A Framework for Estimating Risk

Rodney Stephen Kroon

Dissertation presented for the degree of
Doctor of Philosophy
at Stellenbosch University

Promoter: Prof. S. J. Steel
April 2008

Declaration

By submitting this dissertation electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the owner of the copyright thereof (unless to the extent explicitly stated otherwise) and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: February 19, 2008

Opsomming

Ons beskou evaluering van gepaste modelle deur middel van risikoberaming. Verskeie benaderings tot risikoberaming word in ’n verenigde raamwerk oorweeg. Hierdie raamwerk is ’n uitbreiding van ’n beslissingsteoretiese raamwerk oorspronklik deur David Haussler voorgestel. Punt- en intervalberaming gebaseer op toets- en evalueringssteekproewe word bespreek, met ’n onderverdeling van intervalberamers op grond van die afwykingsmaatstaf wat die beramer probeer begrens.

Die hoofbydrae van die tesis is in die gebied van evalueringssteekproef intervalberamers, spesifiek bedekkingsgetal-gebaseerde en PAC-Bayesiaanse intervalberamers. Die tesis bespreek ’n aantal benaderings tot die verkryging van sulke beramers. Die eerste tipe evalueringssteekproef intervalberamer om aandag te ontvang is beramers gebaseer op klassieke bedekkingsgetal argumente. ’n Aantal sulke beramers is op verskeie maniere veralgemeen. Tipiese veralgemenings het die volgende ingesluit: uitbreiding van resultate vir misklassifikasieverlies na algemene verliesfunksies; uitbreiding van resultate om ’n arbitrêre spooksteekproefgrootte toe te laat; uitbreiding van resultate om arbitrêre resolusie van die betrokke bedekkingsgetalle toe te laat; en uitbreiding van resultate om arbitrêre keuse van β in die gebruik van simmetriseringslemmas toe te laat.

Hierdie uitbreidings is op bedekkingsgetal-gebaseerde beramers vir verskeie afwykingsmaatstawwe toegepas, asook vir die spesiale gevalle van misklassifikasieverlies beramers, beramers vir die haalbare geval, en spelingsgrense. Uitgebreide resultate is ook vir die geval van besluitklasse gestratifiseer op grond van (algoritme- en data-afhanklike) funksiekompleksiteit.

Om toepassing van hierdie bedekkingsgetal-gebaseerde intervalberamers aan te help, is ’n oorsig van verskeie kompleksiteitsdimensies en benaderings tot die verkryging van bogrense op bedekkingsgetalle aangebied.

Die tweede tipe evalueringssteekproef intervalberamer wat in die tesis bespreek word is Rademacher-bogrense. Hierdie resultate gebruik gevorderde ophopingsongelykhede, wat ons in ’n aparte hoofstuk bespreek. Ons bespreking van Rademacher-bogrense lei tot die aanbieding van ’n alternatiewe, effens sterker, vorm van die kernresultaat wat gebruik word om plaaslike Rademacher-bogrense af te lei, deur ’n paar onnodige verslappings in die afleiding te omseil.

Daarna begin ons met ’n bespreking van PAC-Bayesiaanse bogrense. Ons gebruik ’n metode ontwikkel deur Olivier Catoni om nuwe PAC-Bayesianse bogrense gebaseer op Hoeffding se ongelykeheid af te lei. Deur Catoni se idee van “verwisselbare priors” te gebruik, kon ons hierdie resultate verder veralgemeen om ’n uitbreiding van ’n bedekkingsgetal-gebaseerde resultaat te verkry wat ook op middelmaatklassifikasietegnieke toegepas kan word. Verder kon die ooreenstemmende algoritme- en data-afhanklike resultate soortgelyk uitgebrei word.

Die laaste bydrae van hierdie tesis is die ontwikkeling van ’n meer buigsame peulontbindingsgrens: deur Hoeffding se stertongelykheid in plaas van Hoeffding se relatiewe entropie ongelykheid te gebruik het ons die grens uitgebrei om op algemene verliesfunksies van toepassing te wees, om die gebruik van ’n arbitrêre hoeveelheid busse toe te laat, en deur tussen-bus en binnebus “priors” in te voer.

Laastens, om die berekening van hierdie intervalberamers te illustreer, het ons van hierdie grense bereken vir beslissingsbome en opgejaagde stompe toegepas op die UCI gemorspos klassifikasie probleem.

Abstract

We consider the problem of model assessment by risk estimation. Various approaches to risk estimation are considered in a unified framework. This framework is an extension of a decision-theoretic framework proposed by David Haussler. Point and interval estimation based on test samples and training samples is discussed, with interval estimators being classified based on the measure of deviation they attempt to bound.

The main contribution of this thesis is in the realm of training sample interval estimators, particularly covering number-based and PAC-Bayesian interval estimators. The thesis discusses a number of approaches to obtaining such estimators. The first type of training sample interval estimator to receive attention is estimators based on classical covering number arguments. A number of these estimators were generalized in various directions. Typical generalizations included: extension of results from misclassification loss to other loss functions; extending results to allow arbitrary ghost sample size; extending results to allow arbitrary scale in the relevant covering numbers; and extending results to allow arbitrary choice of β in the use of symmetrization lemmas.

These extensions were applied to covering number-based estimators for various measures of deviation, as well as for the special cases of misclassification loss estimators, realizable case estimators, and margin bounds. Extended results were also provided for stratification by (algorithm- and data-dependent) complexity of the decision class.

In order to facilitate application of these covering number-based bounds,

a discussion of various complexity dimensions and approaches to obtaining bounds on covering numbers is also presented.

The second type of training sample interval estimator discussed in the thesis is Rademacher bounds. These bounds use advanced concentration inequalities, so a chapter discussing such inequalities is provided. Our discussion of Rademacher bounds leads to the presentation of an alternative, slightly stronger, form of the core result used for deriving local Rademacher bounds, by avoiding a few unnecessary relaxations.

Next, we turn to a discussion of PAC-Bayesian bounds. Using an approach developed by Olivier Catoni, we develop new PAC-Bayesian bounds based on results underlying Hoeffding's inequality. By utilizing Catoni's concept of "exchangeable priors", these results allowed the extension of a covering number-based result to averaging classifiers, as well as its corresponding algorithm- and data-dependent result.

The last contribution of the thesis is the development of a more flexible shell decomposition bound: by using Hoeffding's tail inequality rather than Hoeffding's relative entropy inequality, we extended the bound to general loss functions, allowed the use of an arbitrary number of bins, and introduced between-bin and within-bin "priors".

Finally, to illustrate the calculation of these bounds, we applied some of them to the UCI spam classification problem, using decision trees and boosted stumps.

Acknowledgements

Most notably, I would like to thank my wife, Dalene de Beer, for her support, encouragement and faith in me. Without you, I would probably never have stuck to my guns and finished this thing off. Thanks for the sacrifices you made to encourage my progress, and for showing just the right mix of regret and resignation at my long hours away from home. Although words can never say thank you enough, I dedicate all the words in this thesis to you.

I'd also like to thank a number of my friends. My cell group (including the members that have left town during my studies) has provided consistent fellowship, and I'd like to thank them for their compassion, prayers and motivating words. I'd like to single out Andries Kruger, whose companionship in the last month helped me come out of this sane. I was also fortunate to meet Hugo van der Merwe during my doctoral studies. Hugo is somewhat of a kindred spirit, and our developing friendship featured an inspirational mix of philosophical, religious, scientific and political discussions. I'd also like to thank Florian Breuer, who was kind enough to loan me his mathematical expertise to explore unfamiliar terrain when necessary.

Thank you to my family for their patience while my student career stretched interminably. You can now tell the people who ask what I am doing that I have finished studying and started working — neither of which is true. I'm also grateful to you all, especially my mother-in-law, Ria de Beer, for your encouraging words.

My promoter, Sarel Steel, was of course instrumental in the completion of this thesis. His guidance and feedback were invaluable, and my productivity

was roughly proportional to the frequency of our meetings.

In addition to Sarel, I would like to thank my other colleagues at the Department of Statistics for their friendliness and helpfulness, particularly Surette Oosthuizen, Morne Lamont, Pieta van Deventer, and Brigott Bruintjies.

The Department of Computer Science was extremely kind to provide me with an office during the second half of my studies. I'd like to thank my future colleagues there for welcoming me to the fold, and for the many fascinating tea-time discussions. Here I'd like to single out Deon Borman, the technical assistant, who patiently helped me with a variety of problems, and Jaco Geldenhuys, for his help with TeX.

I am also grateful to Ulrich Paquet for pointers to useful resources and the long-term loan of a textbook; to Andrew Inggs for the use of his LaTeXthesis template; to the maintainers of the UCI machine learning repository, whose behind-the-scenes work is not mentioned often enough; and to the legion of contributors who make Wikipedia such an excellent resource.

Finally, a number of sources provided me with financial support. Stellenbosch University awarded me the Harry Crossley bursary, the HB Thom bursary, and a merit bursary, and I would like to thank the organizations who provide funding for these bursaries. In addition, I would like to thank the university's Department of Statistics, the Ernst and Ethel Erikson Trust and the Skye Foundation for financial assistance.

“We are all pencils in the hand of God.”
— Mother Teresa of Calcutta

Contents

List of Tables

Chapter 1

Introduction

1.1 Motivation

The practice of fitting models to data is ubiquitous in modern science, engineering and business. Predictions made using such fitted models are often employed for automated decision-making, or as a source of information for higher level manual decision making. As a result, the problem of assessing the quality of fitted models is vitally important.

The perceived quality of a fitted model is dependent on the purpose for which it is employed. This is particularly understandable in certain business contexts, where one may be able to calculate an exact financial cost incurred by a suboptimal prediction. Perhaps the most natural approach in this context is to attempt to select a fitted model which almost minimizes the expected cost of future predictions. This expected cost, which we shall call the *risk* of the fitted model, seems a good indicator of the quality of the fitted model.

In this thesis, we investigate various approaches to estimating the risk of a fitted model.

Traditionally, the risk of a fitted model is estimated by means of a test sample, or holdout sample: a data set which is representative of expected future data points, and which is independent of the fitted model. Since risk assessment is so important, it is standard practice to remove such a test sample from an initial data sample, fit a model on the remainder of the data, and then assess the model on the test sample. A number of good, well-established estimators of the fitted model's risk are available in this scenario.

A criticism of this approach is that one would generally expect a model fitted on the full data sample to perform better than the one fitted on the reduced data sample. If one could effectively assess the model fitted on the entire data sample without reserving a portion for testing, better models could be employed in practice.

Methods for assessing models in this way do exist. However, mere existence of such techniques is not adequate. The quality of these estimators needs to be competitive with those obtained using the test sample, if such an approach to model fitting and assessment is to be widely adopted.

However, even in cases where such estimators are not competitive with traditional estimators based on a test sample, these estimators are practically useful: such approaches have been employed for showing consistency of various model-fitting procedures, model selection, and designing new model-fitting procedures. Recent model-fitting procedures inspired by such estimators feature among the most successful model-fitting procedures available.

1.2 Problem statement

The statistical problem we face is that of estimation: we desire an estimate for the risk of a fitted model. More generally, we may be interested in the expected risk of a model fitting procedure (or algorithm) in a given context.

The main problem we shall consider can be stated as:

Generate an estimator of the risk $r_D(w, L)$ *of a fitted model (or* decision rule*)* w*, given that* w *was determined by employing an algorithm* Θ *on a sample* S*.*

Note that this problem statement includes estimating the risk of regression models and classifiers.

In the statement above, L is an encoding of the cost of suboptimal predictions, known as a *loss function*, and D denotes the distribution of future data points. Furthermore, we are primarily interested in estimators which are functions of S — what we shall call training sample estimators.

The simplest form of estimation is point estimation. However, point estimators do not reflect the level of confidence one has in an estimate, whether from its desirable statistical properties, or the reliability obtained from using a large sample to obtain the estimate. For this reason, an estimate of the variance of the estimator is often provided together with an estimator, to give an indication of the variability of the estimator. This naturally leads one to consider interval estimation: providing an interval which typically[1] contains the value being estimated. Such interval estimators are generally called *confidence intervals*.

Construction of training sample interval estimators is a difficult problem, and an extensive literature exists, with further focus on the simpler case of interval estimation of the misclassification rate (or *error*) of the model. Much of our work is also focused on this case, although more general loss functions will not be neglected.

Generally, we are more interested in cases where the risk is unusually large, rather than cases where it is unusually small. As a result, we will often be interested in one-sided confidence intervals for the risk: bounding the risk from above with a certain level of confidence.

1.3 Objectives

The first objective of this thesis is to provide an overview of various approaches to risk estimation within a common framework. This involves the

[1] By typically, we mean that the probability of obtaining a sample for which the interval contains the underlying value exceeds some prescribed value.

development of a single framework in which various approaches to risk estimation can be described, and the presentation of a number of tools which are used for developing these estimators.

An auxiliary objective is to attempt to make training sample interval estimators more accessible to the newcomer by presenting as much of the relevant background as is practical, and attempting to keep the learning curve shallow. Until recently, most of the research in this direction has been done by the theoretical computer science and machine learning communities. My aim in this regard is to make the material more statistician-friendly by presenting the material from a different viewpoint.

Other objectives of the thesis require a little historical background. The foundation of training sample interval estimators was laid with the development of *statistical learning theory* by Vladimir Vapnik and Alexey Chervonenkis from the late 1960s. Their work focused on obtaining analogues of the laws of large numbers which hold uniformly over an infinite function class. Such results only hold under certain conditions, and the theorems stating these conditions can be restated to obtain training sample interval estimators.

The key realization here is that early workers were interested in (various modes of) convergence of sequences of empirical quantities to corresponding quantities of an underlying distribution. Later work investigated the asymptotic rate of convergence of these sequences, but at no point were the precise value of constant factors considered important. When precise constants were presented, they were not generally as tight as possible. While this had no impact on their investigations, the values of the constants are relevant for obtaining training sample interval estimators.

A related issue is that bounds were derived for their asymptotic form, rather than for finite sample purposes. As a result, many variable parameters were set at values convenient for the derivations under consideration, but which may not be near optimal for finite sample considerations.

The third objective of this thesis is therefore to present generalized forms of such results, where the variable parameters can be specified by the practitioner. In addition, we will pay much more attention to values of constants than is typically done in much of the classical literature.

Our fourth objective will be to evaluate the impact of these extra parameters, and the focus on constants, on the bounds which can be achieved. Many classical training sample interval estimators are traditionally summarily dismissed in practice, since they are thought to invariably yield trivial bounds on risk. We will investigate if our generalizations help to address the situation.

Our fifth objective is to compare the performance of training sample interval estimators to interval estimators based on an independent test sample.

Hopefully, these investigations will make it clear whether training sample bounds are competitive enough for practical use yet.

1.4 Thesis outline

Regarding the scope of the thesis, we consider only the case where samples contain independent, identically distributed observations from the same distribution generating future data points. Furthermore, we restrict ourselves to the case of bounded loss.

Chapter 2 introduces the concepts of risk estimation, and presents a framework for considering risk estimation problems. This framework is a generalization of a framework presented in Haussler (1992), the main modification being the introduction of a *strategy*, which allows one to deal with stochastic decision rules and thresholding classifiers. We show that this framework encompasses traditional results on risk estimation by employing a type of projection argument.

Chapter 3 considers various approaches to risk estimation using a test sample. We focus on the case of misclassification rate, where the number of misclassified points on the test sample has a binomial distribution. We present a view of interval estimators in terms of various *measures of deviation*: an interval estimator is obtained by inverting a bound on some measure of deviation. We discuss various criteria for evaluating interval estimators, and consider various test sample interval estimators.

Before turning to training sample estimators, we introduce inequalities based on the concept of concentration of measure in Chapter 4. The results in the first half of this chapter enable us to obtain test sample interval estimators for other loss functions. The second half of the Chapter provides more sophisticated machinery which we will use for some of the more refined training sample interval estimators.

Chapter 5 is the longest chapter in the thesis. It presents a few approaches to training sample point estimation by employing the bootstrap and the jackknife, before turning to the problem of training sample interval estimation. The Occam's razor method for countable function classes is presented, followed by the idea of approximating a function class by a suitable cover. Combining these two concepts allows one to obtain interval estimators in terms of the size of a cover of the class. We present and generalize a number of such estimators based on bounding various measures of deviation. The chapter also considers margin bounds and bounds based on the (generic) chaining method from empirical process theory[2]. Finally, we consider various approaches to obtaining bounds on the covering numbers employed in the estimators presented in this chapter.

The bounds presented in Chapter 5 are data-independent in the sense that the bounds obtained on the relevant measure of deviation do not depend on the training sample employed. Chapter 6 presents data-dependent bounds, which allow one to take advantage of a "lucky" training sample.

Chapter 7 explores the use of the advanced concentration inequalities presented in Chapter 4 to obtain training sample interval estimators. This approach allows us to replace the mean covering numbers employed in earlier chapters by the realized covering number on the training sample under

[2] Bounding the regular measure of deviation uniformly over a function class can be viewed as bounding the supremum of an empirical process.

consideration. The main focus of the chapter, however, is on Rademacher bounds, which are based on a symmetrization lemma from empirical processes.

The PAC-Bayesian approach to obtaining training sample interval estimators, which began in the late 1990s, is the focus of Chapter 8. This approach provides interval estimators for decision rules employing the Gibbs strategy. Extensions of this approach to obtain margin bounds and data- and algorithm-dependent results are also presented. Shell decomposition bounds, which are based on a similar style of argument, are presented next, before the chapter is concluded with an overview of Occam's hammer, a recently discovered approach due to Gilles Blanchard.

Chapter 9 applies a number of these estimators to risk estimation on a benchmark data set for spam classification and the results of the various approaches are discussed and compared. We review and summarize our findings and contributions in Chapter 10, and suggest a number of avenues for further investigation.

1.5 Technical issues and notation

It would be fair to say that the general attitude toward precision and exceptional cases in the field of statistical learning theory (the basis of training sample interval estimators) could in the past be summarized by the following quote of Hector Hugh Munro, the British writer better known as Saki:

> *A little inaccuracy sometimes saves tons of explanation.*

This observation holds true on two major levels.

First, many of the foundational results are measure theoretic in nature. In order to apply them, a number of technical restrictions on various function classes are necessary in order to ensure measurability of certain sets and functions. Once it was discovered that these restrictions generally hold on the function classes usually considered in practice, it became the norm to simply note that measurability issues would be ignored, and to refer those with a taste for detail to the work of Richard Dudley (e.g. Dudley, 1978) and David Pollard (Pollard, 1984) for conditions under which they would hold. This pragmatic point of view is exemplified by the following excerpt from Talagrand (1995):

> *…measurability questions are well understood, and are irrelevant in the study of inequalities. Since it would be distracting to spend time and energy on routine considerations, we have felt that it would be better to simply ignore all measurability questions, and treat all sets and functions as if they were measurable. This is certainly the case if one should assume that Ω is Polish, μ is a Borel measure, and that one studies only compact sets, which is the only situation that occurs in applications.*

As noted by Talagrand in the same article, results which hold in the measurable case can often be extended to cases where measurability does not hold by replacing integrals/probabilities with outer integrals/probabilities, although proofs are complicated by the lack of an equivalent of Fubini's theorem for outer integrals. We shall also avoid measurability questions as a general rule, but note that this requirement of only studying compact sets is what motivates the requirement of bounded loss functions in our work.

