.3, if `AdaBoost`* runs for k rounds then with probability $1-\delta/2$ it returns a convex combination of hypotheses from $\overline{\mathcal{H}}_+$ such that

$$M(f_\circ, S) \geq \frac{\sup_{f\in\mathrm{co}(\Omega(\overline{\mathcal{H}},S))} M(f,S) - r^2\epsilon}{2r-1} - \sqrt{2\ln m/k}. \tag{8.22}$$

Due to the realizability assumption for D, there is an $h \in \Omega(\overline{\mathcal{H}}, S)$ that classifies correctly the bag sample S. It follows that for any weighting $\mathbf{w} \in \Delta_m$ of S, $\Gamma(h, S_{\mathbf{w}}) = 1$. It is easy to verify that since $\overline{\mathcal{H}}$ is compact with respect to S, then so is $\Omega(\mathcal{H}, S)$. Thus, by Theorem 8.2, $\sup_{f\in\mathrm{co}(\Omega(\overline{\mathcal{H}},S))} M(f,S) = \min_{\mathbf{w}} \sup_{h\in\Omega(\overline{\mathcal{H}},S)} \Gamma(h, S_{\mathbf{w}}) = 1$. Substituting ϵ and k with their values, setting $\sup_{f\in\mathrm{co}(\Omega(\overline{\mathcal{H}},S))} M(f,S) = 1$ in Eq. (8.22) and simplifying, we get that with probability $1-\delta/2$

$$M(f_\circ, S) \geq \frac{1}{8r-4}. \tag{8.23}$$

We would now like to apply the generalization bound in Eq. (8.1), but for this we need to show that Eq. (8.2) holds for $\overline{\mathcal{H}}$. We have the following bound on the covering numbers of $\overline{\mathcal{H}}$, for all $\gamma \in (0,1]$:

$$\mathcal{N}_m(\gamma, \overline{\mathcal{H}}, \infty) \leq \mathcal{N}_{rm}(\gamma, \mathcal{H}, \infty) \leq \left(\frac{erm}{\gamma d}\right)^d.$$

The first inequality is due to Cor. 7.9 and the fact that ψ is 1-Lipschitz, and the second inequality is due to Haussler and Long [1995] and the pseudo-dimension of $\mathcal{H}$ (see Eq. (8.2) above). This implies

$$\mathcal{N}_m(\gamma, \overline{\mathcal{H}}, \infty) \leq \left(\frac{erm}{\gamma d}\right)^d = \left(\frac{em}{\gamma d}\right)^d \cdot e^{d\ln(r)} = \left(\frac{em}{\gamma \cdot 10d\ln(er)}\right)^d \cdot (10\ln(er))^d e^{d\ln(r)}$$
$$= \left(\frac{em}{\gamma \cdot 10d\ln(er)}\right)^d \cdot e^{d(\ln(10\ln(er))+\ln(r))}.$$

Therefore, for $m \geq 10d\ln(er)$

$$\mathcal{N}_m(\gamma, \overline{\mathcal{H}}_+, \infty) \leq 1 + \mathcal{N}_m(\gamma, \overline{\mathcal{H}}, \infty) \leq 1 + \left(\frac{em}{\gamma \cdot 10d\ln(er)}\right)^d \cdot e^{d(\ln(10\ln(er))+\ln(r))}$$
$$\leq \left(\frac{em}{\gamma \cdot 10d\ln(er)}\right)^d \cdot e^{d(\ln(10\ln(er))+\ln(er))}.$$

Now, $\ln(10\ln(er)) + \ln(er) = \ln(10) + \ln(\ln(er)) + \ln(er) \leq \ln(10) + 2\ln(er) \leq 3 + 2\ln(er) \leq 5\ln(er)$. Therefore,

$$\mathcal{N}_m(\gamma, \overline{\mathcal{H}}_+, \infty) \leq \left(\frac{em}{\gamma \cdot 10d\ln(er)}\right)^d \cdot e^{5d\ln(er)} \leq \left(\frac{e^2 m}{\gamma \cdot 10d\ln(er)}\right)^{5d\ln(er)}$$
$$\leq \left(\frac{em}{\gamma \cdot 10d\ln(er)}\right)^{10d\ln(er)}.$$

Thus, for $m \geq 10d\ln(er)$, Eq. (8.2) holds for $\overline{\mathcal{H}}_+$ when substituting d with $d_r = 10d\ln(er)$. This means the generalization bound in Eq. (8.1) holds when substituting d with d_r as well. It follows that with probability $1 - \delta/2$

$$\mathbb{P}[Y f_\circ(X) \leq 0] \leq O\left(\sqrt{\frac{d_r \ln^2(m/d_r)/M^2(f_\circ, S) + \ln(1/\delta)}{m}}\right).$$

Now, with probability $1 - \delta/2$, by Eq. (8.23) we have $M(f_\circ, S) \geq 1/(8r-4)$. Combining the two inequalities and applying the union bound, we have that with probability $1 - \delta$

$$\mathbb{P}[Y f_\circ(X) \leq 0] \leq O\left(\sqrt{\frac{d_r(8r-4)^2 \ln^2(m/d_r) + \ln(2/\delta)}{m}}\right)$$
$$\leq O\left(\sqrt{\frac{10d\ln(er)(8r-4)^2 \ln^2(m) + \ln(2/\delta)}{m}}\right).$$

Due to the O-notation we can simplify the right-hand side to get Eq. (8.21).

□

Similar generalization results for Boosting can be derived for margin-learning as well, using covering-numbers arguments as discussed in Schapire et al. [1998]. The theorem above leads to the following conclusion.

Corollary 8.12. *If there exists a one-sided PAC-learning algorithm for $\mathcal{H}$ with polynomial run-time in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$, then there exists a PAC-learning algorithm for classical MIL on $\mathcal{H}$, which has polynomial run-time in r,$\frac{1}{\epsilon}$ and $\frac{1}{\delta}$.*

Cor. 8.12 is similar in structure to Theorem 6.1: Both state that if the single-instance problem is solvable with one-sided error, then the realizable MIL problem is solvable. Theorem 6.1 applies only to bags with statistically independent instances, while Cor. 8.12 applies to bags drawn from an arbitrary distribution. The assumption of Theorem 6.1 is similarly weaker, as it only requires that the single-instance PAC-learning algorithm handle random one-sided noise, while Cor. 8.12 requires that the single-instance algorithm handle arbitrary one-sided noise. Of course, Cor. 8.12 does not contradict the hardness result provided for APRs in Auer et al. [1998]. Indeed, this hardness result states that if there exists a MIL algorithm for d-dimensional APRs which is polynomial in both r and d, then $\mathcal{RP} = \mathcal{NP}$. Our result does not imply that such an algorithm exists, since there is no known agnostic or one-sided PAC-learning algorithm for APRs which is polynomial in d.

We have shown a simple and general way, independent of hypothesis class, to create a PAC-learning algorithm for classical MIL from a learning algorithm that runs on single instances. Whenever an appropriate polynomial algorithm exists for the non-MIL learning problem, the resulting MIL algorithm will also be polynomial in r. To illustrate, consider for instance the algorithm proposed in Shalev-Shwartz et al. [2010]. This algorithm is an agnostic PAC-learner of fuzzy kernelized half-spaces with an L-Lipschitz transfer function, for some constant $L > 0$. Its time complexity and sample-complexity are at most poly$((\frac{L}{\epsilon})^L \cdot \ln(\frac{1}{\delta}))$. Since this complexity bound is polynomial in $1/\epsilon$ and in $1/\delta$, Cor. 8.12 applies, and we can generate an algorithm for PAC-learning MIL with complexity that depends directly on the complexity of this learner, and is polynomial in r, $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$. More generally, using the construction we proposed here, any advancement in the development of algorithms for agnostic or one-sided learning of any hypothesis class translates immediately to an algorithm for PAC-learning MIL with the same hypothesis class, and with corresponding complexity guarantees.