The second level on which this attitude manifests is a disregard for precise values of constants. Many foundational results in the field were convergence theorems, where constants were of no consequence. Later work investigated the asymptotic rate of convergence, and constants were still of no consequence. The following excerpt from Alexander (1984) clarifies the view at the time:

> *We have not attempted to obtain best numerical constants in the above and following results; techniques which depend on the metric entropy*[3]*, which is usually known only up to an asymptotic rate, do not lend themselves to this. Our results are intended for asymptotic use.*

[3] The metric entropy is the natural logarithm of the covering number.

As a result of this view, many results are presented with very large or unspecified constants. A number of these results in turn form the foundation of modern results, which are still provided with poor or unspecified constants.

However, if one is interested in training sample interval estimators, good constants become valuable. Devroye et al. (1996) present an example where a result from Alexander (1984) is outperformed by an asymptotically weaker result in Devroye (1982) for all sample sizes less than 2^{6144}. The question which is unanswered by this approach is what sample size would be necessary if both results employed optimal constants. Such questions are highly relevant when one makes the transition from asymptotic results to finite samples. In this thesis, we investigate the question of obtaining practical training sample interval estimators.[4] As a result, the values of constants are treated as important. This necessarily means that results which are asymptotically attractive could not be employed in this thesis, unless the underlying constants could be extracted from the proof of the result.

A number of other details are glossed over in this thesis. When an operation is performed on elements of a class, we assume that the appropriate operation is defined on the function class. Similarly, when we work with the density of a measure, we implicitly assume the measure is absolutely continuous w.r.t. an appropriate measure.

1.5.1 Notation

Probability is denoted by $\mathbb{P}$, with a subscript typically indicating the variable and distribution under consideration. Similarly, $\mathbb{E}$ denotes expectation, and $\mathbb{V}$ denotes variance. The mode and median of a random variable E are denoted by $\text{mode}(E)$ and $M(E)$ respectively.

$\text{Ent}(Q)$ denotes the entropy of Q. $\text{KL}(Q_1||Q_2)$, where Q_1 and Q_2 are distributions, denotes the Kullback-Leibler divergence (relative entropy) of

[4] Alexander's quote seems fatalistic with regards to bounds based on covering numbers. It seems to posit that training sample interval estimators based on covering number approaches will never be practically useful. In a sense, then, portions of this thesis can be seen as an investigation of the validity of this claim.

Q_2 from Q_1; $\mathrm{KL}(v_1||v_2)$, where $v_1, v_2 \in [0,1]$, is used as shorthand for $\mathrm{KL}\left(\mathrm{Bin}(1,v_1)||\,\mathrm{Bin}(1,v_2)\right)$, the divergence of a Bernoulli distribution with parameter v_2 from one with parameter v_1.

The uniform distribution on a set A is denoted by $\mathrm{Unif}\,A$, $\mathrm{Bin}(k,p)$ denotes the binomial distribution with parameters k and p, $N(\mu,\sigma^2)$ denotes the normal distribution, and χ_i^2 denotes the chi-square distribution with i degrees of freedom.

$\mathbb{R}$, $\mathbb{N}$ are used for the real and natural numbers, $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ for the counting numbers, and $\mathbb{Z}$ for the integers. ln denotes the natural logarithm, logarithms with base b are denoted by $\log_b$. v_+ and v_- denote the positive and negative parts of v respectively. I is used for the indicator function of a set or predicate. supp denotes the support of a function or a distribution, while domain and range denote the domain and range of a function respectively.

sgn denotes the sign function mapping into $\{-1,1\}$, with $\mathrm{sgn}(0) = -1$. erf denotes the Gauss error function,

$$\mathrm{erf}(v) = \frac{2}{\pi} \int_0^v e^{-t^2}\, dt \ .$$

Lik denotes a likelihood function.

We use $\mathrm{conv}\,A$ to denote the convex hull of a set A, and $\mathrm{absconv}\,A$ to denote the absolute (symmetric) convex hull of A, defined by $\mathrm{absconv}\,A = \mathrm{conv}(A \cup -A)$.

$\sup_{v \in A} \phi(v)$ is used as shorthand for $\sup\{\phi(v) : v \in A\}$, and the infimum is handled similarly. The expression $v \in A$ may be replaced by another condition defining a set of suitable v. When A is clear from the context, it may be omitted.

We denote the cardinality of a set A by $|A|$. The power set of A is written as 2^A, and the set of functions from A_1 to A_2 is written as ${A_2}^{A_1}$. v^T denotes the transpose of a matrix v. $\langle v_1, v_2 \rangle$ denotes the inner product of v_1 and v_2. For a multi-dimensional quantity or sequence v, $v^{(i)}$ denotes the i-th component or coordinate of v. For multi-dimensional quantities or sequences v_1 and

v_2, inequalities such as $v_1 > v_2$ are to be understood as holding for each component of v_1 and v_2. We use traditional interval notation, and extend it to represent sets of integers. To illustrate, $[v_1 : v_2)$ represents the set of integers i such that $v_1 \leq i < v_2$. Finally, $(i \leftrightarrow j)$ indicates the transposition of i and j.

A list of symbols used in the thesis appears in Appendix A, and Appendix B contains a list of abbreviations.

Chapter 2

Risk estimation: the setting

This chapter introduces the model in which we shall consider the problem of risk estimation, and briefly discusses the importance of the risk estimation problem.

2.1 Introduction

Broadly speaking, one can make two major groupings of techniques used for bounding the risk of some statistical procedure. The first, and far more traditional group, is for techniques which make use of performance on a hold-out sample (test sample bounds[5]) to evaluate a model selected using a training sample. The second, more modern group, is based on evaluating models directly on their performance on the training sample (training sample bounds). There is also a hybridization technique which uses the performance on the training sample to improve estimates based on performance on a hold-out sample.

Generally, training and test sample bounds should use any other information available besides the performance on these sets, if possible. This may include, for example, prior knowledge about the distribution generating the

[5] It is traditional to speak of training and test sets. Since nothing in general precludes such samples from having identical entries, I shall however consistently use the term sample instead.

data, knowledge about the distribution generating the data inferred from the training or test sample (or even unlabeled data), and even knowledge about the structure of the set of possible fitted models.

It will become quite clear in this work that using a hold-out sample for assessing fitted models is a much simpler approach than doing so without a hold-out sample. However, in many settings data are simply not plentiful, whether because of financial, natural, or other considerations. In such settings, one would like to be able to use as much of the data as possible to fit an accurate model, rather than having to reserve data for the exclusive purpose of model assessment. With hybrid techniques available to combine training sample bounds and test sample bounds, one is faced with a trade-off in the size of the training and test sample. Clearly, improving either type of bound, or techniques for combining them, will improve the status quo. The ideal however, is to have tight bounds based only on training sample performance.

2.2 Some concepts, definitions and notation

Our focus shall be on the traditional supervised learning scenario in machine learning, but the setting we shall use is based on David Haussler's powerful decision-theoretic generalization (Haussler, 1992) of the *probably approximately correct* (PAC) learning model (Valiant, 1984). This model is closely related to what Vidyasagar (2002) presents as his "model-free" setting. A variety of other models for the learning problem exist, but investigating these is beyond the scope of this work, and in this regard we restrict ourselves to referring the interested reader to Haussler (1996, Part 1), Vidyasagar (2002, Chapter 9), Goldman (1999), and Angluin (1992) for overviews of other models with more extensive references.

In the model we shall employ, the predictors (inputs) are located in a space $\mathcal{X}$, and the response variables (outputs) in a space $\mathcal{Y}$. Predictor-response (input-output) pairs are sampled according to an underlying joint distribution $D \in \mathcal{S}$ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{S}$ is a family of distributions over $\mathcal{Z}$. For most of this work we shall assume that $\mathcal{S}$ is the class of all distributions over $\mathcal{Z}$, that is

$$\mathcal{S} = \mathcal{Q}_{\mathcal{Z}} \ ,$$

where we shall generally write $\mathcal{Q}_A$ for the class of all distributions over a set A. An independent, identically distributed[6] (i.i.d.) sample of input-output pairs (or labelled inputs) is provided, and the goal is to make good decisions based on the inputs, when actions are evaluated with respect to the output.[7] By far the most common example is when the action consists of predicting the output from the input.

Typically, we have an action class, $\mathcal{A}$, consisting of possible actions. The quality of an action with respect to an output is evaluated by a loss function, which we shall discuss in the next section.

Example 2.1. In the common example of predicting the output from the input, the action class can be identified with the output space, e.g. we can identify the action "predict 0" with the value 0, and the action "predict 1" with the value 1, so that $\mathcal{A} = \mathcal{Y} = \{0, 1\}$. □

The approach to modeling the relationship of actions, inputs and outputs we shall study employs an *hypothesis class* and a *strategy*. The hypothesis class $\mathcal{H}$ is a class of functions $h : \mathcal{X} \to \mathcal{Q}_R$ called hypotheses[8], each mapping each input $x \in \mathcal{X}$ to a distribution over some set R. If the distribution is entirely concentrated on a single value for every hypothesis in the class, we call the hypothesis class deterministic. Otherwise we call the hypothesis class stochastic. For a deterministic hypothesis class, if we have that $h(x)$ is entirely concentrated on $r \in R$, we shall also write $h(x) = r$. In addition,

[6] For the Bayesian, an assumption of the sample coming from an (infinitely) exchangeable sequence is almost always adequate for our results, thanks to the de Finetti (de Finetti, 1931) and Hewitt-Savage (Hewitt and Savage, 1955) theorems (Lauritzen, 2007). We will not go into these details in this work, however.

[7] The i.i.d. assumption is technically convenient, but rather restrictive. Work has been done on relaxing this assumption by using the concepts of mixing processes. The interested reader is referred to Vidyasagar (2002, Section 2.5) and the references therein for more on this topic.

[8] Note that this is a different concept from the traditional statistical use of the term in hypothesis testing. We will refer extensively to hypothesis testing later, but the context should make it clear which meaning we have in mind.

we shall consider a class of functions into R to be a valid hypothesis class, by assuming that a function mapping to $r \in R$ corresponds to an hypothesis with a distribution concentrated entirely on r. We shall mostly deal with these "deterministic" hypothesis classes.

Together with the hypothesis class, we define a strategy for obtaining an element of the action class from an hypothesis h, an input x, and a potential source of stochasticity. We represent the strategy as a function $g : \mathcal{H} \times [0,1] \times \mathcal{X} \to \mathcal{A}$, mapping an hypothesis h, a value in $[0,1]$, and an input x, to the action class. In practice, it is most common that we have $R = \mathcal{A}$. The value in $[0,1]$ represents an external source of stochasticity, which we assume corresponds to a random variable (r.v.) $U \sim \mathrm{Unif}[0,1]$. Thus, given an hypothesis h and a strategy g, the corresponding action is a r.v. $g(h,U,x)$. The choice of strategy is often linked to the loss function for a given problem. We shall discuss loss functions in Section 2.3.

Example 2.2. The strategy g may be deterministic, such as the common case where we define $g(h,u,x) = \mathbb{E}_{r \sim h(x)}\, r$ for all $u \in [0,1]$ (assuming that such a mean is defined). □

Example 2.3. A very common and important example, which is a special case of the previous example, is when we have $R = \mathcal{A}$, and the hypothesis class is deterministic. In this case, we define the *identity strategy* $\mathrm{id}_{\mathcal{A}}$ by $g(h,u,x) = h(x) \in R$. □

Example 2.4. More generally, g may be the realization of a random variable based on $u \in [0,1]$. In this case, we say g is stochastic. An example is when $g(h,u,x)$ is obtained by sampling from $h(x)$: defining $h_u(x)$ as the $100u$-th percentile of $h(x)$ (assuming a meaningful concept of percentile in this context), the strategy employed is $g(h,u,x) = h_u(x)$. □

For a deterministic class $\mathcal{H}'$ of functions from $\mathcal{X}$ into $\mathcal{A}$, we define an associated stochastic hypothesis class, the *Gibbs class* $\mathcal{G}_{\mathcal{H}'}(\mathcal{Q})$ associated with $\mathcal{H}'$, indexed by $\mathcal{Q}$ (which is a class of distributions over $\mathcal{H}'$):

$$\mathcal{G}_{\mathcal{H}'}(\mathcal{Q}) = \{h_Q : Q \in \mathcal{Q}\} \quad ,$$

where $h_Q : \mathcal{X} \to \mathcal{Q}_{\mathcal{A}}$ is defined by

$$[h_Q(x)]\,(A) = \mathbb{P}_{h' \sim Q}\left\{h'(x) \in A\right\}$$

for all subsets $A \subseteq \mathcal{A}$. Finally, we write $\mathcal{G}_{\mathcal{H}'} = \mathcal{G}_{\mathcal{H}'}(\mathcal{Q}_{\mathcal{H}'})$. We call $h' \in \mathcal{H}'$ a *base hypothesis* of $\mathcal{G}_{\mathcal{H}'}(\mathcal{Q})$, and $\mathcal{H}'$ the base hypothesis class. Gibbs classes are important stochastic hypothesis classes, which we shall consider in Chapter 8.

Example 2.5. A number of strategies shall be relevant in Chapter 8 when the hypothesis class is a Gibbs class.

A deterministic strategy in this scenario is the *maximum a posteriori (MAP) strategy*: let mode(·) denote the mode[9] of a distribution. Then the MAP strategy corresponds to the choice[10]

$$g(h_Q, u, x) = \operatorname{mode}(h_Q(x)) \ .$$

A second deterministic alternative is the *Bayes strategy*, determined by

$$g(h_Q, u, x) = \mathbb{E}_{r \sim h_Q(x)}\, r \ .$$

It can be shown that this strategy corresponds to making the average prediction of h' on x when the base hypothesis h' is sampled from the distribution Q, i.e.

$$g(h_Q, u, x) = \mathbb{E}_{h' \sim Q}\, h'(x) \ .$$

A third strategy is the stochastic *Gibbs strategy*, which shall receive plenty of attention later. In this case, we use u to sample from the distribution h_Q. Let $r_u \in R$ be the $100u$-th percentile of $h_Q(x)$. Then the Gibbs strategy, defined by

$$g(h_Q, u, x) = r_u \ ,$$

corresponds to sampling an hypothesis h' according to the distribution Q, and predicting $h'(x)$. □

[9] We assume a deterministic method for selecting a unique mode.

[10] This does not correspond exactly to the MAP strategy outlined in, for example, Definition 3.6 of Herbrich (2002). Their definition could be exressed as $h_{\text{mode}(Q)}(x)$, but we may not have direct access to Q, since more than one Q may map to the same h_Q. The differences between our approaches disappear for the Bayes and Gibbs strategy we consider next.

We call the combination of an hypothesis with such a strategy,

$$g_h(x, u) = g(h, u, x) \ ,$$

a *decision rule* (since we can make a decision, i.e. select an action from $\mathcal{A}$, based on an input x by applying the strategy to h, u and x). We call the set of decision rules $\mathcal{W} = \{g_h : h \in \mathcal{H}\}$ the *decision class*. A decision class is said to be stochastic when the decision rules g_h are stochastic. When the decision rules are deterministic, the value of u is irrelevant, and we shall simply write $g_h(x) = g_h(x, u)$.

Example 2.6. A large class of statistical and machine learning techniques known as *thresholding classifiers* perform binary classification by calculating a real value from the input, and then comparing the real value to a specified threshold s. One notable class is the class of (binary) voting classifiers, which includes (the two-class versions of) the well-known techniques of bagging (Breiman, 1996) and boosting (Schapire, 1999).

For such a prediction problem, we can assume $\mathcal{A} = \mathcal{Y} = \{0, 1\}$. However, the hypotheses output real values.[11] In this case, the strategy function is simply $g(h, u, x) = I(h(x) \geq s)$, and the decision rule is thus

$$g_h(x) = I(h(x) \geq s) \ .$$

□

In many cases, for a given hypothesis class $\mathcal{H}$ and strategy g, we can find a transformation ϕ such that, for every $h \in \mathcal{H}$, we have $g_h = g_{\phi(h)}$. In such a case, we call $\phi(\mathcal{H}) = \{\phi(h) : h \in \mathcal{H}\}$ a surrogate hypothesis class for $\mathcal{H}$.

In many cases, it will turn out to be useful to consider a simpler hypothesis class obtained as a surrogate hypothesis class. The intuition underlying the use of surrogate classes is based on the fact that many results we derive for training sample based estimates include a term reflecting complexity of the hypothesis class. Modifying each function by replacing it with a “less complex” function which has identical loss behaviour, but a simpler structure, thus provides one with tighter bounds.

[11] Equivalently, they output distributions concentrated entirely on a real value.

Example 2.7. In the case of thresholding classifiers with a threshold s, a surrogate hypothesis class that is often used is a trimmed class corresponding to $\mathcal{H}$.

In general, consider a class $\mathcal{H}$ of functions over $\mathcal{X}$, and two functions, $\gamma^- \leq \gamma^+$. Define the (γ^-, γ^+)-trimming of a class $\mathcal{H}$ by

$$\pi_{(\gamma^-,\gamma^+)}(\mathcal{H}) = \left\{\pi_{(\gamma^-,\gamma^+)}(h) : h \in \mathcal{H}\right\} \ ,$$

where $\pi_{(\gamma^-,\gamma^+)}(h)$ is defined pointwise by

$$\pi_{(\gamma^-,\gamma^+)}(h)(x) = \begin{cases} \gamma_-(x), & h(x) \leq \gamma_-(x) \\ h(x), & \gamma_-(x) < h(x) < \gamma_+(x) \\ \gamma_+(x), & h(x) \geq \gamma_+(x) \end{cases} \ . \tag{2.1}$$

Effectively, we trim the functions h to lie in a band specified by γ^- and γ^+. When $\gamma^- = -\gamma^+$, we shorten $\pi_{(\gamma^-,\gamma^+)}$ to π_{γ^+}.

One common choice is a constant γ. This is a suitable surrogate class for thresholding classifiers when the strategy involves thresholding at zero. More generally, if the threshold is at s, a suitable surrogate class is $\pi_{(s-\gamma,s+\gamma)}(\mathcal{H})$ for a constant γ. □

Example 2.8. Trimmed classes with constant upper and lower functions can be viewed in another light. Specifically, we can write

$$\pi_{(\gamma^-,\gamma^+)}(h) = \pi_{(\gamma^-,\gamma^+)} \circ h \ ,$$

where the function $\pi_{(\gamma^-,\gamma^+)}$ is a piecewise linear function mapping into $[\gamma^-, \gamma^+]$.

More generally, for an arbitrary function ϕ, one may be interested in the class of functions obtained by composing each function in $\mathcal{H}$ with ϕ.

When ϕ maps into a subset of the range of the functions in $\mathcal{H}$, we call ϕ a *squashing function.* A useful property for squashing functions is Lipschitz continuity[12], as this typically means the squashed function class can not be much worse behaved than the original class.

An example: consider a thresholded classifier thresholding real values of functions in $\mathcal{H}$ at 0. Then composing the hypothesis class with the translated logistic function

$$\phi(v) = \frac{1}{1+e^{-v}} - \frac{1}{2}$$

[12] A function ϕ is said to be Lipschitz continuous if, for some constant K, $|\phi(v_1) - \phi(v_2)| \leq K\|v_1 - v_2\|$ for all $v_1, v_2 \in \text{domain}(\phi)$. If this holds, we also say that ϕ satsifies a Lipschitz condition with (Lipschitz) constant K.

before applying the strategy does not change any of the decision rules.

In addition ϕ maps into $[-\frac{1}{2}, \frac{1}{2}]$, so that the squashed class $\phi(\mathcal{H})$ is a surrogate class for $\mathcal{H}$. □

For an hypothesis class $\mathcal{H}$, an $\mathcal{H}$-algorithm is any procedure which selects an hypothesis h in $\mathcal{H}$ together with a strategy g. A wide variety of classical statistical techniques and machine learning approaches match this description. Specifically, an $\mathcal{H}$-algorithm Θ is a mapping $\Theta : \bigcup_{i=1}^{\infty} \mathcal{Z}^i \to \mathcal{H} \times \mathcal{A}^{\mathcal{Q}_R \times [0,1]}$. Θ is called stochastic or deterministic based on the nature of the decision class. Note that a technique may be an $\mathcal{H}$-algorithm for a certain class $\mathcal{H}$, but not for another. In fact, many procedures inherently specify an hypothesis class for which they are an $\mathcal{H}$-algorithm. Generally the hypothesis class is clear from the context, and we shall simply refer to an algorithm. More generally however, an algorithm is any procedure which is an $\mathcal{H}$-algorithm for some hypothesis class $\mathcal{H}$.

Generally, the labelled inputs are used to guide the selection of an hypothesis from the hypothesis class, but they are typically also used to assess the quality of the selected hypothesis. A sample of l input-output pairs is typically split into a so-called *training sample* S and a *test (hold-out) sample* T. Traditionally, the training sample is used to select an hypothesis, and the test sample is used to assess the quality of the selected hypothesis. More recent advances in techniques for assessing hypothesis quality means that these names may soon be rendered outdated: now, methods for using the training sample to assess the hypothesis selected on the base of the same data, as well as bootstrap and cross-validation (CV) approaches of the past decades, mean that explicit hold-out samples are becoming less common for problems where data sets are small. Such problems are being tackled much more often by modern statisticians, as theoretical advances and more powerful computers allow high-dimensional problems to be tackled, even with relatively small samples — a typical example is the analysis of microarray data in genetics, where there are often thousands of predictors (typically 20000 or more), but the sample sizes are typically less than a hundred (Dougherty, 2001). In any case, the training sample size will be denoted by m, while the hold-out sample size, $k = l - m$, may be zero.

Choosing an appropriate hypothesis from an hypothesis class is the subject of the *learning problem*. There are hundreds, if not thousands, of proposed approaches to the learning problem, but the relative merits of these techniques are outside the scope of this study. This study will focus on the problem of assessing decision rules.

Example 2.9. Consider the linear model $Y = \beta_0 + \beta^T X + \varepsilon$ with $\mathbb{E}\,\varepsilon = 0$. If we assume that X and ε are independent and distributed normally, it follows that Y is distributed normally. Suppose furthermore that we know the variance of ε, and the covariance matrix of X. In that case, the distribution of $Y|X$ is a function only of β_0 and β. In this scenario, we could regard each distribution for $Y|X$ implied by a specific (β_0, β) as an hypothesis, and the hypothesis class could be the collection of distributions for all combinations of (β_0, β).

Suppose an hypothesis $h_{(\beta_0^*, \beta^*)}$ corresponding to (β_0^*, β^*) is selected, and that the action class is $\mathcal{Y}$. Now we consider two strategies, and the resulting decision rules. The most familiar to statisticians will be to select the mean of the conditional distribution, $\mathbb{E}\, h_{(\beta_0^*, \beta^*)}(X)$. This strategy results in regression. The other strategy mentioned briefly above involves obtaining a decision by sampling from $h_{(\beta_0^*, \beta^*)}(X)$. □

The framework sketched so far is rather more general than is commonly needed. Particularly, it is very common that $R = \mathcal{A} = \mathcal{Y}$, that we use the identitity strategy, and that $\mathcal{H}$ is deterministic. In this case, the hypotheses simply map into $\mathcal{Y}$ rather than to a distribution over $\mathcal{Y}$. This is, of course, a special case of the general framework, where the conditional distributions place all their mass on single points. In this scenario, the decision class is of course deterministic.

When the cardinality of $\mathcal{A}$ is two, and the decison class $\mathcal{W}$ is deterministic, we can define the *concept classes* $\mathcal{C}_0(\mathcal{W})$ and $\mathcal{C}_1(\mathcal{W})$ corresponding to $\mathcal{W}$ (without loss of generality, we assume $\mathcal{A} = \{0, 1\}$). Then $\mathcal{C}_0(\mathcal{W}) = \{c_w : w \in \mathcal{W}\}$, where $c_w = \{x \in \mathcal{X} : w(x) = 0\}$. $\mathcal{C}_1(\mathcal{W})$ is identical, but with $c_w = \{x \in \mathcal{X} : w(x) = 1\}$. The sets c_w are called 0-concepts and 1-concepts respectively. It is common that $\mathcal{C}_0(\mathcal{W}) = \mathcal{C}_1(\mathcal{W})$, in which case we write

$\mathcal{C}(\mathcal{W})$ for both — this is the concept class corresponding to $\mathcal{W}$. When $\mathcal{W}$ is clear from the context, it is often omitted.

2.3 Loss functions

To assess the quality of a decision rule, we make use of a loss function. In real-world situations, the cost involved in deviations between predicted and actual outcomes can sometimes be quantified exactly, and often at least estimated. More generally, we can quantify the loss incurred for an actual outcome when a certain action or decision was made.

Specifying these costs is usually done by means of a loss function L, mapping an action-output combination to the associated cost: $L : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$. Note that since the loss function is defined on the decision rules, replacing an hypothesis class by a surrogate hypothesis class does not affect the loss on any point. When L is bounded, we shall (without loss of generality) assume its range is $[0, 1]$ unless it is explicitly stated otherwise. A one-to-one correspondence of the range and this interval is easily achieved by translation and scaling, and all the results we derive for loss functions mapping into $[0, 1]$ apply to more general loss functions by merely appropriately scaling and translating any estimates obtained. If the loss function is unbounded above, we assume its range is $[0, \infty)$, again without loss of generality. Unless stated otherwise, we shall assume that the loss function is bounded. This is necessary in order to obtain the results we desire for arbitrary distributions on $\mathcal{Z}$ (Talagrand, 1994).

If two loss functions L_1, L_2 satisfy

$$L_1(r, y) \geq L_2(r, y)$$

for all actions r and outputs y, we say that L_1 *dominates* L_2.

In the common case when the action class equals the output space, the loss function often has a form mapping to zero when both elements are equal, i.e. $L(y, y) = 0$ for all $y \in \mathcal{Y}$. In this case, L is a prametric on the output space: a generalization of the concept of a metric, requiring only positivity and that (y, y) be mapped to zero for all y.

Specifying a loss function is not always practical, though, and complicated loss functions are mathematically inconvenient. Thus it is common practice to use simpler loss functions than the real-world ones, functions which are better behaved mathematically, and still generally give a good indication of the relative quality of hypotheses. Some examples of such simplifications follow.

Example 2.10. In standard regression techniques, the sum of squared errors criterion is minimized. This corresponds to minimizing empirical risk (see below) with the squared-error loss function $L(y_1, y_2) = (y_1 - y_2)^2$. In many cases this loss function may not be appropriate, but it is still commonly used and accepted since it has desirable mathematical properties, and it is believed that small empirical risk with this loss function usually corresponds to small empirical risk for most other real-world loss functions. If this is not the case for some problem, standard regression approaches may yield a very poor solution.

Note that the strategy discussed in Example 2.9, of using the mean of $h_{(\beta_0^*, \beta^*)}$, flows naturally from this approach: the underlying strategy is to minimize the empirical risk under an appropriate loss function. Using the same underlying strategy with other loss functions will lead to other strategies, as the next example shows. □

Example 2.11. In classification problems, the misclassification rate is often used to compare hypotheses. This corresponds (as we shall see in what follows) to the use of the loss function $L(y_1, y_2) = I(y_1 \neq y_2)$. Clearly this loss function would not be very sensible in a regression setting.

Minimizing the empirical risk here leads to the strategy of selecting the mode of the selected hypothesis evaluated at the input. □

Example 2.12. Consider the loss function $L_\varepsilon(y_1, y_2) = (|y_1 - y_2| - \varepsilon)_+$. This is called the ε-insensitive loss function, where ε is an accuracy parameter which can be selected depending on the problem. This loss function is popular in robust approaches to regression, and support vector (SV) regression. It should be clear that this loss function may be a more suitable approximation to the real-world situation than the squared error loss in some cases, although its mathematical behaviour is more inconvenient. □

Example 2.13. Now consider $L_\varepsilon(y_1, y_2) = I(|y_1 - y_2| > \varepsilon)$. This loss function corresponds to the previous loss function, except that any positive loss in the previous case is now assigned a value of 1.

This loss function and that in Example 2.11 are related in that they are both indicator functions for some event. All loss functions of this form "punish" all predictions which do not meet some criterion equally, while those that do meet the criterion are not punished. For the loss function of Example 2.11, the criterion is that the predicted value must be exactly correct. This criterion is common in situations where Y is a finite set, usually with small cardinality (typically two elements). Such problems are called *classification problems.*

On the other hand, the loss functions in this example, as well as the first and third example above, are more appropriate for situations where $\mathcal{Y}$ is infinite, such as the extremely common case $\mathcal{Y} = \mathbb{R}$. These problems are generally called regression problems[13]. The criterion to be met in this example is that the predicted value must be sufficiently accurate (where the required accuracy is determined by ε). □

Loss functions of the form in Examples 2.11 and 2.13 are referred to as zero-one loss functions.

Example 2.14. Selecting an hypothesis by optimizing risk with respect to a zero-one loss function usually involves a combinatorial problem which is computationally intractible.[14] Thus, many algorithms in use today make use of a so-called *proxy loss* or *dominating loss* — this is an alternative function which is an upper bound on the original loss function. The proxy loss function is then used as a replacement loss function in order to simplify the search for good hypotheses. One desirable property of such proxy loss functions is convexity. The convexity yields many computational and theoretical advantages, but at the expense of the loss function being less representative of the underlying problem.

There are a number of popular convex proxy loss functions, an example being the hinge loss $L(y_1, y_2) = (1 - y_1 y_2)_+$ used in SV classification. A cost-benefit analysis of using some of these convex proxy loss functions, which actually led to a suggestion for an alternative to the hinge loss traditionally used for SV classification, was performed in Bartlett et al. (2003a). Further work by Peter Bartlett and his co-workers in this regard is Bartlett (2003) and Bartlett et al. (2003b). □

[13] This terminology is actually a misnomer, since regression technically refers to finding the mean response given the predictors.

[14] In general, this type of problem is *NP-hard* (Goldman, 1999).

2.3.1 Loss classes and the modified learning problem

Given a loss function L and a decision rule $w = g_h$, one can define a function $f_{w,u} : \mathcal{Z} \times [0,1] \to \text{range}(L)$ by $f_{w,u}(z) = L(w(x,u), y)$, where $z = (x,y)$. We can also define the stochastic function $f_w : \mathcal{Z} \to \mathcal{Q}_{\text{range}(L)}$ with $f_w(z)$ the distribution of $L(w(x,U), y)$, when $U \sim \text{Unif}[0,1]$. If $f_w(z)$ is entirely concentrated on $v \in \text{range}(L)$, we will also write $f_w(z) = v$.

We define the *loss class* $\mathcal{F}$ associated with the decision class $\mathcal{W}$ as

$$\mathcal{F}_{\mathcal{W}} = \{f_w : w \in \mathcal{W}\} \ .$$

If $\mathcal{W}$ is clear from the context, the subscript may be omitted. Note that although the notation does not make it explicit, the loss class $\mathcal{F}_{\mathcal{W}}$ is also dependent on the loss function. Once again, the loss class is unaffected when an hypothesis class is replaced by a surrogate.

The modified learning problem

In the framework we have sketched so far, a learning problem can be specified by a tuple $\{\mathcal{X}, \mathcal{Y}, \mathcal{S}, \mathcal{A}, \mathcal{H}, L\}$. An algorithm then selects a strategy g and an hypothesis $h \in \mathcal{H}$.

It is often useful to consider a specific transformation of a general learning problem, which we shall describe in this section. The resulting modified learning problem can almost be seen as a projection of the problem into a manageable portion of the framework sketched above (by fixing certain choices). A simple analogy in classical statistics is when results can be derived for variables with zero mean without loss of generality: similarly, results derived with the fixed choices arising from this transformation yield results for the entire framework without loss of generality.

This modified setting is convenient because, regardless of the structure of the original output space and action class, the modified output space and action class lie on the real line, allowing the use of analytic tools which may not be available for general sets.

The modified learning problem is a tuple $\{\mathcal{X}', \mathcal{Y}', \mathcal{S}', \mathcal{A}', \mathcal{H}', L'\}$ obtained from the original problem as follows. We set $\mathcal{X}' = \mathcal{Z}$, $\mathcal{Y}' = \mathcal{A}' = \mathrm{range}(L)$, and $\mathcal{H}' = \mathcal{F}_{\mathcal{W}}$ (remembering that $\mathcal{W}$ is defined in terms of $\mathcal{H}$ and g). As such, a modified hypothesis in $\mathcal{H}'$ evaluated on a point $x' = (x, y) \in \mathcal{X}'$ is a distribution over range(L) (or a specific value in range(L) if $\mathcal{H}$ and g are deterministic). We shall employ the identity strategy $g' = \mathrm{id}_{\mathcal{A}'}$, so that $\mathcal{W}' = \mathcal{H}'$. The modified loss function, L', has domain range(L) × range(L): we define $L'(l_1, l_2) = l_1$ (so the second argument is irrelevant). Finally, $\mathcal{S}'$ is the set of all couplings between an element of $\mathcal{S}$ and any distribution over range(L). In this modified problem, it is assumed the modified input-output pairs in $\mathcal{Z}' = \mathcal{Z} \times \mathrm{range}(L)$ are generated by a distribution D' such that the marginal distribution of the modified input is D. Finally, we associate any modified predictor-response pair $((x, y), y') \in \mathcal{Z}'$ with the predictor-response pair (x, y). In the modified setting, the strategy is fixed as $\mathrm{id}_{\mathcal{A}'}$, and an algorithm need only select an $h \in \mathcal{H}'$.

In this setting, consider an arbitrary predictor-response pair $(x, y) \in \mathcal{Z}$, an arbitrary $h \in \mathcal{H}$, and an arbitrary $y' \in \mathcal{Y}'$. Then,

$$\begin{aligned} L'({g'}_{h'}(x'), y') &= g'(h'(x')) \\ &= h'(x') \\ &= f_{g,h}(x, y) \\ &= L(g_h(x), y) \ , \end{aligned}$$

showing that both approaches behave identically with respect to their losses on points in $\mathcal{Z}$ (regardless of the value of y', and thus the exact form of the distribution D').

This result means that estimates of risk for the modified learning problem apply directly to estimates of risk for the original problem. This is very useful, especially because many results have been obtained for problems of the form of the modified learning problem.

2.4 Risk and error

For $m, k \in \mathbb{N}$, consider the training sample

$$S = [\![(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)]\!]$$

and the test sample

$$T = [\![(x_1^*, y_1^*), (x_2^*, y_2^*), \cdots, (x_k^*, y_k^*)]\!] \quad .$$

We also denote the empirical distribution w.r.t. the elements of the training sample by S, and w.r.t. the elements of the test sample by T, i.e.

$$S(x, y) = \frac{1}{m} \sum_{i=1}^{m} I\left((x, y) = (x_i, y_i)\right) \quad ,$$

and

$$T(x, y) = \frac{1}{k} \sum_{i=1}^{k} I\left((x, y) = (x_i^*, y_i^*)\right) \quad .$$

In general, we shall use the same symbol for a sample and the empirical distribution w.r.t. the elements of that sample, in order to reduce notational clutter. In addition, we shall sometimes use the symbol to denote the set of elements of the sample. In all cases, the context should make it clear which use of the symbol we are employing.

First consider a deterministic hypothesis class $\mathcal{H}$ with strategy function $g = \mathrm{id}_{\mathcal{A}}$. For an hypothesis $h \in \mathcal{H}$, we define the (true) risk $r_D(h, L)$ of h as the expected loss of h, with the expectation over the distribution D,

$$r_D(h, L) = \mathbb{E}_{(x,y)\sim D}\, f_h(x, y) = \mathbb{E}_{(x,y)\sim D}\, L(h(x), y) \quad .$$

We define the apparent (or training) risk $r_S(h, L)$ and the test risk $r_T(h, L)$ of h in the same way, but with the expectation over S and T respectively. The training and test risk are sometimes also known as the *holdout* and *resubstitution* estimates, respectively (e.g. Devroye et al., 1996).

We define the (true) error $e_D(h, \mathscr{E})$ of h w.r.t. the predicate $\mathscr{E}$ as the risk of h when using the zero-one loss function $I(\mathscr{E}(h(x), y))$, with the apparent (or training) error $e_S(h, \mathscr{E})$ and test error $e_T(h, \mathscr{E})$ of h similarly as the apparent and test risk of h using the zero-one loss function $I(\mathscr{E}(h(x),y))$, respectively. Since the criterion $\mathscr{E}(h(x),y)$ within the indicator function of the zero-one loss function is generally an indication of inadequate performance for a decision rule, then the error of h,

$$\begin{aligned} e_D(h,\mathscr{E}) &= r_D(h, I(\mathscr{E}(h(x),y))) \\ &= \mathbb{E}_{(x,y)\sim D}\, I(\mathscr{E}(h(x),y)) \\ &= \mathbb{P}_{(x,y)\sim D}\left\{\mathscr{E}(h(x),y)\right\} \quad , \end{aligned}$$

is simply the probability that the decision rule h performs inadequately on a future point, or the long-term proportion of predictions which are inadequate.

Example 2.15. The most common choice of $\mathscr{E}$ is

$$\mathscr{E}(y_1,y_2) = [y_1 \neq y_2] \quad .$$

This choice is appropriate when $R = \mathcal{A} = \mathcal{Y}$, and the corresponding indicator function is the misclassification loss — see Example 2.11.

Another example is the choice of $\mathscr{E}$ corresponding to the ε-insensitive loss of Example 2.13. Again, this choice is appropriate for $R = \mathcal{A} = \mathcal{Y}$, and the corresponding choice of $\mathscr{E}$ is

$$\mathscr{E}(y_1,y_2) = [|y_1 - y_2| > \varepsilon] \quad .$$

□

Similarly,

$$\begin{aligned} e_S(h,\mathscr{E}) &= \mathbb{P}_{(x,y)\sim S}\left\{\mathscr{E}(h(x),y)\right\} \\ &= \frac{1}{m}\sum_{i=1}^{m} I(\mathscr{E}(h(x_i),y_i)) \end{aligned}$$

is the proportion of training sample points inadequately predicted, and

$$\begin{aligned} e_T(h,\mathscr{E}) &= \mathbb{P}_{(x,y)\sim T}\left\{\mathscr{E}(h(x),y)\right\} \\ &= \frac{1}{k}\sum_{i=1}^{k} I(\mathscr{E}(h(x_i^*),y_i^*)) \end{aligned}$$

is the proportion of test sample points inadequately predicted.