Chapter 9

Using MIL in a non-MIL Setting

Consider three applications from three different domains: In the first, you want to conduct market research using online ads, to identify which products are attractive. You can put up ads featuring products, but your only feedback is whether or not the ad was clicked. In the second application, consider some chemical or biological problem where the goal is to learn to classify chemical samples based on the result of a chemical experiment. Each experiment is costly, but is possible to conduct an experiment with numerous types of molecules at the same time, and to identify only if a reaction has occurred or not. In the third application, suppose the purpose is to learn a classifier that identifies images with faces, using a large set of labeled images. To obtain this labeled set, one introduces a large set of images to human labelers, who indicate whether the image contains a face or not. We would like to minimize the cost of the human work by reducing the labeling time to a minimum.

These examples come from different domains, but share a common feature: In all of them we have access to practically unlimited data which we can present to a teacher (a human labeler, or some experimental machinery for obtaining a label), but there is a high cost for each label obtained from the teacher. In addition, it is possible to obtain from the teacher a *single label* for *a set of examples* at essentially the same cost as a label for a single example. The single label indicates only if there exists a positive example in the examined set: In the market research application, it is possible to feature several products in one ad. In the chemical experiment task, it may be possible to conduct one large experiment testing several different samples, instead of several experiments, one for each sample. In the face recognition task, one can present test subjects an array of images instead of a single image (see Figure 9.1) and ask them to indicate whether there is a face anywhere in the array of images[1]. In these example application, the main cost of training is the number of

[1] There might be other possibilities, such as asking the labeler to click the exact location of the face in the array, however this might produce a much slower labeling rate than if the labeler clicks only Yes or No buttons

labels, and not the total number of examples.

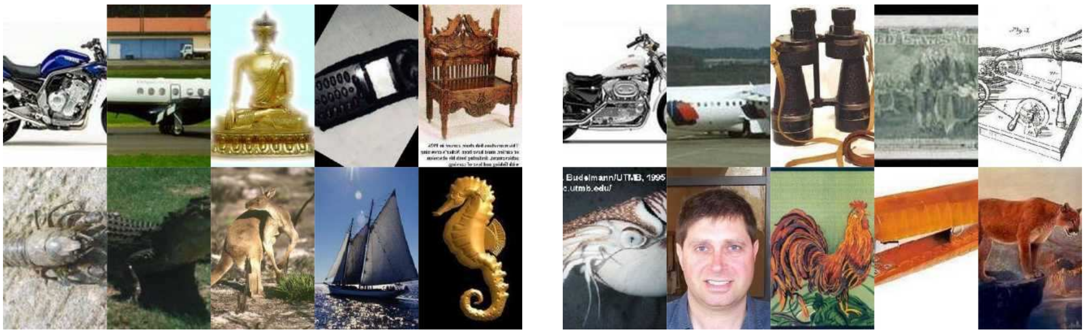

Figure 9.1: A person easily identifies whether there is a face in a bag of images. Left: Negative Label. Right: Positive Label. Images from CALTECH101 [L. Fei-Fei and Perona., 2004].

We consider learning in the setting illustrated by the three example applications, and investigate when it is worthwhile to present a teacher with sets of examples instead of individual examples in this setting. In our model we assume that the cost of obtaining a label does not depend on the size of the set for which the label was obtained, and that obtaining examples to present to the teacher incurs no cost. Therefore, the cost of learning depends only on the number of obtained labels, and the goal is to reduce this number as much as possible using sets of examples of an optimal size.

In the proposed setting, the teacher labels sets of examples using a single label. This can be thought of as a form of Multi-Instance Learning, in which the bags are created by the algorithm, and not by the environment. Moreover, the goal is to learn to classify individual instances and not whole bags. The bag-labeling function in this setting is the classical Boolean OR.

Intuitively, there is an inherent trade-off when obtaining one label for a whole bag: On the one hand, this allows one label to provide information on a large number of examples. On the other hand, this information can be ambiguous, since if the label is positive we do not know which examples in the bag are the positive ones. In this work we investigate this trade-off, and show that it is possible to reduce the number of required labels by presenting bags of examples to the teacher instead of individual examples. After describing the formal setting (Section 9.1), we show, both analytically and experimentally, that using bags can indeed improve performance considerably, for a wide range of problem parameters. We show analytically (Section 9.2) how to *select the bag size* presented to the teacher for optimal performance. In addition, we propose (Section 9.3) a simple and practical algorithm along the lines of Felzenszwalb et al. [2008] for finding a separating hyperplane for individual examples from a training sample composed of labeled bags. Several types of experiments were performed (Section 9.4), on synthetic data sets and on real data sets. The experiments demonstrate the success of the proposed approach for an even wider range of parameters than guaranteed by the analysis.

9.1 Problem Setting

Let $\mathcal{X}$ be the domain of examples, and let $\mathcal{H} \subseteq \{\pm 1\}^{\mathcal{X}}$ be the hypothesis class. We assume a realizable distribution D over labeled examples, and select $c \in \mathcal{H}$ such that $\mathbb{P}_{(X,Y)\sim D}[Y = c(X)] = 1$. The goal of the learner is to return a classifier $h : \mathcal{X} \to \{\pm 1\}$ such that $\ell_{0/1}(h, D)$ is small. The marginal of D over $\mathcal{X}$ and the function c determine the distribution D. Thus in the sequel we identify D with its marginal on $\mathcal{X}$.

We assume that the learner has unlimited access to samples in $\mathcal{X}$ drawn according to D. We consider the case where the main cost incurred in the learning procedure is that of obtaining labels from the teacher, while the cost of presenting examples to the teacher is negligible. We assume that one can ask the teacher to label *bags* of examples using a single label. The teacher's label indicates whether at least one of the examples in the bag is positive. Formally, we fix the bag-labeling function $\psi = \mathrm{OR} \equiv \max$. For every bag $\bar{\mathbf{x}}$ presented to the teacher, the teacher returns a single binary label $\bar{c}(\bar{\mathbf{x}})$. We wish to get low error over *individual examples*, using the smallest possible number of labels. Note that unlike active learning, here the entire sample is generated in advance, with no feedback from the teacher. The following procedure is proposed:

1. Select a bag size r and a sample size m_r;

2. Create m_r bags of size r from $r \cdot m_r$ examples drawn independently from D;

3. Present the bags $\{\bar{\mathbf{x}}_i\}_{i=1}^{m_r}$ to the teacher, and receive m_r labels $\{y_i\}_{i=1}^{m_r}$ such that $y_i = \bar{\mathbf{c}}(\bar{\mathbf{x}}_i)$.