Note that we can also define the risk and error according to any other distribution Q over $\mathcal{Z}$ similarly: $r_Q(h, L) = \mathbb{E}_Q\, L(h(x), y)$, and $e_Q(h, \mathscr{E})$ as the risk using the zero-one loss function $I(\mathscr{E}(h(x), y))$. When Q is an empirical distribution w.r.t. a sample Q, we refer to $r_Q(h, L)$ ($e_Q(h, \mathscr{E})$) as the empirical risk (empirical loss) w.r.t. the sample Q. It is clear that if L_1 dominates L_2, we have

$$r_Q(h, L_1) \geq r_Q(h, L_2)$$

for every hypothesis h.

When L or $\mathscr{E}$ is implicit in the context, or arguments hold for all loss or indicator functions, we shall often omit them, referring simply to $r_Q(h)$ or $e_Q(h)$.

Next, we consider the general case when $\mathcal{H}$ is stochastic, and the strategy need not be the identity function. The resulting definitions generalize those above. Consider a stochastic decision rule $w(x, u)$. We define the true risk of a decision rule $w \in \mathcal{W}$ as

$$r_D(w, L) = \mathbb{E}_{(x,y)\sim D}\, \mathbb{E}_{u\sim \mathrm{Unif}[0,1]}\, L(w(x, u), y) \ \ .$$

The rest of the definitions above can be simply extended similarly. For example,

$$e_Q(w, \mathscr{E}) = \mathbb{E}_{(x,y)\sim Q}\, \mathbb{E}_{u\sim \mathrm{Unif}[0,1]}\, I(\mathscr{E}(w(x, u), y)) \ \ .$$

We will also be interested in the risk associated with an algorithm $\mathit{\Theta}$. Consider $r_D(w_S)$, where w_S is the decision rule selected by $\mathit{\Theta}$ for a training sample S. Then the risk of an algorithm $\mathit{\Theta}$ is the mean of $r_D(w_S)$ over training samples drawn from D^m:

$$r_D(\mathit{\Theta}) = \mathbb{E}_{S\sim D^m}\, r_D(w_S) \ \ .$$

With these definitions in mind, and our finding in Section 2.3 that the modified form of the original problem behaves identically to the original problem in terms of evaluations of loss, it follows that any results for the risk or error of the modified form of the problem can immediately be converted directly into results on the original problem. Which approach is more useful shall depend on the situation. In particular, some cases where the strategy function is not the identity function (such as thresholded classifiers, to be studied later) will benefit from the (more powerful) original setting. When the strategy function is the identity function, the modified setting is generally the preferable approach. Furthermore, we can in all cases replace any hypothesis class with a surrogate hypothesis class, since this does not affect any decisions made.

We conclude this chapter by referring the reader to the problem statement in Section 1.2, now that the relevant concepts have been properly introduced.

Chapter 3

Test sample estimators

Suppose an algorithm selected an hypothesis h and strategy g using a training sample S. The decision rule $w = g_h$ has true risk $r_D(w)$ with respect to a loss function L.

Test sample estimators focus on the use of the behaviour of h and w on a test sample T to obtain information about $r_D(w)$.

For general loss functions, the distribution of the loss on future data points is generally unknown, so that traditional statistical methods can not easily be applied. Therefore, most of the results presented focus on the error with respect to zero-one loss functions. In this case, the number of errors on a sample consisting of $k = |T|$ data points, has a binomial distribution with parameters k and $p = e_D(w)$: each prediction has a fixed probability of "success" (error), p, so the error status of each prediction is a Bernoulli variable; these predictions are independent, so the total number of errors is a sum of independent Bernoulli random variables. Specifically, if r of the k points in the test sample yield errors, then r is the realization of a binomial variable with parameters k and p, i.e. $r \sim \text{Bin}(k, p)$. This situation corresponds to the large volume of classical statistical literature on estimating a proportion.

3.1 Test sample point estimators

An excellent resource on point estimation in classical statistics is Lehmann and Casella (1998).

3.1.1 UMVU estimator

The most well-known estimator of a population mean from a sample is undoubtedly the sample mean. This estimator is unbiased as the method of moments (MM) estimator of the mean. Furthermore, it is the maximum likelihood (ML) estimator of the mean for a number of distributions, including the binomial distribution.

In this case, the test sample mean loss is $r_T(w)$, the test risk. For a zero-one loss function, we write $\hat{p} = \frac{r}{k} = e_T(w)$, the test error of w. Besides $\hat{p}$ being the unique ML estimator of p, we note that it is a function of r, which is a complete, sufficient statistic for the Bernoulli proportion (Dudewicz and Mishra, 1988, Example 8.2.21). Since $\hat{p}$ is an unbiased estimator of p derived from a complete, sufficient statistic, it follows from the Lehmann-Scheffé theorem (Lehmann and Scheffé, 1950; Lehmann and Casella, 1998, Lemma 2.1.10) that $\hat{p}$ is the unique best (or uniform minimum-variance) unbiased (BU or UMVU) estimator[15] of p.

3.1.2 The bias-variance trade-off

The analysis above does not consider the so-called *bias-variance trade-off* (Geman et al., 1992). In general, the most traditional measure of the quality of a point estimator is the mean square error (MSE). For an estimator $\widehat{r_D(w)}$

[15] Here "best" means that $\hat{p}$ exhibits minimum expected risk of all unbiased estimators, for any convex loss function on the parameter.

of $r_D(w)$, we have

$$\begin{aligned}\mathrm{MSE}\left(\widehat{r_D(w)}\right) &= \mathbb{E}_{T\sim D^k}\left(\widehat{r_D(w)} - r_D(w)\right)^2 \\ &= \mathbb{E}_{T\sim D^k}\left[(\widehat{r_D(w)} - \mathbb{E}_{T\sim D^k}\widehat{r_D(w)}) + (\mathbb{E}_{T\sim D^k}\widehat{r_D(w)} - r_D(w))\right]^2 \\ &= \mathbb{E}_{T\sim D^k}\left(\widehat{r_D(w)} - \mathbb{E}_{T\sim D^k}\widehat{r_D(w)}\right)^2 + \left(\mathbb{E}_{T\sim D^k}\widehat{r_D(w)} - r_D(w)\right)^2\end{aligned} \tag{3.1}$$

since the cross-product term falls away after expansion. In the last line, the first term is the variance of the estimator, $\mathbb{V}_{T\sim D^k}(\widehat{r_D(w)})$, while the second term is the square of the *bias* of the estimator,

$$\mathrm{Bias}(\widehat{r_D(w)}) = \mathbb{E}_{T\sim D^k}\widehat{r_D(w)} - r_D(w) \ .$$

More generally, decision theory specifies a loss function, which we shall call an *optimality function*[16], for evaluating the suitability of an estimator. In that scenario, the MSE corresponds to the optimality function $L(y_1, y_2) = (y_1 - y_2)^2$.

An UMVU estimator may not necessarily be the best estimator in terms of MSE (or mean error with respect to another optimality function): in our case, a biased estimator of p with lower variance than $\hat{p}$ may have a lower MSE. To complicate matters, the MSE of an estimator is usually a function of the unknown parameter $r_D(w)$. For example, $\hat{p}$ is unbiased, so its MSE is simply its variance. Since the number of errors, r, is a binomial variable, its variance is $kp(1-p)$ and the variance of $\hat{p}$ is thus

$$\frac{kp(1-p)}{k^2} = \frac{p(1-p)}{k} \ ,$$

a function of p. For fixed k, we see that the MSE of $\hat{p}$ is largest when $p = 0.5$, and smallest when p is zero or one.

A number of alternative point estimators have been introduced, many employing Bayesian ideas to "shrink" or bias the estimator to some prior estimate of the parameter. These estimators are usually called shrinkage estimators. Two other important groups of point estimators are minimax estimators and minimum risk equivariant (MRE) estimators. All the techniques

[16] The term loss function here is distinct from our usage so far. However, the decision-theoretic foundations of our framework are the reason why the two concepts share a name.

which follow are in principle capable of dealing with general optimality functions, but we shall restrict ourselves to the squared error optimality function and MSE.

3.1.3 Bayes and minimax estimators

The starting point for a Bayes estimator is a prior distribution for $r_D(w)$. Since the flexible family of Beta(α, β) distributions is a conjugate family of priors for the binomial distribution, it is a very popular choice of prior for estimating error.[17]

To obtain a Bayes estimator, the parameters α and β of the Beta distribution need to be specified in some way. There are three major approaches to doing this:

- traditional Bayes, where α and β must be fully specified before obtaining the sample;
- empirical Bayes, where the sample itself is used to determine values of α and β in the prior; and
- hierarchical Bayes, where further priors, known as hyperpriors, with fully specified distributions, are used to model α and β.

Since the Beta distribution is a conjugate family of priors for the binomial distribution, the posterior distribution of p after seeing the test sample is also a Beta distribution. If the prior has parameters α and β, the posterior distribution's parameters are $r + \alpha$ and $(k - r) + \beta$. To obtain a point estimate, we minimize the MSE using the posterior distribution. It can be shown quite easily that the optimal point estimate from this perspective is the mean of the posterior distribution[18]. This mean turns out to be a weighted average of the empirical error $\hat{p}$, and the prior estimate of the error $\frac{\alpha}{\alpha+\beta}$. The weight depends on k, with an increase in k causing more

[17] For risk, the Bayes estimation procedure will not generally be practical, unless one has information on the distribution of $r_T(w)$.

[18] The proof is analogous to the derivation of the bias-variance trade-off above.

weight to be assigned to the empirical error. Specifically, the mean of the posterior Beta$(r+\alpha,(k-r)+\beta)$ distribution is

$$\frac{r+\alpha}{r+\alpha+(k-r)+\beta}=\frac{r+\alpha}{k+\alpha+\beta}=\frac{k}{k+\alpha+\beta}\frac{r}{k}+\frac{\alpha+\beta}{k+\alpha+\beta}\frac{\alpha}{\alpha+\beta}\quad .$$

Another approach to deriving a point estimator from a Bayesian approach is the MAP estimator. Although this estimator is not optimal with respect to MSE in a Bayesian scenario, it is popular due to its similarities to the ML estimator. The MAP estimator of a parameter is, as the name suggests, the value of that parameter maximizing the posterior likelihood function derived in the Bayesian framework.

In our case, assuming a Beta(α,β) prior, the posterior likelihood of p is (from the Beta posterior distribution)

$$\mathrm{Lik}(p)=p^{\alpha+r-1}(1-p)^{\beta+(k-r)-1}\quad .$$

Setting the partial derivative of the log-likelihood to zero and solving for p reveals an extremum at $\hat{p}_{MAP}=\frac{\alpha+r-1}{\alpha+\beta+k-2}$. Further differentiation reveals the extremum to be a maximum at least when $\alpha+r>1$ and $\beta+(k-r)>1$.[19]

We see that the MAP estimate for given parameters α,β, is equal to the posterior mean estimate with parameters $(\alpha-1),(\beta-1)$. As such, in what follows, we shall restrict our attention to estimators based on the posterior mean.

Traditional Bayes

If there is no prior information available about the distribution of p, various approaches to selecting a prior exist. The *maximum entropy* (ME) continuous distribution on $[0,1]$ is the uniform distribution on $[0,1]$, which is also a Beta$(1,1)$ distribution. As a result, the ME principle for selecting a non-informative prior suggests using this prior. This leads to the maximum entropy point estimator $\hat{p}^{ME}=\frac{r+1}{k+2}$.

[19] In fact, the extremum is a maximum when $(\alpha+r-1)^{-1}+(\beta+(k-r)-1)^{-1}>0$.

Another popular non-informative prior is the *Jeffreys prior*, since it is invariant to reparameterization (Agresti and Min, 2005). Once again, it turns out that this prior is a Beta distribution, specifically a Beta$(0.5, 0.5)$ distribution, yielding the estimator $\hat{p}^J = \frac{r+0.5}{k+1}$. Note that the Jeffreys prior shrinks $\hat{p}$ towards the same value as the ME estimator (0.5), but the extent of the shrinkage is less. In addition, the *reference prior*, which is determined by maximizing the expected information gain (or Kullback-Leibler (KL) divergence) of the posterior relative to the prior, happens to yield the same distribution as the Jeffreys prior in our case (Agresti and Min, 2005).

Finally, it is worth noting that the physicist and Bayesian statistician, Edwin T. Jaynes, derived an improper prior (corresponding roughly to a "Beta(0,0)" distribution), based on an argument involving Lie groups (Jaynes, 1968). This prior results in the usual unbiased estimator $\frac{r}{k}$.

Empirical Bayes

The empirical Bayes approach uses the data sample under consideration to guide selection of appropriate parameters for the prior distribution, i.e. hyperparameters. In this scenario, one considers the marginal distribution of the data sample. This can be written as the integral of the joint distribution of the parameters and the data sample, with respect to the parameters. Since the parameters are based on the prior distribution with certain hyperparameters, the marginal distribution of the data sample is a function of these hyperparameters. Empirical Bayes then selects the hyperparameters to maximize the marginal probability of the observed sample, so that this is effectively a kind of ML approach.

In our case, assume the random parameter W of a Bin(k, W) variable V has a Beta distribution, with parameters α and β, and k is known. Then the joint distribution function of V and W is

$$f_{V,W}(v, w) = \binom{k}{v} \frac{1}{B(\alpha, \beta)} w^{v+\alpha-1} (1-w)^{(k-v)+\beta-1} \quad ,$$

where B denotes the Beta function. Integrating this with respect to w yields a marginal distribution for V which is a function of α and β. Maximizing this function is easily seen to be equivalent to maximizing the ratio of Beta functions,

$$\frac{B(v+\alpha, (k-v)+\beta)}{B(\alpha,\beta)} \ .$$

Finding the optimum choice of α and β is then generally done numerically.

Hierarchical Bayes

One can take the Bayesian approach even further: rather than specifying α and β in a Beta prior distribution, we can consider α and β as realizations of random variables $\mathscr{A}$ and $\mathscr{B}$, each with their own underlying (usually parametrized) distributions. One can then try to estimate the parameters of these distributions instead of α and β themselves. This approach, called *hierarchical Bayes*, is clearly somewhat more complicated than the Bayes methodologies outlined above, but in turn yields a more flexible model.

3.1.4 Minimax estimator

The minimax estimator is an estimator which minimizes the maximum value of the optimality function over all possible values of the estimand, in our case, $r_D(w)$. That is, a minimax estimator $\hat{p}_{MM}$ of p satisfies

$$\sup_p \mathrm{MSE}(\hat{p}_{MM}) = \inf_{p^\star} \sup_p \mathrm{MSE}(p^\star) \ ,$$

where the infimum on the right is over all point estimators $p^\star$. Thus, a minimax estimator, as the name implies, tries to optimize the worst-case scenario. It turns out that the minimax estimator typically has a Bayesian interpretation: the minimax estimator is the Bayesian estimator under a *least favourable prior*. When a Bayesian prior generates a constant MSE for all values of p, the prior is least favourable (Lehmann and Casella, 1998, Chapter 5). In the binomial estimation case, this occurs for $\alpha = \beta = \frac{\sqrt{k}}{2}$. As such, the minimax estimator of p is

$$\hat{p}_{MM} = \frac{r + \frac{\sqrt{k}}{2}}{k + \sqrt{k}} \ .$$

3.1.5 Minimum risk equivariant estimators

For many loss functions, it may be reasonable to expect an estimator of risk to be equivariant with respect to $\phi(v) = 1 - v$. Put another way, we may expect an estimator to satisfy

$$1 - \widehat{r_D(w, L)} = \widehat{r_D(w, 1 - L)} \ . \tag{3.2}$$

The implications of this assumption depend on the distribution of the loss, but in the case of a zero-one loss function, one obtains that the estimator satisfying (3.2) with minimum MSE is simply $\hat{p}$.[20]

3.1.6 Estimators for thresholded classifiers

Further point estimates are available for thresholded classifiers. These estimates make use of the value of $h(x)$ before the strategy g thresholds it at s.

The first estimator we shall discuss is the Glick smoothed estimate. This estimator, based on work in Glick (1978), is presented as a training set point estimator in Chapter 31 of Devroye et al. (1996), but the definition applied there could equally well be applied to a test set. We further modify their presentation by generalizing the smoothing function to cases where the underlying value is not necessarily a probability estimate.

Note that the regular estimator $e_T(w)$ is not robust to the classification of a point near the threshold in the following sense: a small perturbation of the input can lead to $e_T(w)$ abruptly changing by an amount of $\frac{1}{k}$. The Glick smoothed estimate is defined in terms of a smoothing function ϕ. The idea of the smoothed estimate is that the classification of sample points for which $h(x)$ is close to the threshold are sensitive to perturbations, so their contribution to the determination of the error rate should not be as large as that of those for which $h(x)$ is distant from s. This smoothed estimate is thus more robust and exhibits a lower variance than $e_T(w)$. This is the same

[20] This follows since $\hat{p}$ is an equivariant UMVU estimator — see Lehmann and Casella (1998, Lemma 1.23).

intuition we shall encounter in Section 5.6 when we discuss margin bounds. The smoothing function ϕ encodes our understanding of the relationship between robustness of classification and distance from the threshold.

Specifically, for a decision rule $w = g_h$, we define

$$\hat{p}_G = \frac{1}{k}\sum_{i=1}^{k} \left(y_i^* \left(1 - \phi(h(x_i^*))\right) + (1 - y_i^*)\phi(h(x_i^*))\right) \quad ,$$

where the G stands for Glick, and the (x_i^*, y_i^*) are the elements of the test set T. Generally, we expect ϕ to be monotonically increasing with range $[0, 1]$. Note that setting $\phi(v) = I(v \geq s)$ yields the standard test sample estimate.

One common application area of this approach is when $h(x)$ is an estimate of the regression function $\mathbb{E}(Y|X = x)$, and $s = \frac{1}{2}$. Since s is the midpoint of possible values of $h(x)$, and the distance of $h(x)$ from s seems equally relevant on both sides, it is typical to expect ϕ to satisfy $\phi(s - v) = 1 - \phi(s + v)$ in this case. It is common to choose ϕ to be the identity function or a sigmoid function. In general, this estimator is only as good as the choice of the smoothing function.

The final estimate of error we shall consider is also for the case where $h(x)$ is an estimate of the regression function, and the strategy involves thresholding $h(x)$ at $\frac{1}{2}$. Put another way, $h(x)$ is an estimate of $\mathbb{P}(Y = 1|X = x)$. Suppose we had access to the true conditional probabilities. Then

$$\begin{aligned} p = e_D(w) = e_D(g_h) &= \mathbb{E}_{(X,Y)\sim D}\, L(g(h(X)), Y) \\ &= \mathbb{E}_{(X,Y)\sim D}\, I(g(h(X)) \neq Y) \\ &= \mathbb{E}_{X\sim D_X}\, \mathbb{E}_{Y\sim D_{Y|X}}\, I\left(I\left(h(X) \geq \frac{1}{2}\right) \neq Y\right) \quad , \end{aligned}$$

where D_X denotes the marginal distribution of the input, and $D_{Y|X}$ denotes the conditional distribution of the output for a given input. Expanding the inside expectation, the expression equals

$$\mathbb{E}_{X\sim D_X}\left[\mathbb{P}\{Y = 0|X\} I\left(I\left(h(X) \geq \frac{1}{2}\right) = 1\right) + \mathbb{P}\{Y = 1|X\} I\left(I\left(h(X) \geq \frac{1}{2}\right) = 0\right)\right] \quad .$$

It is easy to verify that $I(I(\mathscr{E}) = 1)$ reduces to $I(\mathscr{E})$, and similarly that $I(I(\mathscr{E}) = 0)$ reduces to $I(\bar{\mathscr{E}})$, where the bar denotes the complement of the predicate $\mathscr{E}$. Thus, we obtain

$$\mathbb{E}_{X \sim D_X}\left[\mathbb{P}\{Y = 0|X\} I\left(h(X) \geq \frac{1}{2}\right) + \mathbb{P}\{Y = 1|X\} I\left(h(X) < \frac{1}{2}\right)\right] \ .$$

This formulation leads to the so-called *posterior probability estimator* (PPE): the estimate is based on two approximations: first, approximating D_X by T_X, the marginal of X with respect to the uniform distribution over the test set; second, we use $h(x)$ to approximate $\mathbb{P}\{Y = 1|X = x\}$. These approximations result in the estimate

$$\hat{p}_{PPE} = \frac{1}{k} \sum_{i=1}^{k} \left[(1 - h(x_i^*))\, w(x_i^*) + h(x_i^*)\, (1 - w(x_i^*)) \right] \ .$$

For more on the origin and development of this and similar estimators, the reader is referred to Chapter 31 of Devroye et al. (1996) (but note that the context there is for training sample point estimators).

In this section, we have introduced and discussed a number of point estimators. As mentioned, the main criterion for evaluation of a point estimator is its MSE (or more generally, its mean optimality), while some other criteria are briefly mentioned. There are however, a variety of other criteria (outside the scope of this work) for evaluating point estimators, such as the efficiency, consistency, admissibility, stability, and the (asymptotic) distribution of the estimator.

3.1.7 Summary

Each of the point estimators discussed here has its merits and drawbacks, being more suitable in some situations, and less so in others. The most relevant issue for us in determining which point estimator is most suitable is the approximate range one expects p to lie in: since the MSE of an estimator is usually a function of p, the best estimator (in terms of MSE) will depend on where we think p lies (Botha, 1992): p close to 0.5 will yield a different selection of good estimators to a situation where p is probably smaller than 0.05. When there is no indication of which value of p is more likely, the minimax estimator is typically the best option. However, if you do expect your parameter to lie in a certain region, it's no good having an estimator that performs poorly there. The idea is then to rather improve it there, at the cost of other regions (assuming one can not get uniform improvement). This is what the Bayes estimator does, where the prior represents the statistician's opinion of the relevant values of p. The minimax estimator also does this — it improves the performance of an estimator at all the values of p where it performs poorly, at the expense of the values of p exhibiting better performance: this is why a minimax estimator typically has constant risk.

3.2 Test sample interval estimators

As discussed earlier, point estimates are often not sufficient indicators of locality of a parameter, even in combination with the variance of the estimator. Interval estimators are an alternative method of specifying locality, which address some of the shortcomings of point estimators.

In what follows, we shall be considering various approaches to constructing interval estimators for risk on the basis of a test sample T. That is, for a given decision rule w, we hope to be able to construct a region $A(T, w)$ such that, with high confidence, the region contains the true risk $r_D(w)$ of the decision rule.

Thus, we seek a statement of the form

$$\mathbb{P}_{T\sim D^k}\left\{r_D(w) \in A(T, w)\right\} \geq 1 - \delta \ .$$

Concepts from statistical hypothesis testing will be prominent in what follows, since there is a duality between statistical hypothesis testing and confidence intervals. Loosely speaking, a statistical hypothesis test for a parameter t determines a confidence interval: the $100(1 - \delta)\%$ confidence interval associated with a statistical hypothesis test consists of those values v for which a null hypothesis $t = v$ would not be rejected. More details on this duality are available in Dudewicz and Mishra (1988, Section 10.8). In addition, an excellent resource on statistical hypothesis testing is Lehmann and Romano (2005).

3.2.1 Measures of deviation

The expression $r_D(w) \in A(T, w)$ will often be represented by

$$\psi\left(r_D(w), r_T(w)\right) \leq \epsilon(T, w) \tag{3.3}$$

or a similar expression, where ψ is some *measure of deviation* (typically a prametric) between $r_D(w)$ and $r_T(w)$. Furthermore, $\epsilon(T, w)$ is not a function of $r_D(w)$ or $r_T(w)$, and ψ does not depend on T or D. ψ need not be analytically invertible, but to be practically useful, some technique for obtaining the corresponding $A(T, w)$ from $\epsilon(T, w)$, given ψ, is necessary. Some of the measures of deviation ψ we shall consider are presented in what follows, along with a derivation of the corresponding intervals for t implied by $\psi(t, v) \leq \epsilon$.

- *(Upper) regular deviation*: $\psi\left(t, v\right) = t - v$; this leads to a one-sided interval $(-\infty, v + \epsilon]$. A lower regular deviation and absolute regular deviation can also be used, yielding the intervals $[v - \epsilon, \infty)$ and $[v - \epsilon, v + \epsilon]$ respectively.

- *(Upper) relative deviation*: $\psi\left(t, v\right) = \frac{t-v}{\sqrt{t}}$; this measure of deviation assumes that $t > 0$. Inverting this measure of deviation involves solving a quadratic equation in $\sqrt{t}$. The resulting interval is

$$\left(0, v + \frac{\epsilon^2\left(1 + \sqrt{1 + \frac{4v}{\epsilon^2}}\right)}{2}\right] . \tag{3.4}$$

  A lower and absolute relative deviation are defined similarly. The upper relative deviation is a special case of the upper Bartlett-Lugosi (B-L) ν-deviation, which we shall discuss later.

- *(Upper) Rao deviation*: $\psi\left(t,v\right) = \frac{t-v}{\sqrt{t(1-t)}}$; this measure of deviation assumes that $t \in (0,1)$. Inverting it once again involves solving a quadratic equation, which results from multiplying ϵ by the denominator here, and squaring. The resulting interval is

$$\left(0, \frac{v+\frac{\epsilon^2}{2}+\epsilon\sqrt{v(1-v)+\frac{\epsilon^2}{4}}}{1+\epsilon^2}\right] . \tag{3.5}$$

Once again a lower and absolute deviation can be defined. The resulting interval for the lower deviation is

$$\left[\frac{v+\frac{\epsilon^2}{2}-\epsilon\sqrt{v(1-v)+\frac{\epsilon^2}{4}}}{1+\epsilon^2}, 1\right) .$$

This is the measure of deviation used, for example in the Wilson score interval, which is based on the Rao score hypothesis test.

- *(Upper) Wald deviation*: $\psi\left(t,v\right) = \frac{t-v}{\sqrt{v(1-v)}}$; results for this measure only apply for $v \in (0,1)$. Multiplying ϵ by the denominator leads to the interval $(-\infty, v+\epsilon\sqrt{v(1-v)}]$. Results for a lower and absolute deviation follow similarly. This is the measure of deviation employed in the Wald interval for a binomial proportion, based on the corresponding Wald hypothesis test.

- *(Upper) Pollard ν-deviation*: $\psi_\nu\left(t,v\right) = \frac{t-v}{\nu+\sqrt{t}+\sqrt{v}}$, with $\nu > 0$. Here, t and v are assumed to be positive. Lower and absolute deviations are analogous. This measure of deviation was apparently proposed in Pollard (1986), together with the following measure of deviation. The intervals resulting from the upper and lower deviations can be obtained by solving quadratic equations.

- *(Upper) Pollard-Haussler (P-H) ν-deviation*: $\psi_\nu\left(t,v\right) = \frac{t-v}{\nu+t+v}$, with $\nu > 0$. Again lower and absolute versions can be defined. This deviation measure, proposed by Pollard to address *inter alia* his concern with the behaviour of relative deviation when $r_D(w) = 0$, was further investigated in Haussler (1992). The interval resulting from the upper deviation is $(0, \frac{(1+\epsilon)v+\epsilon\nu}{1-\epsilon}]$. The lower interval can be obtained similarly.

One useful feature of this measure of deviation is that the two-sided P-H deviation measure is a metric on $\mathbb{R}^+$.

- *(Upper) B-L ν-deviation*: $\psi_\nu(t,v) = \frac{(t-v)-\nu}{\sqrt{t}}$. Again, it is assumed that $t > 0$. The lower B-L ν-deviation is $\psi_\nu(t,v) = \frac{(v-t)-\nu}{\sqrt{v}}$. An absolute B-L deviation is not defined. Note that the upper B-L 0-deviation is the upper relative deviation above. These deviation measures were proposed in Bartlett and Lugosi (1999). By the upper bound's similarity to the upper relative deviation, we easily see that substituting $v+\nu$ for v in (3.4) will provide an interval here, yielding

$$\left(0, (v+\nu) + \frac{\epsilon^2(1+\sqrt{1+\frac{4(v+\nu)}{\epsilon^2}})}{2}\right] \quad .$$

An interval for the lower B-L ν-deviation is easily obtained as

$$[v - \epsilon\sqrt{v} - \nu, \infty) \quad .$$

The following inequality derived in Corollary 1 of Bartlett and Lugosi (1999) is useful[21]: if the upper B-L ν-deviation $\psi_\nu(t,v) \le \epsilon$, then, for any $\eta > 0$,

$$t \le (1+\eta)\left[v + \nu + \frac{\epsilon^2}{\eta}\right] \quad .$$

A similar result can be obtained from the lower B-L ν-deviation. A notable use of these inequalities is converting results for the P-H ν-deviation to results for the B-L ν-deviation.

- *Inverse distribution deviation*:

$$\psi(t,v) = \mathbb{P}_{V\sim Q_t}\{V \le v\} \quad .$$

This measure seems rather obscure at first. However, generally it is applied where Q_t is the (approximate) distribution of a statistic, when the value of some unknown parameter is t. It follows that $\psi(t,v) \le \epsilon$ when $F_{Q_t}(v) \le \epsilon$, where F_{Q_t} denotes the cumulative distribution function (c.d.f.) of Q_t. Thus one obtains the interval $\{a : Q_t(v) \le \epsilon\}$.

[21] The result in the reference is actually a little stronger.

There are no specific one-sided or absolute versions of this deviation measure: instead, the nature of the resulting interval will depend on the statistic used. We note that the inverse distribution deviation generally underpins confidence sets which are generated by inverting hypothesis tests. More details on this are in Section 3.2.3.

- *(Lower) binomial tail deviation*: $\psi^L(t,v) = \mathbb{P}_{V\sim\mathrm{Bin}(k,\frac{t}{k})}\{V \leq v\}$. Clearly, this is just a special case of the inverse distribution deviation, where Q_t assumes a binomial distribution. This can be seen as the most natural measure of deviation for the number of errors of a decision rule. This measure of deviation naturally leads to a (lower) one-sided interval — however, this interval must be found numerically for a specific value of t[22]. Since the binomial tail deviation is non-increasing in t, one can find the root of $\psi(t,v) = \epsilon$ by employing a line search technique such as the secant method or Brent's method (Brent, 1973). We denote this root by $\mathrm{LBT}(\frac{v}{k}, k, \epsilon)$ where LBT stands for lower binomial tail. The upper binomial tail deviation is

  $$\psi^U(t,v) = \mathbb{P}_{V\sim\mathrm{Bin}(k,\frac{t}{k})}\{V \geq v\} \quad .$$

  This is obtained from the inverse distribution deviation for a transformation of the binomial variable: the interval obtained corresponds to that obtained by the lower binomial tail deviation $\psi^L(k-t,v)$. One can obtain an upper interval from an upper binomial tail deviation, and the upper bound is denoted by $\mathrm{UBT}(\frac{v}{k}, k, \epsilon)$.

- *Kullback-Leibler deviation*: $\psi(t,v) = \mathrm{KL}(t||v)$; here it is assumed that $t \in [0,1]$, and to apply it, we need $v \in (0,1)$. Once again, this measure of deviation is not analytically invertible, but can be inverted with a line search. This measure of deviation inherently provides two-sided intervals, but the interval is usually not symmetric. We denote

[22] In order to evaluate the deviation for a given t, a cumulative binomial probability must be evaluated. For large k, many of these binomial probabilities are extremely small, leading to potential underflow problems when evaluating the probabilities on a computer. One of the most important techniques for addressing this issue is that of instead evaluating the log of the cumulative probability function. Another technique for speeding up the calculation of these intervals is the use of Stirling's approximation to calculate factorials.

the lower and upper endpoints obtained by inverting $\mathrm{KL}(t||v) = \epsilon$ by $\mathrm{LKL}(v||\epsilon)$ and $\mathrm{UKL}(v||\epsilon)$ respectively.

Given a deviation measure ψ, the problem of obtaining a confidence set reduces to finding appropriate (approximate) value of $\epsilon(T, w)$.[23] Generally, a standard approach to constructing confidence intervals will indeed specify ψ — we shall see this later. First we turn our attention to the desirable properties of an interval estimator.

3.2.2 Criteria for interval estimators

There are a wide variety of generic interval estimation techniques, and for most common problems, there are a number of additional techniques designed to generate suitable interval estimators. The reason for this plethora of interval estimators is that the ideal interval estimator is typically unattainable. Instead, there are a variety of characteristics which are considered desirable for an interval estimator (some parameter- or problem-specific), and most interval estimators offer a trade-off between these criteria.

Coverage

The first issue is a rather philosophical one, and is related to the concept of *coverage* of an interval. An interval estimator typically contains a parameter of interest with a certain probability, but often that probability is a function of the underlying parameter value. Thus for a given parameter value, we can calculate the probability of the interval estimator containing that value. This is the coverage of the interval at that parameter value. The confidence coefficient of the interval estimator is the infimum of the coverage over all parameter values, and traditionally, an interval estimator is said to be a $100(1-\delta)\%$ confidence interval if the confidence coefficient is at least $1-\delta$. However, many interval estimators which are actually only approximate confidence intervals, due to the use of asymptotic results, for example, have

[23] Although our notation does not make it explicit, note that this value is dependent on the value of δ.

become ubiquitous in practice. Many of these approximate intervals may have a confidence coefficient well below $1-\delta$, yet, for various reasons (such as ease of use) they remain popular. When an interval estimator is described as a $100(1-\delta)\%$ confidence interval, but its confidence coefficient is actually less than $1-\delta$, we say *undercoverage* occurs. In this case, for certain parameter values, the interval estimator will not contain the actual parameter with probability at least $1-\delta$.

At this stage it may seem clear-cut to support the traditional view of interval estimation. But sometimes, as in the case of the problem under investigation here, the issue is not so clear: sometimes the available interval estimators which exhibit a confidence coefficient of at least $1-\delta$ have a significantly higher coverage for almost all of the possible parameter values, while an interval estimator with an only slightly lower confidence coefficient may remain much closer to a coverage of $1-\delta$ for most of the possible parameter values. Surely, as supporters of this view say, the second type of estimator should be preferable, especially since the use of a frequentist probability over a sample space, when only one sample is available, is a rather arbitrary concept for evaluating an interval estimator. Proponents of the second view suggest other measures of coverage quality, such as the mean squared deviation between the coverage and the nominal confidence level[24], or the maximum deviation in coverage from the nominal confidence level.[25] We shall not go into detail here, but note that representative statements of both groups can be found in the discussion of Brown et al. (2001).

Troubling issues in estimators exhibiting undercoverage are: gross undercoverage, where the coverage for certain parameter values are far below the nominal confidence level, especially when the coverage approaches zero for some parameter values; and consistent undercoverage, where an estimator's coverage is consistently below the nominal confidence level for an extensive continuous range of a parameter value. The first issue is self-explanatory. The second issue is less clearly defined: in the binomial proportion case,

[24] A generalization of this idea is a weighted mean squared deviation based on a Bayesian prior over the parameter space.

[25] A related idea is the use of Bayesian credible intervals instead of classical confidence intervals.

suppose an estimator exhibits only coverage of 0.85 when the nominal confidence level is 95%, for all $p \leq 0.05$ and $p \geq 0.95$, and has adequate coverage for all other p. This is not likely to be an issue for a problem seeking an estimator for the proportion of males in a typical mammal population. However, for the problem of estimating the proportion of defects in a manufacturing process, this interval estimator is unlikely to be suitable. (Note that this is effectively the same consideration which should guide one's hand in choosing a point estimator.) Thus, the nature of the undercoverage, and the practitioner's expectations of a problem are factors in the acceptance of estimators exhibiting undercoverage.[26]

Two important concepts for the study of coverage of intervals employing distributional assumptions are n-th order correctness and n-th order accuracy (Efron and Tibshirani, 1993, Section 22.2). An interval endpoint estimator is said to be n-th order correct if the asymptotic deviation of the endpoint estimator and the ideal endpoint estimator (based on the exact distribution of the statistic used as the basis for the interval) is $O_p(k^{\frac{-n-1}{2}})$. n-th order correctness is usually studied by employing Edgeworth expansions. It is generally expected that a confidence interval be first-order correct, and preferably second-order correct. Obtaining third-order correctness (or more) is almost always impractical. n-th order accuracy is a weaker condition than n-th order correctness, which considers the deviation in coverage of an interval from the nominal coverage, regardless of the relationship of the endpoint estimators to the ideal endpoint estimators. An example of these concepts being employed to study bootstrap confidence intervals is Hall (1988). Due to our focus on smaller sample sizes, and the asymptotic nature of these concepts, we will not be considering correctness and accuracy as factors in what follows.

[26] However, estimators exhibiting overcoverage may also be unsuitable if they lead to over-conservatism and/or financial losses.

Length

Clearly, for a given confidence level, the shorter of two interval estimators is preferable.

Symmetry

This refers to estimators which are centred around some point estimator of the parameter under investigation. In the one-sided case, this property obviously becomes irrelevant. Symmetry is desirable for an interval estimator if the underlying statistic has a symmetric distribution, otherwise it is not. In the case of a binomial proportion, the statistic $\hat{p}$ generally has a non-symmetric distribution with mean p and variance $\frac{p(1-p)}{k}$. In this case, it is not desirable to obtain a symmetric interval estimator.

Equal tails

This is another property which is not relevant in the one-sided scenario. It refers to an estimator where the probability that the parameter falls below the lower bound of the interval is (sometimes approximately) equal to the probability that the parameter exceeds the upper bound.

There is sometimes an interesting interplay between this requirement and that of length, when a shorter interval can be obtained by violating this condition.

It is customary to either set the tail probabilities equal, or to construct one-sided intervals. However, there is generally no reason why this must be the case, and intervals for arbitrary upper and lower tail probabilities can generally be constructed just as easily[27].

Example 3.1. To illustrate, consider the case where it is desired that the lower tail probability be 0.025 and the upper be 0.075. Then, given a method for constructing an equal-tailed confidence interval, an appropriate interval

[27] At least in any case where two-sided intervals can be constructed.

will be $[\mathscr{L}_{0.95}, \mathscr{U}_{0.85}]$, where $[\mathscr{L}_{0.95}, \mathscr{U}_{0.95}]$ and $[\mathscr{L}_{0.85}, \mathscr{U}_{0.85}]$ are 95% and 85% equal-tailed intervals respectively. □

Bayesian highest posterior density (HPD) credible intervals are perhaps the most well-known form of interval estimators which do not generally have equal tails.