4. Return the hypothesis $h_\circ \in \mathcal{H}$ such that $\overline{h}_\circ$ minimizes the training error over bags:

$$h_\circ = \operatorname*{argmin}_{h \in \mathcal{H}} \sum_{i=1}^{m_r} |\overline{h}(\bar{\mathbf{x}}_i) - y_i|.$$

This procedure is a generalization of the classical empirical risk minimization (ERM) strategy, where the learner finds the hypothesis with minimal training error: For $r = 1$ this procedure is exactly ERM over an i.i.d. sample drawn from the distribution D. For a general r, we use an i.i.d. sample drawn from the distribution D^r. Importantly, regardless of the chosen r, our goal is to minimize $\ell_{0/1}(h_\circ, D)$, the error on *individual examples* drawn from D, and we will measure success based on this goal.

We denote by α the probability of a single example having a positive label, i.e. the frequency of positive examples in D. As we will see, the methods we describe are relevant when α is substantially smaller then half. That is, when positive examples are relatively rare. When the frequency α of positive examples is small, measuring the error becomes tricky: a hypothesis which labels everything as negative has error α, but we typically want a hypothesis that better balances type I and type II errors. In our analysis we assume for simplicity that all the hypotheses in $\mathcal{H}$ have the same probability for a positive label:

$$\forall h \in \mathcal{H}, \mathbb{P}_{(X,Y)\sim D}[h(X) = 1] = \mathbb{P}_{(X,Y)\sim D}[Y = 1] = \alpha. \tag{9.1}$$

That is, all hypothesis are calibrated by the known positive example rate. This assumption implies that the probability of type I errors is identical to the probability of type II errors, and allows us to use the overall error as a single objective even for very small α. In particular, if the learner balances type I and type II errors, or in the realizable case, if the learner seeks a zero empirical error hypothesis, then the hypotheses chosen by the learner satisfies this condition at least approximately. This assumption also implies that the true error of h is in the range $[0, 2\alpha]$.

9.2 Theoretical Analysis

In this section we analyze the procedure described above, and show how it can reduce the required number of labels. We start by analyzing the relationship between the bag size, the sample size, and the resulting true error over individual examples, based on theoretical error bounds. We then use these bounds to choose a bag size r and study the reduction in sample size achieved by the proposed procedure.

9.2.1 The Sample Complexity of Training on Bags

We will base our analysis on standard results, that bound the true error when using ERM on a training sample with a given sample size. These bounds do not suffice by themselves, since they refer to the true error over examples drawn from the same distribution as the training sample. In our case, these results will bound the error over *bags* drawn from D^r, while we wish to bound the true error over *individual examples* drawn from D. We thus start with the following theorem, which provides the relationship between the true error on bags and the true error on individual examples.

Theorem 9.1. *For any $h : \mathcal{X} \to \{0, 1\}$ such that Eq. (9.1) holds, we have*

$$\mathbb{P}[\overline{h}(\bar{\mathbf{X}}) \neq \overline{c}(\bar{\mathbf{X}})] = \kappa_r^\alpha(\mathbb{P}[c(X) \neq h(X)]) \tag{9.2}$$

where $\kappa_r^\alpha(\epsilon) \triangleq 2((1-\alpha)^r - (1-\alpha-\epsilon/2)^r)$.

Proof. Let $X \sim D$ be a random variable over individual examples and $\bar{\mathbf{X}} \sim D^r$ be a random variable over bags. We have

$$\begin{aligned}
&\mathbb{P}[\overline{h}(\bar{\mathbf{X}}) \neq \overline{c}(\bar{\mathbf{X}})] = \\
&= \mathbb{P}[\overline{h}(\bar{\mathbf{X}}) = 0 \wedge \overline{c}(\bar{\mathbf{X}}) = 1] + \mathbb{P}[\overline{h}(\bar{\mathbf{X}}) = 1 \wedge \overline{c}(\bar{\mathbf{X}}) = 0] \\
&= \mathbb{P}[\overline{h}(\bar{\mathbf{X}}) = 0] - \mathbb{P}[\overline{h}(\bar{\mathbf{X}}) = 0 \wedge \overline{c}(\bar{\mathbf{X}}) = 0] + \mathbb{P}[\overline{c}(\bar{\mathbf{X}}) = 0] - \mathbb{P}[\overline{c}(\bar{\mathbf{X}}) = 0 \wedge \overline{h}(\bar{\mathbf{X}}) = 0] \\
&= \mathbb{P}[\overline{h}(\bar{\mathbf{X}}) = 0] + \mathbb{P}[\overline{c}(\bar{\mathbf{X}}) = 0] - 2\mathbb{P}[\overline{c}(\bar{\mathbf{X}}) = 0 \wedge \overline{h}(\bar{\mathbf{X}}) = 0].
\end{aligned} \tag{9.3}$$

Since the instances in $\bar{\mathbf{X}}$ are statistically independent we have

$$\mathbb{P}[\overline{c}(\bar{\mathbf{X}}) = 0] = (\mathbb{P}[c(X) = 0])^r = (1 - \alpha)^r.$$

From Eq. (9.1) we also have

$$\mathbb{P}[\overline{h}(\bar{\mathbf{X}}) = 0] = \mathbb{P}[\overline{c}(\bar{\mathbf{X}}) = 0] = (1 - \alpha)^r.$$

In addition,

$$\mathbb{P}[\overline{c}(\bar{\mathbf{X}}) = 0 \wedge \overline{h}(\bar{\mathbf{X}}) = 0] = (\mathbb{P}[c(X) = 0 \wedge h(X) = 0])^r = (1 - \alpha - \mathbb{P}[c(X) \neq h(X)]/2)^r.$$

The last equality above follows from Eq. (9.1) via simple calculations. Using the three equations above in Eq. (9.3), we get

$$\mathbb{P}[\overline{h}(\bar{\mathbf{X}}) \neq \overline{c}(\bar{\mathbf{X}})] = 2((1 - \alpha)^r - (1 - \alpha - \mathbb{P}[c(X) \neq h(X)]/2)^r).$$

Eq. (9.2) follows from this equality by setting $\epsilon = \mathbb{P}[c(X) \neq h(X)]$. □

To bound the true error on bags achieved by $\overline{\hat{h}}$, we invoke the VC-bound for the realizable case [Vapnik and Chervonenkis, 1971], which states that with probability at least $1 - \delta$ over a sample of m_r bags:

$$\mathbb{P}[\overline{\hat{h}}(\bar{\mathbf{X}}) \neq \overline{c}(\bar{\mathbf{X}})] \leq 2\frac{d_r}{m_r}(\log \frac{2em_r}{d_r} + \log \frac{2}{\delta}) \triangleq \text{VC-BOUND}(m_r, d_r), \tag{9.4}$$

where d_r denotes the VC-dimension of $\overline{\mathcal{H}}$, the class of hypotheses over bags of size r.