Equivariance

This section is related to Section 3.1.5.

Equivariance refers to a kind of desired symmetry in the construction of the estimator. In the binomial proportion case, equivariance means that a $100(1-\delta)\%$ interval estimator for p constructed on the basis of r "successes", will also generate a $100(1-\delta)\%$ interval estimator for $1-p$ if $k-r$ is substituted for r. Informally, one could say that an interval is equivariant in this sense if the interval does not depend on which of two outcomes is considered a "success" when taking a sample or performing an experiment.

Equivariance with respect to other transformations can also be desirable. Once again, the inherent asymmetry of one-sided intervals typically makes this property irrelevant in that case.

Monotonicity

Generally, monotonicity means that it is undesirable for the interval endpoints to move in the opposite direction to the statistic they are based on.

Consider again the case of interval estimation for the true error. First, monotonicity in the number of successes means than an increase in the number of successes for a fixed sample size (and hence in $\hat{p}$) should not lead to a decrease in either endpoint of the estimator. Second, monotonicity in the sample size means that an increase in sample size while the number of successes remains constant (hence a decrease in $\hat{p}$) should not lead to an increase in either endpoint of the estimator.

Clearly, this property is still relevant in the one-sided case.

Summary

Our main criteria for evaluating interval estimators are coverage, length, and monotonicity. We shall disregard equality of tails: if they are desired, they can be obtained from any two-sided interval. Symmetry will not be considered a benefit for interval estimators of error, since it is generally unusual for the statistics we consider to have a symmetric distribution. Equivariance with respect to $\phi(v) = 1 - v$ is desirable for error estimation, while for general loss functions we shall disregard equivariance.

We briefly mention three other considerations for interval estimators.

- *Overshooting* refers to the situation where the interval estimator contains a value which it is impossible for the estimand to attain. A classic example is an interval estimator for variance with a negative endpoint. Since the true risk in our case is assumed to lie in $[0, 1]$, any interval generating endpoints outside this interval can be improved by trimming the endpoints.
- *Degenerate* or (zero-width) intervals occasionally occur. This may not seem very useful, but theoretically these degenerate intervals are perfectly valid.
- The assumptions underlying an interval estimator are, of course, very important. The most common assumption used in constructing common confidence intervals are distributional assumptions, typically regarding asymptotic normality. The validity of these assumptions should be considered.

3.2.3 Employing the inverse distribution deviation

Consider an hypothesis test for the simple null hypothesis $H_0 : r_D(w) = t_0$, relative to the composite alternative hypothesis $H_a : r_D(w) < t_0$. Suppose the hypothesis test employs a statistic V such that $(V|r_D(w) = t)$ has a distribution Q_t with c.d.f. F_{Q_t}. Generally, the realization v of V on the sample is compared to a threshold value s_δ for the test. This threshold value is typically the $100(1-\delta)$-th quantile of Q_{t_0}. It is common that the null hypothesis is rejected if $v > s_\delta$. One can obtain a confidence set by inverting this hypothesis test. That is, if the test would not reject the null hypothesis $r_D(w) = t$ for the realized statistic, t is included in the confidence set, since it can be considered "reasonable" in such a sense. The confidence interval thus consists of all values of t for which $\mathbb{P}_{V \sim Q_t}\{V \leq v\} \leq 1-\delta$. But this expression is the inverse distribution deviation, so we can rewrite this expression simply as

$$\psi(t, v) \leq 1 - \delta \ ,$$

where v is the realization of V for the sample T.

In practice, Q_t is generally not known. When the loss function is zero-one, Q_t is related to a binomial distribution, and we will focus on that case.

3.2.4 Employing the binomial tail deviation

The upper and lower binomial tail deviations are applications of the inverse distribution deviation, where Q_t is a binomial distribution. Results employing the binomial tail deviation follow from inverting the binomial test. For a zero-one loss function, the number of errors has a binomial distribution with parameters k and $e_D(w)$, so this method can be employed to obtain a confidence set for $e_D(w)$.

The resulting interval is known as the Clopper-Pearson interval or the max-P interval (Clopper and Pearson, 1934, Vollset, 1993). Since the result is derived from the exact distribution of the hypothesis statistic $e_T(w)$, the interval is sometimes called *exact* (as opposed to approximate/asymptotic).

We use the following reasoning to construct an upper $100(1-\delta)\%$ confidence interval: the binomial test employs $r = ke_T(w)$ as a statistic, and compares it to s_δ, the $100(1-\delta)$-th quantile of the binomial distribution with parameters k and t_0, where t_0 is the null hypothesis probability of error. Thus, if

ψ^U denotes the upper binomial tail deviation, the corresponding confidence interval consists of the points t satisfying

$$\psi^U(kt, r) \leq 1 - \delta \ ,$$

namely $[0, \mathrm{UBT}(e_T(w), k, 1-\delta)]$.

For a given δ, this interval has a guaranteed coverage of $1-\delta$, i.e. the probability that $e_T(w)$ lies outside the interval is strictly less than δ.

The binomial test is theoretically well-motivated: if the c.d.f. of the statistic V used in an hypothesis test has a monotone likelihood ratio (LR), the Neyman-Pearson lemma implies that the test is a uniformly most powerful (UMP) test at the specified level, if the critical function is of the form $I(V > s_\delta)$.

It is known (see, for example, Dudewicz and Mishra, 1988, Theorem 9.3.70) that the c.d.f. of any distribution in the one-parameter exponential family has a monotone LR in V, where V is the coefficient of the natural parameter of the distribution. This family includes the binomial, Poisson, normal, and one-parameter Gamma and Beta distributions[28]. For the binomial distribution in particular and our one-sided alternative hypothesis, this means that the number of errors r employed above is an appropriate statistic.

Following this reasoning to its conclusion yields in essence the binomial test described above. The complication is that a UMP test needs to have level exactly δ, but the discrete nature of the binomial distribution means that this cannot be done. Theoretically, therefore, there may be a uniformly better test than the binomial test above, with corresponding improved confidence intervals. However, it is more likely that a number of alternative tests exist, with improved power for only a restricted set of parameter values. Inverting such tests would generally yield confidence sets which are only improvements for certain parameter values. However, due to its relationship to an "almost" UMP test, the Clopper-Pearson interval is generally considered the "gold standard" for binomial confidence intervals.

[28] These one-parameter distributions are special cases of the regular distributions where the two parameters are equal.

Next, we consider three popular general hypothesis tests and the intervals they produce: the LR test, the Rao score test, and the Wald test.

3.2.5 The likelihood ratio test

Along with the Wald interval and the score interval, the LR method for interval estimation is very popular (Brown et al., 2001). This method is based on inverting the LR test.

In our case of the simple null hypothesis $H_0 : r_D(w) = t_0$, and the composite alternative hypothesis $H_a : r_D(w) < t_0$, the test statistic is $-2 \ln \mathit{\Lambda}(T)$, where

$$\mathit{\Lambda}(T) = \frac{\mathrm{Lik}(t_0|T)}{\sup_{t<t_0} \mathrm{Lik}(t|T)}$$

is the LR.

Clearly we cannot directly apply this test for general loss functions, as we do not have a parametric form for the distribution of $L(w(x), y)$ when $(x, y) \sim D$. Once again, we consider the zero-one loss function, where that distribution is known to be Bernoulli with parameter $e_D(w)$.

The resulting likelihood is

$$\mathrm{Lik}(t|T) = t^r(1-t)^{k-r} \quad , \tag{3.6}$$

so that the LR is

$$\mathit{\Lambda}(T) = \frac{{t_0}^r(1-t_0)^{k-r}}{\sup_{t<t_0}(t^r(1-t)^{k-r})} \quad .$$

The supremum in the denominator occurs at $\hat{p}$ if $\hat{p} < t_0$, else as $t \to t_0$. The resulting statistic for $\hat{p} \geq t_0$ is

$$\begin{aligned} V(T) &= -2 \ln \mathit{\Lambda}(T) \\ &= -2 \ln \left(\frac{{t_0}^r(1-t_0)^{k-r}}{{t_0}^r(1-t_0)^{k-r}} \right) \\ &= 0 \quad , \end{aligned}$$

so that no $t_0 \leq \hat{p}$ will be rejected. Assuming $\hat{p} < t_0$,

$$V(T) = -2[r(\ln t_0 - \ln \hat{p}) + (k-r)(\ln(1-t_0) - \ln(1-\hat{p}))] \quad .$$

We see that $V(T)$, given t_0, depends on a k-sample T only through r, so that we can equivalently write $V(r)$. Thus, the distribution of $V(T)$ given t_0 can easily be obtained. The LR test then proceeeds by comparing $V(T)$ to an appropriate quantile of the distribution of $V(T)$. The $100(1-\delta)\%$ confidence set obtained by inverting this test is then

$$[0,\hat{p}] \cup \left\{ t : r\ln t + (k-r)\ln(1-t) \geq r\ln\hat{p} + (k-r)\ln(1-\hat{p}) - \frac{1}{2}s_\delta(t) \right\} \ ,$$

where $s_\delta(t)$ denotes the $100(1-\delta)$-th quantile of $V(T)$ given t.

Obtaining the exact distribution of the statistic $-2\ln\Lambda$ for general LR tests is generally very difficult. Traditionally, applications of the LR test employ an asymptotic approximation to the distribution of $-2\ln\Lambda$. In our case of testing a single parameter, $-2\ln\Lambda$ asymptotically has a central χ_1^2 distribution. This asymptotic approximation is so prevalent that it is assumed as the *de facto* standard when speaking of a LR test. The test using the exact distribution of $-2\ln\Lambda$ is then typically called an exact LR test.

The hypothesis test using the asymptotic distribution is performed by comparing the test statistic to an appropriate quantile of the χ_1^2 distribution. The resulting upper interval with confidence $1-\delta$ is then described by

$$[0,\hat{p}]\cup\left\{ t : r\ln t + (k-r)\ln(1-t) \geq r\ln\hat{p} + (k-r)\ln(1-\hat{p}) - \frac{1}{2}\chi_1^2(1-\delta) \right\} \ ,$$

where $\chi_1^2(1-\delta)$ denotes the $100(1-\delta)$-th quantile of the χ_1^2 distribution. The solution to this equation is generally found numerically, using a line search technique.

We next turn our attention to a closely related test, the Rao score test.

3.2.6 The score interval

The Rao score test is based on the fact that, for a small value of c,

$$\ln\frac{\mathrm{Lik}(t|T)}{\mathrm{Lik}(t-c|T)} = \ln\mathrm{Lik}(t|T) - \ln\mathrm{Lik}(t-c|T)$$

$$\approx \ln\mathrm{Lik}(t|T) - (\ln\mathrm{Lik}(t|T) - c\frac{\partial}{\partial t}\ln\mathrm{Lik}(t|T))$$

$$= c\frac{\partial}{\partial t}\ln \mathrm{Lik}(t|T)) \ ,$$

by means of a first-order Taylor expansion. The test is then constructed using $V(T) = \frac{\partial}{\partial t}\ln \mathrm{Lik}(t|T)|_{t=t_0}$, the score at $t = t_0$, as the statistic. Since the statistic is based on an approximation to a ratio of log-likelihoods, it is not surprising that the intervals obtained by inverting the score test are generally very similar to those obtained from the LR test.

Once again, we note that this cannot be applied to general loss functions, but that it can be done for zero-one loss functions.

By taking the logarithm of (3.6), one obtains the log-likelihood

$$r\ln t + (k-r)\ln(1-t) \ .$$

Taking the partial derivative and evaluating at t_0 yields

$$V(T) = \frac{\partial}{\partial t}\ln \mathrm{Lik}(t|T))|_{t=t_0} = \frac{r}{t_0} - \frac{k-r}{1-t_0} \ .$$

A similar argument to that in the previous section allows the derivation of an exact score interval, employing quantiles of the exact distribution of $V(T)$.

For an i.i.d. k-sample, the likelihood can be expanded as a k-product; hence the log-likelihood can be expanded as a k-sum. The result is that the central limit theorem can be applied to conclude that the asymptotic distribution of a log-likelihood is normal. Since the log-LR here is simply the difference between two log-likelihoods, it also has an asymptotic normal distribution. Non-exact (i.e. standard) score tests thus use a normal distribution to approximate the distribution of $V(T)$.

Given $t = t_0$, by rewriting $V(T)$ as $\frac{r-kt_0}{t_0(1-t_0)}$, we note that $\mathbb{E}_{T\sim D^k} V(T) = 0$ and $\mathbb{V}_{T\sim D^k} V(T) = \frac{kt_0(1-t_0)}{{t_0}^2(1-t_0)^2} = \frac{k}{t_0(1-t_0)}$. (Note that $\mathbb{V}_{T\sim D^k} V(T)$ is the Fisher information given $t = t_0$.) In this case, the score test rejects the null hypothesis if $V(T)$ is less than the 100δ-th quantile of the normal distribution with mean 0 and variance $\frac{k}{t_0(1-t_0)}$, or equivalently, if

$$\sqrt{\frac{t_0(1-t_0)}{k}}\left[\frac{r-kt_0}{t_0(1-t_0)}\right] < -z_\delta \ ,$$

where z_δ denotes the level-δ critical value of the standard normal distribution. Rewriting this condition, we obtain that the t satisfying

$$\frac{t-\hat{p}}{\sqrt{t(1-t)}} \leq \frac{z_\delta}{\sqrt{k}}$$

are those which will not be rejected by the test, and thus form the score interval. Note that this is of the form

$$\psi(t,\hat{p}) \leq \frac{z_\delta}{\sqrt{k}}$$

where ψ is the upper Rao deviation.

The resulting score interval (see (3.5)), also known as Wilson's score interval (Wilson, 1927), is

$$\left(0, \frac{\hat{p} + \frac{{z_\delta}^2}{2k} + \frac{z_\delta}{\sqrt{k}}\sqrt{\hat{p}(1-\hat{p}) + \frac{{z_\delta}^2}{4k}}}{1+\frac{{z_\delta}^2}{k}}\right] ,$$

or the more popular equivalent formulation

$$\left(0, \frac{r + \frac{{z_\delta}^2}{2} + z_\delta\sqrt{r - \frac{r^2}{k} + \frac{{z_\delta}^2}{4}}}{k+{z_\delta}^2}\right] .$$

Due to the symmetry of the normal distribution, the corresponding lower interval is of the same form, with upper bound 1 and lower bound the same form as the upper bound, except that the term containing the square root is subtracted. As a result, the centre of a two-sided score interval is $\frac{r+\frac{{z_\delta}^2}{2}}{k+{z_\delta}^2}$. This expression has a clear interpretation in terms of shrinkage: it is a weighted average of $\hat{p}$ and $\frac{1}{2}$, with the weights dependent on k and the critical value $z_{\frac{\delta}{2}}$ (and hence the required confidence):

$$\frac{r+\frac{1}{2}z^2_{\frac{\delta}{2}}}{k+z^2_{\frac{\delta}{2}}} = \hat{p}\left(\frac{k}{k+z^2_{\frac{\delta}{2}}}\right) + \frac{1}{2}\left(\frac{z^2_{\frac{\delta}{2}}}{k+z^2_{\frac{\delta}{2}}}\right) .$$

A similar shrinkage interpretation applies to the width of the interval — for details, see Agresti and Coull (1998). As such, the $1-\delta$ confidence level score interval displays a remarkable similarity to the Bayesian credible interval for the Beta$\left(\frac{1}{2}z^2_{\frac{\delta}{2}}, \frac{1}{2}z^2_{\frac{\delta}{2}}\right)$ prior distribution. Bayesian credible intervals will be discussed later.

The last well known interval we shall discuss is the Wald interval. We shall see that in its standard application, the resulting formulae are much simpler than the score interval, but the resulting interval is also less accurate.

3.2.7 The Wald interval

This section discusses the interval estimators arising by (approximately) inverting the Wald test. The Wald test for a parameter t begins with the statistic

$$V(T) = \frac{\hat{t} - t_0}{\sqrt{\mathbb{V}\,\hat{t}}} \ ,$$

where $\hat{t}$ is the ML estimate of t.

Yet again, we are restricted to particular loss functions, most notably zero-one loss functions. As a result, we focus on obtaining an interval by inverting the hypothesis test for the case of estimating $e_D(w)$. In this case, $\hat{t} = \frac{r}{k}$ is the ML estimate of t, and $\mathbb{V}\,\hat{t} = \frac{t(1-t)}{k}$. Thus

$$V(T) = \sqrt{k}\frac{t - \hat{t}}{\sqrt{t(1-t)}} = \sqrt{k}\psi(t, \hat{t}) \ ,$$

where ψ denotes the upper Rao deviation. We note that this is the same as the test statistic for the (regular) score interval. Inverting this test directly is possible in this case because $\mathbb{V}\,\hat{t}$ is a simple function of t. For general parameters, however, this is not the case, and the Wald test would have to be inverted numerically, even if normality of the test statistic is assumed. We shall refer to the test outlined here as the *exact-variance* Wald test, and the resulting intervals as either an exact exact-variance Wald interval, or a (standard) exact-variance Wald interval, to distinguish it from the (regular) Wald interval which follows.

To simplify the inversion of the test for general parameters, it is customary to assume that $\mathbb{V}\,\hat{t}$ is constant for the hypothesis test, regardless of the choice of t, and that the appropriate constant is $\mathbb{V}\,\hat{t} = \mathbb{V}\,\hat{t}|_{t=\hat{t}}$, which for our purposes is $\frac{\hat{t}(1-\hat{t})}{k}$. Inverting the resulting test is much easier, since the test statistic

$$V'(t) = \sqrt{k}\frac{t-\hat{t}}{\sqrt{\hat{t}(1-\hat{t})}} = \sqrt{k}\psi(t,\hat{t})$$

(where ψ is the upper Wald deviation) is merely a linear transformation of t.

The exact (regular) Wald interval can be derived by considering the exact distribution of $V'(T)$. If normality of the test statistic is assumed, as is customary, the (regular) Wald interval consists of those t satisfying

$$\psi(t,\hat{t}) \leq \frac{z_\delta}{\sqrt{k}} \quad .$$

Inverting ψ, one obtains the interval

$$\left[0, \hat{t} + z_\delta\sqrt{\frac{\hat{t}(1-\hat{t})}{k}}\right] \quad .$$

3.2.8 Discussion

In our discussion of the LR test, the score test, and the Wald test, we have outlined four types of potential intervals: the LR interval, the score interval, the exact-variance Wald interval, and the (regular) Wald interval. Each interval can be derived in a one-sided manner to obtain upper or lower bounds, which can be combined to obtain two-sided intervals. Furthermore each interval can be obtained based on the exact distribution of the test statistic — the exact test/interval; or as is more customary, assuming that an asymptotic distribution result is accurate — the (standard) test/interval. All the exact intervals have guaranteed coverage of at least $1-\delta$, by their construction, while the standard intervals may potentially exhibit undercoverage.

We have seen that it is not generally possible to calculate these intervals for an arbitrary loss function, since the distribution of the loss, given the true risk, is unknown. The important class of zero-one loss functions allows calculation of these intervals, though.