Combining Theorem 9.1 with Eq. (9.4), and taking the inverse of κ_r^α, yields the following learning bound for the proposed procedure:

Corollary 9.2. *If Eq. (9.1) holds and $c \in \mathcal{H}$, and the procedure described in Section 9.1 is used,* *then with probability $1-\delta$ over the samples of bags,*

$$\mathbb{P}[\hat{h}(X) \neq c(X)] \leq 2(1-\alpha) - 2((1-\alpha)^r - \text{VC-BOUND}(m_r, d_r)/2)^{1/r}.$$

In order to understand the effect of using bags, it will be useful to study the relationship between the bag size and the sample complexity, based on the bound in Corollary 9.2 (Note that the sample size is equal to the number of labels, which is the cost we wish to minimize). We will thus fix a target error rate, and ask how the sample complexity for this error rate changes as a function of the bag size. To this end, define $\tilde{m}_r(\epsilon)$ as the number of bags of size r required to obtain a bound of ϵ on the true error of individual examples, based on Corollary 9.2. This is an upper bound on the sample complexity when using bags of size r. In particular, $\tilde{m}_1(\epsilon)$ is the "standard" VC-bound sample complexity, when using a regular sample with individual examples. The following theorem bounds the reduction in sample complexity when bags of size r are used instead of a regular sample:

Theorem 9.3. *Let d be the VC-dimension of $\mathcal{H}$, and let d_r be the VC-dimension of the class $\overline{\mathcal{H}}$ of hypotheses over bags of size r. We have:*

$$\frac{\tilde{m}_r(\epsilon)}{\tilde{m}_1(\epsilon)} \leq \frac{\epsilon}{\kappa_r^\alpha(\epsilon)} \cdot \frac{d_r}{d}. \tag{9.5}$$

Proof. Let $m_r = \min\{\tilde{m}_1(\epsilon)\frac{\epsilon}{\kappa_r^\alpha(\epsilon)} \cdot \frac{d_r}{d}, \tilde{m}_1(\epsilon)\}$. We have

$$\begin{aligned}
\mathbb{P}[\bar{\hat{h}}(\bar{\mathbf{X}}) \neq \bar{c}(\bar{\mathbf{X}})] &\leq \text{VC-BOUND}(m_r, d_r) \\
&= \frac{d_r \tilde{m}_1(\epsilon)}{d m_r} \cdot 2\frac{d}{\tilde{m}_1(\epsilon)}(\log\frac{2em_r}{d_r} + \log\frac{2}{\delta}) \\
&\leq \frac{d_r \tilde{m}_1(\epsilon)}{d m_r}\text{VC-BOUND}(\tilde{m}_1(\epsilon), d) = \frac{d_r \tilde{m}_1(\epsilon)\cdot\epsilon}{d m_r} \leq \kappa_r^\alpha(\epsilon).
\end{aligned}$$

From Theorem 9.1 it follows that $\mathbb{P}[\hat{h}(X) \neq c(X)] \leq \epsilon$. Therefore the minimal sample size to achieve ϵ using bags of size r is no more than m_r, and Eq. (9.5) follows. □

Examining Eq. (9.5), it is obvious that $\frac{d_r}{d} \geq 1$, since the hypotheses class over bags cannot have a lower VC-dimension then the hypotheses class over individual examples. Therefore a reduction in sample complexity will only be attained if $\kappa_r^\alpha(\epsilon) > \epsilon$. That is, only if the error rate on bags is *higher* then the error rate on individual examples. This may seem counterintuitive—why would we gain from using bags if it causes an *increase* in the error rate? The key point is that we are interested in the implied error rate on individual examples, and so we can allow ourselves a higher error rate on bags, if it implies a lower error on individual examples. Note, however, that any reduction in the sample complexity due to $\kappa_r^\alpha(\epsilon) > \epsilon$ might be canceled if the VC-dimension d_r grows very fast with r. Fortunately, this is not the case, as the following theorem shows:

Theorem 9.4.

$$d_r \leq -\frac{d}{\ln(2)} \cdot W_{-1}\left(-\frac{\ln(2)}{er}\right) = O(d \log r). \tag{9.6}$$

Where W_{-1} denotes the negative branch of the Lambert W function, $x = W(x)e^{W(x)}$.

Proof. We start with the following bound from Eq. (7.1) in the proof of Theorem 7.2:

$$d_r \leq d(\log_2(erd_r/d)).$$

We reorganize this bound to find an upper bound for d_r:

$$\begin{aligned}
d_r &\leq d(\log_2(erd_r/d)) && \Rightarrow \\
d_r \ln(2) &\leq d(\ln(er/d) + \ln(d_r)) && \Rightarrow \\
d_r \ln(2) - d\ln(d_r) &\leq d\ln(er/d) && \Rightarrow \\
d\ln(d_r) - d_r \ln(2) &\geq -d\ln(er/d) && \Rightarrow \\
\ln(d_r) - \tfrac{\ln(2)}{d} d_r &\geq -\ln(er/d) && \Rightarrow \\
d_r \exp(-\tfrac{\ln(2)}{d} d_r) &\leq d/er && \Rightarrow \\
-\tfrac{\ln(2)}{d} d_r \exp(-\tfrac{\ln(2)}{d} d_r) &\leq -\ln(2)/er.
\end{aligned}$$

Since $r \geq 1$, we have that $-\frac{1}{e} \leq -\ln(2)/er \leq 0$. From the properties of the Lambert function we have that for $-\frac{1}{e} \leq x \leq 0$, $we^w \leq x \Rightarrow w \geq W_{-1}(x)$. Therefore

$$\begin{aligned}
-\tfrac{\ln(2)}{d} d_r &\geq W_{-1}(-\ln(2)/er) && \Rightarrow \\
d_r &\leq -\tfrac{d}{\ln(2)} W_{-1}(-\ln(2)/er).
\end{aligned}$$

□

Equipped with Theorems 9.3 and 9.4, we can now study the optimal bag size and the reduction in sample complexity it affords.

9.2.2 Choosing the Bag Size

We now turn to the question of how to choose a bag size r so as to minimize the sample complexity $\tilde{m}_r(\epsilon)$. The two important parameters here are the positive example rate α and the desired error guarantee ϵ. Intuitively, it can be speculated that a good size for a bag is such that the labels on bags are distributed more or less evenly, such that every label received from the teacher conveys a large amount of information to the learner. Thus r should be larger for smaller α. The bag size r should also grow as ϵ is reduced, since larger bags imply a higher sensitivity to error. The following analysis corroborates this intuition, and quantifies the dependence on both ϵ and α.

Following Theorem 9.3, we would like to choose r such that $\kappa_r^\alpha(\epsilon)/d_r$ is maximal. However, since $d \le d_r \le O(d \log r)$, i.e. d_r grows relatively slowly with r, we ignore the exact value of d_r, and define our choice for the bag size as the value of r that maximizes $\kappa_r^\alpha(\epsilon)$:

$$\begin{aligned} r^*(\alpha,\epsilon) &\triangleq \operatorname*{argmax}_r \kappa_r^\alpha(\epsilon) \\ &\equiv \operatorname*{argmax}_r \left[(1-\alpha)^r - (1-\alpha-\epsilon/2)^r\right]. \end{aligned}$$

We shall see that though this choice is not necessarily optimal, it provides a substantial reduction in sample size. Numerical calculations show that using the upper bound for d_r does not change the resulting sample size significantly.

Differentiating $\kappa_r^\alpha(\epsilon)$ we obtain a single maximum in r for all $0 < \alpha < 0.5, 0 < \epsilon < 2\alpha$:

$$r^*(\alpha,\epsilon) = \ln\left(\frac{\ln(1-\alpha-\epsilon/2)}{\ln(1-\alpha)}\right) \Big/ \ln\left(\frac{1-\alpha}{1-\alpha-\epsilon/2}\right). \tag{9.7}$$

As our preliminary intuition implied, $r^*(\alpha,\epsilon)$ is monotonic decreasing in α and in ϵ. We also speculated that the labels on bags of an optimal size should be balanced. Defining $r^*(\alpha,0) \triangleq \lim_{\epsilon\to 0^+} r^*(\alpha,\epsilon)$, we have $r^*(\alpha,0) = -1/\ln(1-\alpha) \approx 1/\alpha$. For this value of r^*, $\mathbb{P}[\bar{c}(\bar{\mathbf{X}}) = 1] = 1 - 1/e$ and the expected number of positive examples in each bag is approximately one. Figure 9.2 plots $\mathbb{P}[\bar{c}(\bar{\mathbf{X}}) = 1]$ as a function of α. The gray area between the two boundaries corresponds to different values of ϵ, in the range $(0, 2\alpha]$. This plot shows that choosing the bag size to be $r^*(\alpha,\epsilon)$ results in an almost constant probability of obtaining positive labels, confirming our intuition.

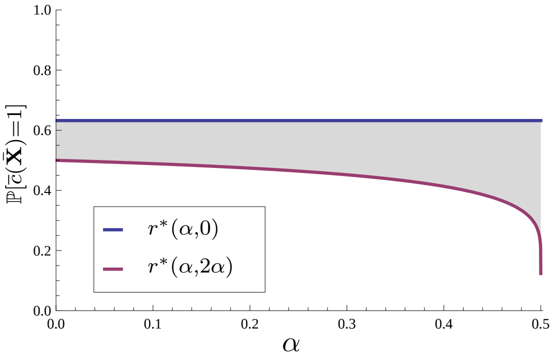

Figure 9.2: The probability for a positive bag.

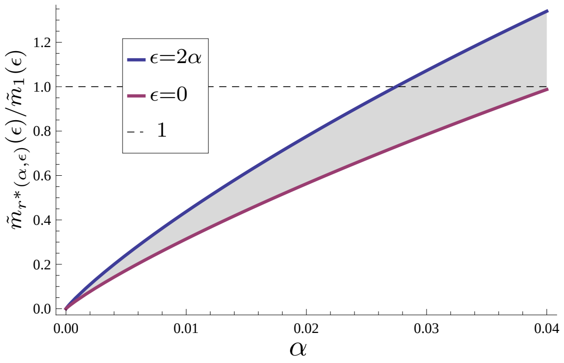

Figure 9.3: Sample size reduction factor. Anything below 1 implies a multiplicative reduction.

9.2.3 The Sample Size Reduction Factor

We can now ask whether our choice of r^* leads to a reduction in the sample size, and how large is this reduction. Substituting Eq. (9.7) and Eq. (9.6) in Eq. (9.5), yields an upper bound on $\tilde{m}_{r^*}(\epsilon)/\tilde{m}_1(\epsilon)$, the sample size reduction factor when using a bag of size r^*. For $\epsilon \to 0$ we have a simplified form:

Corollary 9.5.

$$\lim_{\epsilon \to 0+} \frac{\tilde{m}_{r^*(\alpha,\epsilon)}(\epsilon)}{\tilde{m}_1(\epsilon)} \le (1-\alpha)\ln(1-\alpha) \cdot W_{-1}\big(\ln(2)\ln(1-\alpha)/e\big) \cdot \frac{e}{\ln(2)}.$$

The bound for $\epsilon \in [0, 2\alpha]$ is plotted in Figure 9.3. Whenever the bound is smaller than 1, using bags of size r^* results in a guaranteed sample size reduction. From the figure it can be seen that this holds for $\alpha < 0.04$. This result is only a worst case bound; The experiments described in Section 9.4 show that in practice an even larger reduction is achieved, and that it is achieved for larger α as well.

9.3 Finding a Separating Hyperplane using Bags: The `PMIL` Algorithm

The analysis above provides bounds on the required sample size under the assumption that it is possible to find the hypothesis with the lowest training error on samples of bags of an arbitrary size. We now turn to show how one might find the correct hypothesis efficiently. This problem is not trivial, since it is not known which are the positive examples in a positive bag. Learning from bags with arbitrary distribution is theoretically solvable in the almost realizable case [Sabato and Tishby, 2009], however there is no algorithm that is guaranteed to work with the small sample size that our learning bounds allow. Many heuristic algorithms have also been proposed for MIL [Andrews et al., 2002, Andrews and Hofmann, 2003, Dietterich et al., 1997, Zhi-Hua Zhou, 2007, and others]. These algorithms are typically quite involved, as they must deal with samples of bags with arbitrary dependence between instances. Luckily, though the MIL problem is hard in general, our setting only employs bags with statistically independent instances, which can be expected to be a much easier problem. This case is also provably solvable [Blum and Kalai, 1998], but again only by using a large sample size.

We propose PMIL (Table 9.1), a simple iterative algorithm for finding a separating hyperplane from samples of bags, following ideas from Felzenszwalb et al. [2008]. PMIL executes the basic perceptron algorithm several times on different input samples, using parameters T and L. Though PMIL is a local-search algorithm for a non-convex objective and so might potentially find only a local minimum, it was very successful in our experiments (see Section 9.4), and has almost always found the separating hyperplane with zero or close to zero mistakes. This indicates that it is practically feasible to reduce the number of required labels using bags of independent examples. We defer the comparison of PMIL to other possible heuristics to future work.

Table 9.1: The PMIL algorithm

1. Initialize a separator w randomly;
2. Repeat until T time has passed, or until w classifies the bags with zero training error:
   (a) For each bag $\bar{\mathbf{x}}_k = (x_k^1, \ldots, x_k^r)$, select a representative example from the bag with index $i_k = \mathrm{argmax}_i(w \cdot x_k^i)$,
   (b) Run L epochs of the perceptron algorithm on the sample of individual examples $\{(x_k^{i_k}, y_k)\}_{i=1}^m$.

9.4 Experiments

In this section we present the results of experiments done on several types of learning problems. In the first batch of experiments, presented in Section 9.4.1, the procedure is tested on a finite hypothesis class, using an exhaustive search for the hypothesis with the lowest training error. This allows us to inspect the learning curves of the true $\hat{h}$, without needing to worry about the possible sub-optimality of the PMIL algorithm. Then, in Section 9.4.2, we show that the PMIL algorithm is indeed successful on both synthetic and real data sets. The experiments demonstrate a significant sample size reduction that is even better than the one promised by the analysis. They further demonstrate that using bags improves performance even when the simplifying assumption that $c \in \mathcal{H}$ does not hold. Moreover, it is shown that even using a small bag size yields a significant improvement.

9.4.1 Finite hypothesis class

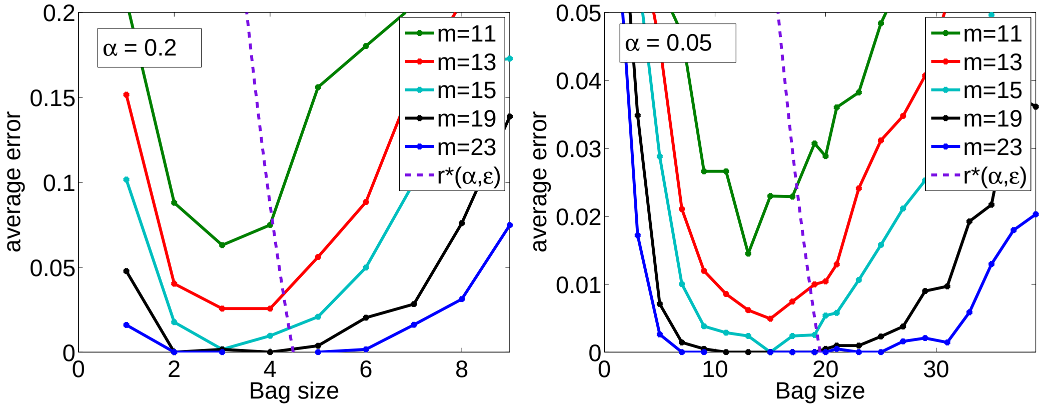

Figure 9.4: Experiments on a finite hypothesis class for two different α. Plots show the error as a function of the bag size, for several sample sizes m.