For all the upper one-sided tests we have outlined here, we find that the statistic $V(T)$ is increasing for zero-one loss functions. It follows that the sets obtained by inverting these tests are upper intervals. Similar results hold for lower and two-sided intervals. Now, suppose some hypothesis test yields a shorter upper interval than the binomial interval when inverted, for some confidence level $1-\delta$. Specifically the new test yields $[0, t_1]$ while the binomial test yields $[0, t_2]$ with $t_2 > t_1$. Since the derivation of t_2 implies

$$t_2 = \inf \left\{t \in [0,1] : \mathbb{P}_{T\sim D^k}\left\{e_T(w) \leq t\right\} \geq 1-\delta\right\} \;\; ,$$

it follows that

$$\mathbb{P}_{T\sim D^k}\left\{e_T(w) \leq t_1\right\} < 1-\delta \;\; ,$$

i.e. the alternative hypothesis test exhibits undercoverage. Since inversions of exact tests can not undercover, any exact interval is no better than the binomial interval. A similar result holds for lower intervals. Furthermore, the intervals for all the tests outlined here turn out to be the same as the binomial test in these cases. *This is not so for two-sided intervals* — we discuss this issue further in Section 3.2.10.

The confidence intervals arising from the inversion of the (regular) tests all rely on distributional assumptions, and hence the quality of the intervals relies on how accurate the approximation is to the exact distribution. Important in this regard is that when using a normal distribution to approximate a discrete distribution, a continuity correction (CC) is necessary. Applying CCs to the tests above is discussed in Section 3.2.9.

For very large n, it would appear we have nothing to worry about, since the central limit theorem would appear to take care of everything (and the effect of the continuity correction would be negligible). However, the central limit theorem does not guarantee uniform convergence for all values of $p \in [0,1]$: in other words, for any n, there are (extreme) values of p for which the normal approximation is not good. This partially explains the poor performance of the approximate tests for values of p near 0 and 1. The regular guidelines given for the required sample size to be sufficient for the normal approximation for the Wald test to be adequate are generally too liberal (Brown et al., 2001). The result is that Wald intervals are consistently too narrow, as has been repeatedly noted by various authors. The development of this discussion can be found in Blyth and Still (1983), Vollset (1993), Agresti and Coull (1998), Brown et al. (2001), and Brown et al. (2002), amongst others. These authors consistently note the superiority of the score interval over the Wald interval for the binomial case, with Blyth and Still (1983, Section 3) showing that the approximation of the variance of $\hat{p}$ by a constant makes the (CC) Wald interval consistently too narrow.

In addition, the (CC) Wald interval may overshoot and yields zero-width intervals when $\hat{p}$ is zero or one. The CC score interval does not have these problems, and exhibits equivariance and monotonicity, among several desirable features (see the discussion in Blyth and Still, 1983, Section 2).

3.2.9 Continuity corrections

The lack of a CC in the development of the (regular) tests above leads to undercoverage: when a $\text{Bin}(k,p)$ variable is approximated by a $N(kp, kp(1-p))$ variable, we note that we are approximating

$$\mathbb{P}_{V \sim \text{Bin}(k,p)}\{V = v\}$$

by

$$\mathbb{P}_{V \sim N(kp,kp(1-p))}\left\{v - \frac{1}{2} < V < v + \frac{1}{2}\right\} \quad .$$

It follows that comparing the statistic to a critical exact value should be equated to comparing the statistic to a *modified* approximate value. Applying such a modification to the tests leads to CC intervals, which remove such consistent undercoverage.

Furthermore, Hall (1982) points out that a correction for skewness is also necessary when approximating the asymmetric binomial distribution with the symmetric normal distribution. For two-sided intervals, this correction is of a lower order than the continuity correction, but when one is constructing one-sided intervals, this correction can have an even larger impact than the traditional continuity correction.

Wald interval

The most accepted form of the CC Wald interval, which we shall employ, is

$$\mathrm{Conf}_{1-\delta}(p) = \left[\hat{p} \pm \left(z_{\frac{\delta}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{k}} + \frac{1}{2k} \right)\right] \quad .$$

As noted in Section 3.2.8, even the CC Wald interval is too narrow. Blyth and Still (1983) provide an improved alternative, which we name the *Blyth-Still-Wald (BSW) interval*

$$\left[\hat{p} \pm \left(\frac{z_{\frac{\delta}{2}}}{\sqrt{k - z^2_{\frac{\delta}{2}} - \frac{2z_{\frac{\delta}{2}}}{\sqrt{k}} - \frac{1}{k}}} \sqrt{\hat{p}(1-\hat{p})} + \frac{1}{2k} \right)\right] \quad .$$

Note that this interval multiplies the variance estimate by a factor. This attempts to undo the damage done by using a constant estimate of the variance, which tends to be an underestimate.

Score interval

The score interval also needs a continuity correction, leading to the CC score interval

$$\mathrm{Conf}_{1-\delta}(p) = \left[\frac{(r \pm 0.5) + \frac{1}{2} z^2_{\frac{\delta}{2}} \pm z_{\frac{\delta}{2}} \sqrt{(r \pm 0.5) - \frac{(r \pm 0.5)^2}{k} + \frac{1}{4} z^2_{\frac{\delta}{2}}}}{k + z^2_{\frac{\delta}{2}}} \right]$$

This modification addresses most of the score interval's undercoverage issues, and is considered far superior to the Wald intervals: it has been popular amongst practitioners for many years (Blyth and Still, 1983). Furthermore, Blyth and Still (1983) recommended the use of the CC score interval for $k > 30$ due to its desirable properties (probably due to the limitations of computers at the time), with just one modification. They suggest using appropriate one-sided binomial tail intervals when $\hat{p} = 0$ or $\hat{p} = 1$ to address undercoverage at the endpoints. We call the resulting interval the Blyth-Still (BS) score interval.

Despite this modification, however, the interval still does not provide guaranteed coverage of $1 - \delta$ for all values of p. On the other hand, in the 95% case, Vollset (1993) report that the coverage is about 94% for all p, with coverage below the nominal 95% level only occuring for p within $\frac{1}{10k}$ of 0 or 1.

3.2.10 Improvements to the two-sided binomial interval

We mentioned above that the binomial interval yields shortest possible one-sided confidence intervals for binomial proportions. It follows that the same holds for two-sided intervals where the minimum tail probabilities for each tail are individually specified. However, two-sided intervals resulting from inverting hypothesis tests do not necessarily have equal tail probabilities. Thus, a better two-sided interval than a simple Bonferroni combination of equal-tailed binomials may be possible.

The two-sided exact LR interval

The question thus arises whether it is possible that the exact intervals from the hypothesis tests we have already considered may be improvements on the binomial interval. It turns out that two-sided exact versions of the Wald, exact-variance Wald, and score tests, are basically a combination of two one-sided tests, so that the resulting two-sided intervals are identical to the two-sided binomial interval. However, the two-sided LR test uses a single comparison to reject the null hypothesis for both too-large and too-small error rates. The result is a test for which the tail probabilities need not be nearly equal.

It turns out that the two-sided exact LR interval is uniformly better than the two-sided binomial interval (Brown et al., 2001, Comments by Corcoran and Mehta). In this regard, we note that the two-sided binomial interval's overcoverage is worst when $p < 0.1$ or $p > 0.9$ (exceeding 98% for large portions of these ranges for a 95% interval), and this overcoverage is heavily reduced by the two-sided exact LR interval.

The Blyth-Still-Casella interval

Improving the Clopper-Pearson interval was also the focus of the work in Blyth and Still (1983): as an alternative, the authors suggested a confidence interval based on Edwin Crow's refinement (Crow, 1956) of Sterne's method (Sterne, 1954) for constructing hypothesis tests. Sterne's method constructs acceptance regions by using a method analagous to Bayesian HPD intervals. Crow's correction ensures that inverting the modified tests yield confidence regions which are intervals.

It is shown that the suggestion of Blyth and Still also guarantees coverage of $1 - \delta$, while generally being of a slightly shorter length. Although this interval is more difficult to understand and implement, implementations should be more widely used in practice as an improvement on the two-sided Clopper-Pearson interval.

In 1986, (Casella, 1986) proposed a method to enhance any equivariant procedure for generating confidence intervals for a binomial parameter. This method can be seen as a generalization of the proposal of Blyth and Still to improve the Clopper-Pearson interval, in the sense that the BS intervals are obtained as a special case of Casella's technique. In addition, Casella's method easily yields confidence intervals for arbitrary confidence levels, while Blyth and Still only performed their construction for 95% and 99% confidence levels (and used the score interval for $k > 30$ as an approximation). Thus, the interval of the BS type obtained by Casella's method is known as the Blyth-Still-Casella (BSC) interval.

3.2.11 Randomized hypothesis tests

Up to now, all the hypothesis tests we have considered have been deterministic; i.e. for a specific value of the test statistic, the decision whether to accept or reject the null hypothesis is always the same. This is the cause of the overcoverage in the case of hypothesis tests for the parameter of the Bernoulli distribution: the overcoverage of exact intervals derives from the fact that the values on the boundary of the acceptance region for the test are (by definition) never used to reject the hypothesis $p = p_0$. However, these boundary values have a non-zero probability of occurring[29], resulting in overcoverage.

One solution to this dilemma is to consider *randomized hypothesis tests*: if the realization of the statistic lies on the boundary of the acceptance region, it is only accepted with a certain probability. Specifically, if v is on the boundary of the acceptance region $R(t_0)$ of an hypothesis test, and we have

$$\mathbb{P}\{V(T) \in R(t_0)\} = 1 - \delta_1 \geq 1 - \delta$$

and

$$\mathbb{P}\{V(T) \in R(t_0) \setminus \{v\}\} = 1 - \delta_2 < 1 - \delta \ ,$$

we wish to reject the hypothesis on realization of $V(T) = v$ with probability

$$\frac{(1-\delta_1) - (1-\delta)}{(1-\delta_1) - (1-\delta_2)} = \frac{\delta - \delta_1}{\delta_2 - \delta_1} \ .$$

The resulting test then has level exactly δ.

Randomized tests present one with a new challenge: the inversion of the randomized test. If one inverts them in the same manner as the deterministic tests discussed earlier, including the boundary values in the acceptance region, one obtains the same intervals as for a corresponding deterministic test, and nothing is gained. However, if in the inversion, at any stage one excludes a boundary value from the rejection region, the resulting confidence interval will sometimes exhibit undercoverage. However, the undercoverage will never exceed the maximum point probability of $V(T)$. Thus the coverage of such an interval is asymptotically equal to the required coverage. A similar argument shows that the overcoverage will also become negligible asymptotically.

The mid-P interval

This serves as a motivation for the so-called mid-P interval, which is obtained from a randomized version of the binomial test. For the purpose of inverting

[29] This is the general problem with inference based on discrete variables

the test, a point on the boundary of the acceptance region is considered included in the acceptance region when the probability of rejecting the point is less than $c = 0.5$. This approach leads to an interval estimator. Using other constants also leads to interval estimators — notably $c = 1$ leads to the max-P interval, and $c = 0$ yields an interval with consistent undercoverage (which we could call the min-P interval).[30]

Mid-P intervals are, as expected, generally slightly shorter than the max-P interval, with substantially less overcoverage, but which exhibits undercoverage.

In principle, one could construct a randomized version of any hypothesis test, and invert it using some constant c. If the original test generated interval estimators, so will this process. We shall call such an interval a c-randomized interval. As an example, the mid-P interval could be called a $\frac{1}{2}$-randomized Clopper-Pearson (or binomial) interval. Specifically, one may be interested in constructing $\frac{1}{2}$-randomized BSC and exact LR intervals.

3.2.12 Approximations to the binomial intervals

Pratt's approximation

Pratt's approximation (Vollset, 1993) is a closed form approximation to the max-P interval: the upper limit of the two-sided Pratt interval is[31]

$$\mathscr{U}_P(r,\delta) = \left[1 + \left(\frac{r+1}{k-r}\right)^2 (\phi(r,k,\delta))^3\right]^{-1} \quad ,$$

where $\phi(r,k,\delta)$ equals

$$\frac{81(r+1)(k-r) - 9k - 8 - 3z_{\frac{\delta}{2}}\sqrt{9(r+1)(k-r)\left(9k+5-z_{\frac{\delta}{2}}^2\right) + k + 1}}{81(r+1)^2 - 9(r+1)\left(2 + z_{\frac{\delta}{2}}^2\right) + 1} \quad .$$

[30] Other methods for inverting the test could be considered, but they may not necessarily yield interval estimators.

[31] The subscript P here refers to Pratt.

The corresponding lower limit is derived from the same formula, as $\mathscr{L}_P(r,\delta) = 1 - \mathscr{U}_P(k-r,\delta)$. A one-sided interval with confidence level $1-\delta$ can be obtained as $[0, \mathscr{U}_P(r, 2\delta)]$.

A modification of Pratt's max-P approximation for the mid-P interval was proposed in Vollset (1993), and essentially consists of linear interpolation between two max-P approximations: the upper bound is

$$\frac{\mathscr{U}_P(r,\delta) + \mathscr{U}_P(r+1,\delta)}{2}$$

and the lower bound is

$$\frac{\mathscr{L}_P(r,\delta) + \mathscr{L}_P(r-1,\delta)}{2} \quad .$$

The realizable case

When $\hat{p} = 0$, we can derive a rather strong bound, since the probability that $\hat{p} = 0$ is simply $(1-p)^k$. Hence, an unlikely p, given zero empirical error, will be one such that $(1-p)^k \leq \delta$. This yields the one-sided (exact) realizable interval

$$\mathrm{Conf}_{1-\delta}(p) = \left[0, 1 - \sqrt[k]{\delta}\right] \quad .$$

It is common to use the bound $(1-p)^k \leq e^{-kp}$, since the resulting interval will be more easily applicable in conjunction with other bounds we shall encounter later. The resulting one-sided exponential realizable interval is

$$\mathrm{Conf}_{1-\delta}(p) = \left[0, \frac{1}{k} \ln \frac{1}{\delta}\right] \quad . \tag{3.7}$$

The motivation for naming these realizable intervals will become apparent when we discuss training sample interval estimators.

3.2.13 Confidence intervals on transformations

A number of techniques involve deriving confidence intervals for a function of p, and then inverting the function on the endpoints of the interval. This method can in principle be used in conjunction with any method for generating confidence intervals. Since this can not improve intervals based on exact hypothesis tests, the techniques are generally only suitable for intervals based on regular hypothesis tests. Here we outline some of the best-known such transformations for estimating a binomial proportion.

This technique has two major benefits. When applying bounds such as the Wald interval, which may overshoot, obtaining bounds on a transformation of the parameter may be beneficial, when the range of the inverse transformation is the valid range of values for the parameter. This is the case for all the transformations we shall consider, so that no intervals obtained based on these transformations can overshoot. Another use of such transformations is that it can simply be easier to obtain intervals for the transformed value: sometimes inverting the hypothesis test for the original parameter is difficult, but much easier for the transformed parameter. We shall see an example of this with the arcsine transformation.

In general, we will refer to an interval constructed by employing a transformation by adding the name of the transformation to the name of the interval. Thus one may speak of an arcsine Wald interval. If the hypothesis test is not specified, it is assumed to be a Wald interval.

The key to obtaining such intervals is being able to find the variance of the transformed empirical error rate. This is done by employing the so-called *delta method*, which employs a first-order Taylor expansion of the transformation ϕ about the true risk — see, for example, Bishop et al. (1975, Section 14.4).

The logit transform

Perhaps the most well-known such transformation is the logit transform of p (Vollset, 1993). The logit transform is a mapping $\phi : [0, 1] \to \mathbb{R}$, where $\phi(p) = \ln(\frac{p}{1-p})$. Suppose an interval $[\mathscr{L}, \mathscr{U}]$ for $t(p)$ can be derived from some hypothesis test. This interval can then be converted into an interval for p by the inverse logit function, $\phi^{-1}(v) = 1 - (1 + e^{v})^{-1}$, yielding

$$[\phi^{-1}(\mathscr{L}), \phi^{-1}(\mathscr{U})] = \left[1 - \left[1 + \exp\left(\ln\left(\frac{r}{k-r}\right) \pm \frac{z_{\frac{\delta}{2}}}{\sqrt{r(\frac{k-r}{k})}}\right)\right]^{-1}\right] ,$$

in the case of a logit Wald interval.

The probit transform

A similar approach is based on the probit transform. The probit transform also maps $[0, 1]$ onto $\mathbb{R}$, and is the inverse of the cumulative distribution function of the standard normal distribution. Once again, a confidence interval for the transformed p can be inverted to generate a confidence interval for p itself: for an interval $[\mathscr{L}, \mathscr{U}]$, the inverted interval is

$$\left[\frac{\operatorname{erf}\left(\frac{\mathscr{L}}{\sqrt{2}}\right) + 1}{2}, \frac{\operatorname{erf}\left(\frac{\mathscr{U}}{\sqrt{2}}\right) + 1}{2}\right] .$$

The arcsine transform

Finally, the arcsine transform (Brown et al., 2001, Section 4.2.2) of p, $\arcsin(\sqrt{p})$, is popular because the variance of the transformed $\hat{p}$ is independent of r. As a result, this transformation is sometimes said to be *variance-stabilizing* (Brown et al., 2001). Note that this approach also differs from the previous three since the range of the transformation is $[-\pi, \pi]$ rather than $\mathbb{R}$. The resulting Wald interval for the transformed parameter is

$$[\mathscr{L}, \mathscr{U}] = \left[\arcsin\left(\sqrt{\hat{p}}\right) \pm \frac{z_{\frac{\delta}{2}}}{2\sqrt{k}}\right] ,$$

and the final interval is obtained as $[\sin^2(\mathscr{L}), \sin^2(\mathscr{U})]$.[32]

[32] Two classical drawbacks to intervals based on transformations such as these, are their common lack of symmetry and equivariance. However, these issues are not relevant in the one sided scenario.

3.2.14 Bayesian credible regions

Bayesian methods are based on the posterior distribution of p, given the assumption regarding its prior distribution. The Bayesian approach treats the underlying parameter as a r.v. rather than an unknown constant, so that the resulting regions can be interpreted by a probability statement with respect to the parameter. To distinguish this interpretation from traditional confidence regions, the regions resulting from Bayesian techniques are typically called *Bayesian credible regions*.

While classical statisticians may not agree with the methodology and interpretation of the estimators so obtained, these estimators generally have good performance in practice, and their properties are well worth considering.

A number of approaches to obtaining such regions exist:

- choosing the narrowest interval with sufficient coverage;
- constructing the interval by choosing the points with HPD first;
- choosing the narrowest interval with fixed minimum posterior tail probabilities (typically equal); and
- choosing an interval centred on the posterior mean.

When the HPD method yields an interval, it is also a narrowest interval with sufficient coverage. This always occurs when the posterior distribution is unimodal. Note also that the HPD and narrowest interval methods are in conflict with the requirement of fixed tail probabilities: these methods tend to shorten the interval at the price of making the tail probabilities unequal. In comparison with the classical techniques above, the fixed tail probabilities technique is analogous to the two-sided binomial interval, while the HPD method is similar to the exact LR and BSC intervals. In the one-sided case, only the narrowest one-sided interval option makes sense. This helps explain why the binomial interval is the best in the one-sided case, but not for general two-sided intervals.

As discussed in Section 3.1.3, we shall consider a Beta(α, β) prior distribution for the case of zero-one loss functions. In this case, the posterior has a Beta$(\alpha{+}r, \beta{+}(k{-}r))$ distribution, which is unimodal for $\alpha{+}r, \beta{+}(k{-}r) < 1$.