We start by examining the actual sample complexity behavior, with experiments on a finite hypothesis class, where the hypothesis with lowest training error is found using exhaustive search. We generated random examples from the domain $\mathcal{X} = \{0,1\}^{1000}$, with each of the 1000 features drawn independently with a positive example rate of α, for various values of α. The examples were labeled with a hypothesis from the class $\mathcal{H} = \{h_1, \ldots, h_{1000}\}$, where $h_i(x)$ is the value of the i'th coordinate of x. Each experiment reported was repeated either 100 or 1000 times. The plots show the average true error that was achieved.

First, we wanted to check the effect of the proposed bagging strategy on the output error on individual examples: If we fix the sample size, is there an optimal bag size $r > 1$ that achieves the lowest error? How close is the empirical optimal r to our $r^*(\alpha, \epsilon)$? Figure 9.4 shows the average true error of the learned hypothesis as a function of the bag size, for different sample sizes, and for two values of α. Even for α as large as 0.2, using bags reduces the achieved error with a fixed sample size. The dips in the plot lines indicate the existence of an optimal bag size, as predicted by the theoretical analysis. The calculated $r^*(\alpha, \epsilon)$, indicated with the dashed line, is quite close to the empirical optimum in both plots, and yields almost optimal performance.

To visualize the improvement in learning performance compared to regular supervised learning, we plotted the learning curves for selected bag sizes. The plots in Figure 9.6 compare the achieved error as a function of the sample size, for three bag sizes: one, two, and $r^*(\alpha, 0)$ (rounded). The left and middle plots show results for two values of α, with no label noise. We see a sharp improvement in performance for $r \sim r^*(\alpha, 0)$. The improvement is sharper for the smaller α. Note also, that even a bag with only two examples delivers a much better result than when using no bags. This means that a considerable improvement can be achieved even in an application that allows only small bag sizes.

One of the assumptions in our theoretical analysis was that $c \in \mathcal{H}$. We now deviate from this assumption by adding randomly flipping some of the labels creating a situation where the optimal hypothesis has error $0.017 = \alpha/3$. The right plot in Figure 9.6 shows that even when label noise is high compared to α, bagging improves the achieved error rate considerably.

Finally, we show a striking comparison between the required sample size when learning with no bags, to the required sample size when bags of optimal size are used. We have seen in the analysis that the positive example rate α is a significant parameter affecting optimal bag size and expected improvement when using bags. As α decreases, labels on single examples become less balanced. In regular learning, this means that more examples are required for effective learning. Since it is less informative to compare absolute error for varying α, Figure 9.5 examines the effect of α on the outcome *recall* (the fraction of positive examples which are identified by the output hypothesis; Note that by Eq. (9.1), the precision is also controlled). When learning without bags (dashed lines), the required sample size for a fixed recall value grows fast as α decreases. In contrast, when bags of size $r^*(\alpha, 0)$ are used (solid lines), the effect of α disappears completely. Thus, the use of bags almost eliminates the effects of unbalanced labels, by changing the bag size according to α.

9.4.2 Experiments Using PMIL

Having investigated the sample complexity effects of the use of bags, we now turn to more realistic experiments, where $\mathcal{H}$ is the set of separating hyperplanes, and PMIL is used to find a separator. In each setting we applied the procedure in Table 9.1 several times, until a separator with perfect classification on the sample of bags was found, or one second of runtime had passed. If a second had passed, we selected the separator that produced the lowest number of errors. L was set to 10.

The first set of experiments was on synthetic examples with no label noise, drawn uniformly from $\mathcal{X} = [0, 1]^{10}$. A positive label was a assigned to a fraction of size α of the cube. We performed the experiments with different sample sizes, bag sizes, and values of α. PMIL usually succeeded

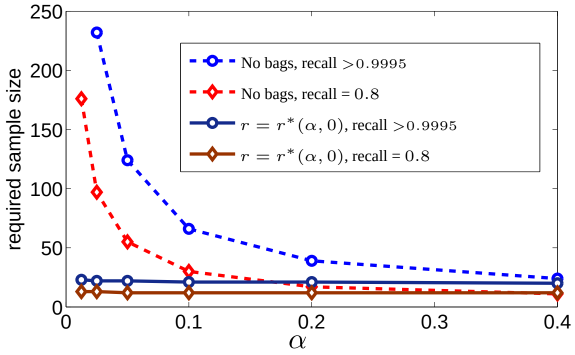

Figure 9.5: The sample size to achieve a fixed recall. Compare dashed lines (no bags) to solid lines ($r = r^*$).

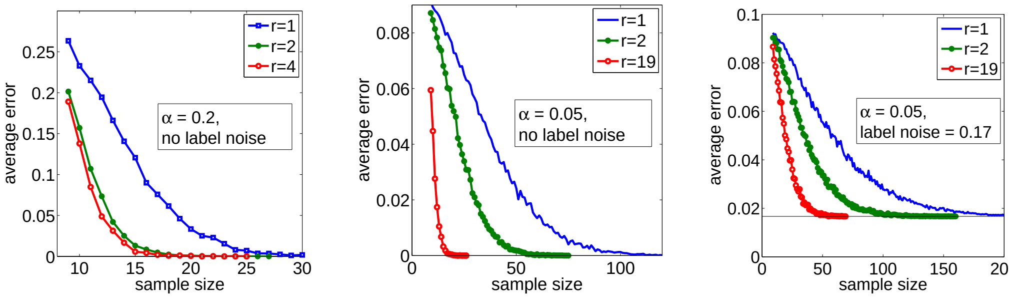

Figure 9.6: Learning curves for the finite hypothesis class, with different values of α: comparing no use of bags, bags of size 2, and bags of size $r^*(0,\alpha)$. In the right plot, some of the labels were randomly flipped.

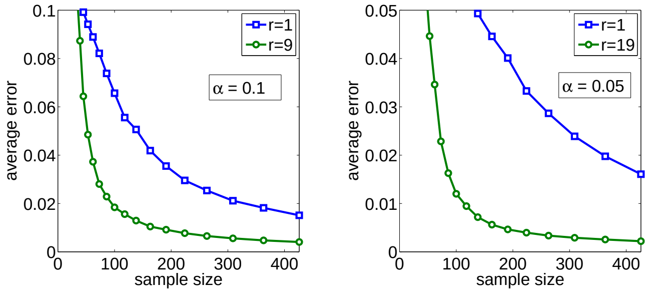

Figure 9.7: Learning synthetic data using PMIL, For two different α. The optimal bag size produces a significant improvement over $r = 1$

in achieving zero or almost zero error on the training set. Even for a bag size of 19, the algorithm usually finished with a negligible number of errors. Figure 9.7 compares the learning curves when using bags and without the use of bags for two values of α. Each dot in is the average of 1000 experiments. Here too the improvement in performance when using bags is clearly visible.

Next, we tested our learning procedure on real data sets, using samples of bags created from the original labeled examples. The first data set is the **Statlog (Shuttle) dataset** [Asuncion and Newman, 2007]. It was chosen due to the relative ease of classification using regular supervised learning, which allowed us to investigate the results of using bags in multiple experiments. To make the original multi-class problem into a binary classification problem, we selected from the training set and from the test set only examples with class 1 and 5. Class 5 was mapped to a positive label. Its occurrence in the data set is $\alpha = 0.067$, thus $r^*(\alpha, 0) \sim 14.5$. The results are plotted in Figure 9.8. On the left is the error as a function of the bag size for different sample sizes, showing that the lowest error is achieved, as expected, around $r = 14$. In the middle we compare the learning curve between learning with no bags, with bags of size 2, and with $r = 14$. Here too even a bag size of 2 provides a large improvement in the error.

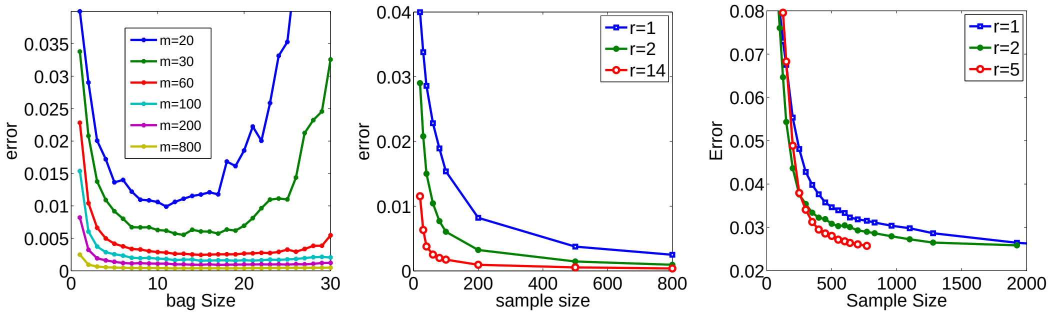

Figure 9.8: Left and Middle: Experiments on the Statlog data set ($\alpha = 0.067$). Left: the error as a function of the bag size. Each line is a sample size. Middle: Learning curves, comparing bag sizes of 1 (no bags), 2, and 14. Right: Classifying images with faces ($\alpha = 0.1$) – learning curves, comparing three bag sizes.

The second real data set we learned with PMIL was the **Caltech101 image data set** [L. Fei-Fei and Perona., 2004], exemplified in Figure 9.1. The positive class was the Faces_easy category. The negative class was all the categories except for Faces and BACKGROUND_Google, since they contain images of faces. We built a random training set of 3850 images and a random holdout set of 500 images. In both sets the we set the fraction of faces to $\alpha = 0.1$. We extracted 1000 features from the training images using k-means clustering on interest points detected as in Mikolajczyk and Schmid [2004], with default parameters. PMIL was applied to the resulting feature vectors with several bag sizes and sample sizes. Because of the default feature extraction methodology and the relatively small number of examples of faces, the best error rate that could be reached using individual examples was quite high compared to α, and only small bag sizes could be tested. Figure 9.8 (Right) compares the learning curves for $r = 1$, $r = 2$ and $r = 5$, which are lower than $r^*(\alpha, 0) \sim 9.5$. An interesting effect can be seen: When the sample size is small, it is better to use bags of a smaller size. As the sample grows, larger bags become more beneficial.

Chapter 10

Discussion (Part II)

In this part of the thesis we have provided a new theoretical analysis for Multiple Instance Learning with any underlying hypothesis class. We have shown that the dependence of the sample complexity of generalized MIL on the number of instances in a bag is only poly-logarithmic, thus implying that the statistical performance of MIL is only mildly sensitive to the size of the bag. The analysis includes binary hypotheses, real-valued hypotheses, and margin learning, all of which are used in practice in MIL applications. For classical MIL, where the bag-labeling function is the Boolean OR, and for its natural extension to $\max$, we have presented a new learning algorithm, that classifies bags by executing a learning algorithm designed for single instances. This algorithm provably PAC-learns MIL. In both the sample complexity analysis and the computational analysis, we have shown tight connections between classical supervised learning and Multiple Instance Learning, which holds regardless of the underlying hypothesis class.

Many interesting open problems remain for the generic analysis of MIL. In particular, our results hold under certain assumptions on the bag functions. An interesting open question is whether these assumptions are necessary, or whether useful results can be achieved for other classes of bag functions. Another interesting question is how additional structure within a bag, such as sparsity, may affect the statistical and computational feasibility of MIL. These interesting problems are left for future research.

We further studied a novel paradigm for learning from a labeled sample using a teacher that can provide OR-labels, when the cost of obtaining labels from the teacher is high, while the cost of presenting examples to the teacher is negligible. We demonstrated that a significant improvement in the error can be achieved with a fixed amount of labels, by presenting to the teacher bags of examples instead of individual examples. We have shown that the size of the bag that should be used has an optimum and that an almost optimal bag size can be analytically found. The `PMIL` algorithm was proposed for finding a separating hyperplane with low training error from a sample of bags. Experiments on various types of data sets demonstrate that the proposed method and learning algorithm work well in practice, and that the method can be used even if the exact problem parameters are not known.

Bibliography

N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. In *Foundations of Computer Science, 1993. Proceedings., 34th Annual Symposium on*, pages 292–301. IEEE, 1993.

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

S. Andrews and T. Hofmann. Multiple-instance learning via disjunctive programming boosting. In *NIPS*, 2003.

S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, pages 561–568, 2002.

S. J. D. Andrews. *Learning from ambiguous examples*. PhD thesis, Brown University, May 2007.

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

A. Antos and G. Lugosi. Strong minimax lower bounds for learning. *Mach. Learn.*, 30(1):31–56, 1998. ISSN 0885-6125.

A. Asuncion and D. Newman. UCI machine learning repository, 2007.

P. Auer, P. M. Long, and A. Srinivasan. Approximating hyper-rectangles: learning and pseudorandom sets. *J. Comput. Syst. Sci.*, 57(3):376–388, 1998.

B. Babenko, N. Verma, P. Dollar, and S. Belongie. Multiple instance learning with manifold bags. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML '11, pages 81–88, 2011.

Z. Bai and J. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer, second edition edition, 2010.

P. Bartlett. Lecture notes. http://www.cs.berkeley.edu/~bartlett/courses/281b-sp06/lecture25.ps, 2006. unpublished.

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

P. L. Bartlett, S. R. Kulkarni, and S. E. Posner. Covering numbers for real-valued function classes. *IEEE Transactions on Information Theory*, 43(5):1721–1724, 1997.

S. Ben-David, T. Lu, and D. Pál. Does unlabeled data provably help? In *Proceedings of the Twenty-First Annual Conference on Computational Learning Theory*, pages 33–44, 2008.

G. M. Benedek and A. Itai. Learnability with respect to fixed distributions. *Theoretical Computer Science*, 86(2):377–389, Sept. 1991.

G. Bennett, V. Goodman, and C. M. Newman. Norms of random matrices. *Pacific J. Math.*, 59(2): 359–365, 1975.

A. Blum and A. Kalai. A note on learning from multiple-instance examples. *Mach. Learn.*, 30(1): 23–29, 1998.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pages 144–152, 1992.

O. Bousquet. *Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms*. PhD thesis, Ecole Polytechnique, 2002.

V. Buldygin and Y. V. Kozachenko. *Metric Characterization of Random Variables and Random Processes*. American Mathematical Society, 1998.

C. Caramanis and S. Mannor. An inequality for nearly log-concave distributions with applications to learning. *Information Theory, IEEE Transactions on*, 53(3):1043–1057, 2007.

S. Chari, P. Rohatgi, and A. Srinivasan. Improved algorithms via approximations of probability distributions. In *Proceedings of the Twenty-Sixth Annual ACM Symposium on the Theory of Computing*, pages 584–592, 1994.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995.

A. Daniely, S. Sabato, S. Ben-David, and S. Shalev-Shwartz. Multiclass learnability and the ERM principle. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, pages 19:207–232. JMLR Workshop and Conference Proceedings, 2011.

L. Devroye and G. Lugosi. Lower bounds in pattern recognition and learning. *Pattern recognition*, 28(7):1011–1018, 1995.

T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.*, 89(1-2):31–71, 1997.

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

R. M. Dudley. Central limit theorems for empirical measures. *The Annals of Probability*, 6(6): 899–929, 1978.

A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant. A general lower bound on the number of examples needed for learning. In *Proceedings of the First Anuual Workshop on Computational Learning Theory*, pages 139–154, Aug. 1988.

P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1–8, 2008.

Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, Aug. 1997.

T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola. Multi-instance kernels. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 179–186, 2002.

C. Gentile and D. Helmbold. Improved lower bounds for learning from noisy examples: an information-theoretic approach. In *COLT*, pages 104–115, 1998.

A. Gonen, S. Sabato, and S. Shalev-Shwartz. Active learning halfspaces under margin assumptions. *CoRR*, abs/1112.1556, 2011.

L. Gurvits. A note on a scale-sensitive dimension of linear bounded functionals in banach spaces. In *Algorithmic Learning Theory*, pages 352–363. Springer, 1997.

D. Haussler and P. M. Long. A generalization of Sauer's lemma. *Journal of Combinatorial Theory, Series A*, 71(2):219–240, 1995.

K. U. Höffgen, K. S. V. Horn, and H. U. Simon. Robust trainability of single neurons. *Journal of Computer and System Sciences*, 50(1):114–125, 1995.

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.

R. F. L. Fei-Fei and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *CVPR 2004, Workshop on Generative-Model Based Vision*. IEEE, 2004.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer, 1991.

P. Liang and N. Srebro. On the interaction between norm and dimensionality: Multiple regimes in learning. In *ICML*, 2010.

P. M. Long and L. Tan. Pac learning axis-aligned rectangles with respect to product distributions from multiple-instance examples. *Machine Learning*, 30(1):7–21, 1998. ISSN 0885-6125.

O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. In *NIPS '97: Proceedings of the 1997 conference on Advances in neural information processing systems 10*, pages 570–576, Cambridge, MA, USA, 1998. MIT Press.

O. Maron and A. L. Ratan. Multiple-instance learning for natural scene classification. In *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning*, pages 341–349, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

S. Mendelson. Rademacher averages and phase transitions in glivenko-cantelli classes. *IEEE Transactions on Information Theory*, 48(1):251–263, 2002.

K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *Int. J. Comput. Vision*, 60(1):63–86, 2004.

S. Nash and A. Sofer. *Linear and Nonlinear Programming*. McGraw-Hill, New York, 1996.

F. Nazarov and A. Podkorytov. Ball, haagerup, and distribution functions. *Operator Theory: Advances and Applications*, 113 (Complex analysis, operators, and related topics):247–267, 2000.

A. Ng. Feature selection, l_1 vs. l_2 regularization, and rotational invariance. In *ICML*, 2004.

R. Paley and A. Zygmund. A note on analytic functions in the unit circle. *Proceedings of the Cambridge Philosophical Society*, 28:266272, 1932.

L. Pitt and L. G. Valiant. Computational limitations on learning from examples. Technical report, Harvard University Aiken Computation Laboratory, July 1986.

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

L. D. Raedt. Attribute-value learning versus inductive logic programming: The missing links (extended abstract). In *ILP '98: Proceedings of the 8th International Workshop on Inductive Logic Programming*, pages 1–8, London, UK, 1998. Springer-Verlag. ISBN 3-540-64738-4.

G. Rätsch and M. K. Warmuth. Efficient margin maximizing with boosting. *Journal of Machine Learning Research*, 6:2131–2152, Dec. 2005.

F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386–407, 1958. (Reprinted in *Neurocomputing* (MIT Press, 1988).).

M. Rudelson and R. Vershynin. The littlewoodofford problem and invertibility of random matrices. *Advances in Mathematics*, 218(2):600–633, 2008.

M. Rudelson and R. Vershynin. The smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics*, 62:1707–1739, 2009.

S. Sabato and N. Tishby. Homogeneous multi-instance learning with arbitrary dependence. In *COLT*, 2009.

S. Sabato, N. Srebro, and N. Tishby. Reducing label complexity by learning from bags. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9, pages 685–692, 2010a.

S. Sabato, N. Srebro, and N. Tishby. Tight sample complexity of large-margin learning. In *Advances in Neural Information Processing Systems 23 (NIPS)*, pages 2038–2046, 2010b.

N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory Series A*, 13:145–147, 1972.

R. Schapire, Y. Freund, P. Bartlett, and W. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, October 1998.

R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):1–40, 1999.

B. Schölkopf, J. Shawe-Taylor, A. J. Smola, and R. Williamson. Generalization bounds via eigenvalues of the gram matrix. Technical Report NC2-TR-1999-035, NeuroCOLT2, 1999.

S. Shalev-Shwartz, O. Shamir, and K. Sridharan. Learning kernel-based halfspaces with the zero-one loss. In *Proceedings of the Twenty-Third Annual Conference on Computational Learning Theory*, 2010.

O. Shamir, S. Sabato, and N. Tishby. Learning and generalization with the information bottleneck. In *Proceedings of the 19th international conference on Algorithmic Learning Theory (ALT)*, pages 92–107, 2008.

O. Shamir, S. Sabato, and N. Tishby. Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411:2696–2711, June 2010.

N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low-noise and fast rates. *CoRR*, abs/1009.3896, 2010.

I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. *Annals of Statistics*, 35(2):575–607, 2007.

V. N. Sudakov. Gaussian processes and measures of solid angles in hilbert space. *Sov. Math.Dokl.*, 12:412–415, 1971.

N. Tomczak-Jaegermann. *Banach-Mazur Distances and Finite-Dimensional Operator Ideals*, volume 38 of *Pitman Monographs and Surveys in Pure and Applied Mathematics*. Pitman, 1989.

L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, Nov. 1984.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

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

V. N. Vapnik and A. Y. Chervonenkis. *Theory of pattern recognition*. Nauka, Moscow, 1974. (In Russian).

N. Vayatis and R. Azencott. Distribution-dependent vapnik-chervonenkis bounds. In *EuroCOLT '99*, pages 230–240, London, UK, 1999. Springer-Verlag. ISBN 3-540-65701-0.

J. von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100:295–320, 1928.

N. Weidmann, E. Frank, and B. Pfahringer. A two-level learning method for generalized multi-instance problems, 2003.

D. Wolpert and W. Macready. No free lunch theorems for optimization. *Evolutionary Computation, IEEE Transactions on*, 1(1):67–82, 1997.

Q. Zhang and S. Goldman. Em-dd: An improved multiple-instance learning technique. In *Advances in Neural Information Processing Systems 14*, 2001.

T. Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2:527–550, 2002.

M.-L. Z. Zhi-Hua Zhou. Solving multi-instance problems with classifier ensemble based on constructive clustering. *Knowledge and Information Systems*, 11(2):155–170, 2007.

Z.-H. Zhou, K. Jiang, and M. Li. Multi-instance learning based web mining. *Applied Intelligence*, 22(2):135–147, 2005. ISSN 0924-669X.