The most common form of Bayesian credible interval, or posterior probability interval, is the HPD interval (Berger, 1985). This involves choosing those values of p with HPD. If the posterior distribution is not unimodal, this direct approach may not in general yield an interval, so various modifications may be necessary to guarantee an interval.[33]

When the posterior has a Beta distribution, and $\alpha, \beta > 1$, the posterior is unimodal, so the $100(1-\delta)\%$ HPD interval is

$$[B\left(\delta_1; r+\alpha; k-r+\beta\right), B\left(1-\delta_2; r+\alpha; k-r+\beta\right)]$$

where $B(u; \alpha; \beta)$ indicates the $100u$-th percentile of a Beta(α, β) distribution, and we have that $\delta_1 + \delta_2 = \delta$, and that the posterior Beta density is equal at these two percentiles. Such an interval must typically be obtained numerically.

A simpler alternative is the central Bayesian credible interval. This employs equal tails on the posterior, so that the $100(1-\delta)\%$ central interval is

$$\left[B\left(\frac{\delta}{2}; r+\alpha; k-r+\beta\right), B\left(1-\frac{\delta}{2}; r+\alpha; k-r+\beta\right)\right] \quad .$$

Note that for one-sided intervals, the distinctions between these approaches disappear: generally a $100(1-\delta)\%$ lower Bayesian credible interval for p is $[0, B(1-\delta; r+\alpha; k-r+\beta)]$.

Interestingly, the popular HPD interval based on the Jeffreys prior turns out to be almost equivalent to the mid-P interval, and can thus also be seen as a continuity correction to the Clopper-Pearson interval (Brown et al., 2001).

However, this interval has poor coverage properties for very small and very large values of p, because the intervals generated when $r = 0$ or $r = k$ are too narrow. Brown et al. (2001) suggest replacing the intervals for these

[33] Note that this approach has close links to Sterne-Crow intervals (Casella, 1986, Crow, 1956, Sterne, 1954).

two values of r by the Clopper-Pearson values (which do not overcover here), yielding an interval estimator with better properties.

3.2.15 Non-parametric bootstrap confidence intervals

The basic idea of the bootstrap is outlined for point estimation from the training sample in Section 5.1.3. Readers not familiar with the bootstrap would benefit by reading that section (stopping before the paragraph on the .632 estimator) to get a feeling for bootstrap techniques before continuing here. On the other hand, advanced readers familiar with the bootstrap confidence intervals described below may be interested in Peter Hall's comparison of their theoretical properties (Hall, 1988).

One criticism common to the advanced bootstrap confidence intervals is that the computational burden of creating these intervals is so high. In general, two levels of bootstrapping needs to be done. Although some tricks can be used to reduce the total amount of computation needed, bootstrap confidence intervals can still be very time-consuming to calculate.

Intervals from the normal and Student t distributions

The approach used to generate bootstrap confidence intervals is based on an extension of the basic approach used for the well-known interval estimators based on the normal and Student t distributions.

Consider an estimator of p, say $p^\star$, and an estimator of the standard deviation of $p^\star$, $\widehat{\mathrm{se}}$. Then, in essence, the Wald interval is constructed by assuming that the statistic $Z(p^\star, \widehat{\mathrm{se}}) = \frac{p^\star - p}{\widehat{\mathrm{se}}}$ has a standard normal distribution[34]. The interval is then based on the $\frac{\delta}{2}$- and $(1 - \frac{\delta}{2})$-level quantiles of the normal distribution.

In practice, even when $p^\star$ and $\widehat{\mathrm{se}}$ are unbiased estimates of p and the standard deviation of $p^\star$, and $p^\star$ is normally distributed, $Z(p^\star, \widehat{\mathrm{se}})$ still only has a standard normal distribution asymptotically.

[34] For a specific choice of $\widehat{\mathrm{se}}$.

In the above case, $Z(p^\star, \widehat{\mathrm{se}})$ will often have a Student t distribution[35], which for small samples is somewhat different to the normal distribution. Thus, to get a more accurate interval, we should use the $\frac{\delta}{2}$- and $(1 - \frac{\delta}{2})$-level quantiles of the Student t distribution, rather than those of the standard normal distribution.

When $p^\star$ does not have a normal distribution, finding the exact distribution of $Z(p^\star, \widehat{\mathrm{se}})$ is typically not practical. However, it should be clear that if we can accurately obtain the $\frac{\delta}{2}$- and $(1-\frac{\delta}{2})$-level quantiles of that distribution, it will enable us to generate a confidence interval for p. Further note that, due to change of sign involved in the derivation of a confidence interval from a probability statement, the $(1-\frac{\delta}{2})$-level quantile is used for the calculation of the lower confidence limit, while the $\frac{\delta}{2}$-level quantile is used for the upper limit — although for the symmetric normal and t-distributions this usually goes by unnoticed.

Basic bootstrap-t confidence intervals

Consider the real-world procedure: select the $\frac{\delta}{2}$- and $(1-\frac{\delta}{2})$-level quantiles of the distribution of $Z_D(p^\star, \widehat{se})$. In general, we can not perform this directly (not even theoretically, because of the dependence on the unknown distribution D).

In the bootstrap world, calculating the distribution, or the quantiles, of $Z_T(p^\star, \widehat{se})$ is theoretically possible, since there are a finite number of possible bootstrap samples. Practically, one approximates the quantiles by a Monte Carlo approximation: given B bootstrap samples, each with realization $z_{T^{\star b}}(p^\star, \widehat{se})$[36], we naïvely use the corresponding quantiles of the empirical distribution arising from the B values $z_{T^{\star b}}(p^\star, \widehat{se})$ arising from the bootstrap samples.

This approach generates what is known as the bootstrap-t confidence interval (Efron and Tibshirani, 1993). It can be shown that this interval performs

[35] The exact distribution, of course depends on the distribution of $\widehat{\mathrm{se}}$.

[36] It is important to realize that when a bootstrap sample is taken, $p^\star$ and $\widehat{se}$ also need to be recalculated, although the notation does not indicate the dependence.

better than regular intervals based on normality assumptions — one of the major sources of this improvement is that the intervals no longer have to be symmetric about $p^\star$. However, this approach only yields improvements when the Monte Carlo approximation is sufficiently accurate, and this can require a large number of bootstrap iterations.[37] In addition, the technique can be erratic when the test sample T is small, or if it contains outliers. In such cases, choosing a robust estimator $p^\star$ may be useful.

Another problem with the bootstrap-t confidence interval is that it can also overshoot. This is a symptom of the methodology used (i.e. extending the Wald interval approach), but it can be addressed by constructing the confidence intervals on transformations of the data which have a *variance-stabilizing effect* on $p^\star$. Efron and Tibshirani (1993, Section 12.6) discusses a method for finding such transformations automatically using bootstrap methods, which also allows one to construct bootstrap-t confidence intervals using bootstrap estimates of standard error without double bootstrapping (i.e. two nested layers of bootstrapping).

Note that one can also construct confidence intervals by directly estimating the quantiles of $-G(p^\star) = p^\star - p$, resulting in what Carpenter and Bithell (2000) calls the non-Studentized pivotal bootstrap. However, this approach is not recommended, for reasons outlined in the aforementioned article.

Percentile method bootstrap intervals

The motivation for the percentile interval is somewhat more complicated than that of the previous two approaches, and typically this interval does not perform as well as the bootstrap-t interval with variance-stabilizing. On the other hand, it is very simple to compute. In addition this approach yields *transformation-respecting* confidence intervals: if the data is transformed by a monotone transformation, and the percentile method interval is calculated on the transformed data, the resulting interval corresponds to the transformation being applied to the endpoints of the original interval.

[37] Furthermore, when $\widehat{se}$ itself is a bootstrap estimate of the variance of $p^\star$, it is necessary to perform a bootstrap estimate *within* each main bootstrap iteration.

Furthermore, the resulting intervals can not overshoot[38].

Put simply, the percentile method constructs intervals based on the distribution of a centred version of $p^\star$, $G(p^\star) = p - p^\star$, using quantiles obtained by a bootstrap approximation to $G(p^\star)$. The procedure of approximating the ideal bootstrap distribution $G_T(p^\star)$ by an empirical distribution arising from Monte Carlo replications, and obtaining the quantiles from this distribution is analagous to that outlined in the previous section.

Note that the sign of $G(p^\star)$ is the opposite of the statistic used in the non-Studentized pivotal bootstrap. As a result of this, we use the "opposite quantiles" of the distribution of $G_T(p^\star)$. This idea is based on an argument involving a monotone transformation of p, say ϕ, such that $\phi(p) - \phi(p^\star)$ has a standard normal distribution. Since such a transformation often does not exist (and specifically does not exist when $p^\star$ is a biased estimator), the resulting confidence interval can behave quite poorly. In fact, the analysis in Hall (1988) leads the author to say that using the percentile method to construct confidence intervals is akin to constructing a confidence interval by "looking up (critical points in) the wrong (statistical) tables backwards."

BC_a intervals

As mentioned above, the percentile method does not work very well for biased estimators. However, improvements to the percentile method to account for bias and for skewness in the distribution of $G(p^\star)$ have led to a more popular bootstrap interval, the BC_a interval. BC stands for bias-corrected, and the a stands for accelerated. This modified technique still involves "looking up the wrong tables backwards", exactly as before, but now adjusts the specific quantiles to be used in the construction of the interval to account for the bias and the skewness (the value of a, the acceleration constant, is related to the skewness of $G(p^\star)$.

Bias correction is done by effectively compensating for "median bias" of $p^\star$ — an adjustment is made depending on where $p^\star$ fits into the bootstrap

[38] Strictly, they can not overshoot if the underlying estimator can not overshoot.

approximation of the distribution of $p^\star$ (a biased estimator will tend to have $p^\star$ away from the median of the estimated distribution).

Acceleration takes into account that the standard deviation of $p^\star$ may depend on the value of p (something not catered for in the simpler percentile and bias-corrected methods). The acceleration is quantified by an estimate of $\frac{\partial}{\partial p} se(p^\star)$ (traditionally computed as a *jackknife estimate* — see Section 5.2). This value turns out to be related to the skewness of the distribution of $p^\star$ — for details, see Hall (1988), and in particular Peter Bickel's comments on the article.

The BC_a interval performs well in practice, but involves a lot of computation, and the two phases of correction (calculation of the bias correction and the acceleration values). The computational overhead can be addressed by calculating an approximation to the BC_a interval, which Efron and Tibshirani (1993) terms the *ABC* interval. However, Hall (1988) and others maintain that despite its good performance, it is based on a poor approach (the percentile method) which is then being patched up by sophisticated techniques.

The BC_a interval provides consistent coverage close to the desired level, and is transformation-respecting, which has the additional benefit of eliminating overshooting. As a result, the BC_a interval is the recommended non-parametric bootstrap method for constructing confidence intervals (Carpenter and Bithell, 2000).

The parametric bootstrap

The bootstrap techniques described above are all examples of the so-called non-parametric bootstrap: no assumptions are made about the distribution of the points in the sample.

A popular alternative is the parametric bootstrap: in this case, the points in the sample are assumed to come from some (parametrized) class of distributions. The sample is then used to estimate the parameters of the distribution. The resulting distribution (employing the estimated parameters) is then employed as an approximation to the underlying distribution for the calculation of bootstrap statistics, instead of the empirical distribution. Thus, bootstrap samples are taken by sampling from the estimated distribution, rather than the empirical distribution.

Chapter 4

Concentration Inequalities

A final source of interval estimators based on the test sample are concentration inequalities (sometimes known as deviation inequalities). The estimators above generally use simple concentration inequalities, but the more advanced inequalities will be central to the training sample bounds which follow. Furthermore, developments in the fields of concentration inequalities have been a major source of the improvements in training sample bounds over the last 15 years. As such, this chapter will discuss the various such inequalities available to us, while presenting some test set-based interval estimators where applicable. Specifically, Sections 4.4 to 4.7 present test sample interval estimators for general loss functions which provide strict coverage of $1 - \delta$. None of the results of the previous chapter achieved this.

A statistic is said to be *concentrated* about its mean if the statistic is close to its expectation with high probability. Concentration about the median is analagous. The extent of concentration of a statistic is usually specified by means of a *concentration inequality.* In the context of test sample bounds, the statistic under consideration is the test risk of an hypothesis — thus an application of a concentration inequality to this case yields a result of the form[39]

$$\mathbb{P}_{T \sim D^k} \left\{ |r_D(w) - r_T(w)| \geq \epsilon \right\} \leq C(k, \epsilon)$$

[39] Note that $r_D(w) = \mathbb{E}_{T \sim D^k}(r_T(w))$.

for some function C. The ideal is that C exhibits exponential decay in k.[40] We shall also call bounds for the one-sided case, with the absolute value sign removed, concentration inequalities. Confidence intervals can be obtained by setting $C(k, \epsilon) = \delta$, and then solving for ϵ in terms of δ.

The chapter begins with introductory concepts being presented in Sections 4.1 to 4.3. Sections 4.4 to 4.7 consider concentration inequalities for sums of independent r.v.'s. Since the risk on a sample for a given decision rule is such a sum, these results allow us to obtain test sample intervals for a given decision rule. The last part of the chapter, from Section 4.8 to Section 4.11, considers the four major approaches to obtaining concentration inequalities for other functions of independent r.v.'s besides sums.

A large body of complex work has been done in this respect, but we limit our attention to results which will be relevant later in the thesis. For the reader interested in further study of this fascinating topic, we recommend Ledoux and Talagrand (1991), Ledoux (2001), Massart (2006), and the many papers on the topic by Michel Talagrand, notably Talagrand (1988, 1994, 1995, 1996b,c,d).

4.1 Chebyshev's inequality

We begin by noting that almost all variables are concentrated about their means in a way. This is exemplified by Chebyshev's inequality (known in earlier days as the Bienaymé-Chebyshev inequality — see Bennett, 1962, Hoeffding, 1963), which states that

$$\mathbb{P}\{\|V - \mathbb{E}\, V\| \geq \epsilon\} \leq \frac{\mathbb{V}\, V}{\epsilon^2}$$

for any r.v. V and $\epsilon > 0$ (Devroye and Lugosi, 2001).

A corresponding one-sided inequality is the Chebyshev-Cantelli inequality (Devroye and Lugosi, 2001, Exercise 2.1)[41]. One-sided inequalities are often tighter than their two-sided counterparts, helping to offset their more

[40] This reflects the asymptotic behaviour which follows from the central limit theorem.

[41] Bennett (1962) attributes this version of the Chebyshev inequality to J. Uspensky.

limited applicability. This is a good example:

$$\mathbb{P}\{V - \mathbb{E}\,V \geq \epsilon\} \leq \frac{\mathbb{V}\,V}{\mathbb{V}\,V + \epsilon^2} \tag{4.1}$$

for any r.v. V and $\epsilon > 0$. Confusingly, this one-sided version was known in earlier days as Chebyshev's inequality.

These results provide very loose bounds, applicable to highly variable statistics, but the statistics we are interested in are much better behaved. For example, the empirical risk is the sum of independent r.v.'s. Such sums are more concentrated about their means than the individual risk on each element of the sample. This allows the derivation of tighter bounds.

Chebyshev's inequality follows from an application of Markov's inequality using the transformation $\phi(v) = v^2$. A generalization of this approach are the so-called *moment bounds*, which employ the transformation $\phi(v) = v^n$ for larger natural numbers. We will not go into these bounds, but refer the interested reader to Lugosi (2004) for more details.

4.2 The exponential moment method

Most concentration inequalities for sums of independent r.v.'s are based on the *Cramér-Chernoff*, or *exponential moment* method (Chernoff, 1952), which is based on the following inequality:

$$\mathbb{P}\{V \geq \epsilon\} = \mathbb{P}\left\{e^{\lambda V} \geq e^{\lambda\epsilon}\right\} \leq \frac{\mathbb{E}\,e^{\lambda V}}{e^{\lambda\epsilon}} \ .$$

This result, which holds for any r.v. V and $\lambda, \epsilon > 0$, also follows from Markov's inequality[42]. It seems this method for deriving probability inequalities was first used by Sergei Bernstein to derive Bernstein's inequality, which we shall discuss later (Hoeffding, 1963). The exercises in Lugosi (2004) point out that the exponential moment method bound never outperforms the best moment bound, but one cannot generally know what choice of n in the transformation ϕ is optimal a priori.

[42] Since the exponential function is monotonically increasing for $\lambda > 0$.

It follows for the test sample bounds that

$$\mathbb{P}_{T\sim D^k}\left\{r_D(w) - r_T(w) \geq \epsilon\right\} \leq e^{-\lambda\epsilon}\,\mathbb{E}_{T\sim D^k}\exp(\lambda(r_D(w) - r_T(w)))\ .$$

We can expand the right hand side of this to

$$e^{-\lambda\epsilon}\,\mathbb{E}_{T\sim D^k}\exp\left(\frac{\lambda}{k}\left(\sum_{i=1}^{k}[r_D(w) - L(w(x_i^*), y_i^*)]\right)\right)\ .$$

Converting the sum in the exponent to a product, and taking the expectation into the product (since the terms in the sum are independent), we obtain

$$\begin{aligned}\mathbb{P}_{T\sim D^k}\left\{r_D(w) - r_T(w) \geq \epsilon\right\} &\leq e^{-\lambda\epsilon}\prod_{i=1}^{k}\mathbb{E}_{(x_i^*,y_i^*)\sim D}\exp\left(\frac{\lambda}{k}[r_D(w) - L(w(x_i^*), y_i^*)]\right)\\ &= e^{-\lambda\epsilon}\left[\mathbb{E}_{(x_i^*,y_i^*)\sim D}\exp\left(\frac{\lambda}{k}[r_D(w) - L(w(x_i^*), y_i^*)]\right)\right]^k\ .\end{aligned}$$

It follows that a bound on the moment generating function (m.g.f.) of $V_i = \frac{1}{k}[r_D(w) - L(h(x_i^*), y_i^*)]$ would yield a bound on the difference between the empirical and the true risk of w.

Example 4.1. Suppose L is a zero-one loss function. Setting $p = e_D(w)$, V_i assumes the value $\frac{p}{k}$ with probability $1-p$, and the value $\frac{p-1}{k}$ with probability p. The m.g.f. for V_i is thus

$$\begin{aligned}\mathbb{E}\,e^{\lambda V_i} &= (1-p)\exp\left(\frac{\lambda p}{k}\right) + p\exp\left(\frac{\lambda(p-1)}{k}\right)\\ &= \exp\left(\frac{\lambda p}{k}\right)\left[(1-p) + p\exp\left(\frac{-\lambda}{k}\right)\right]\\ &= \exp\left(\frac{\lambda p}{k}\right)\left[1 + p\left(\exp\left(\frac{-\lambda}{k}\right) - 1\right)\right]\ .\end{aligned}$$

Substituting this value into the above result gives us:

$$\begin{aligned}\mathbb{P}_{T\sim D^k}\left\{e_D(w) - e_T(w) \geq \epsilon\right\} &\leq e^{-\lambda\epsilon}\left[\exp\left(\frac{\lambda p}{k}\right)\left(1 + p\left(\exp\left(\frac{-\lambda}{k}\right) - 1\right)\right)\right]^k\\ &= e^{\lambda(p-\epsilon)}\left[1 + p\left(\exp\left(\frac{-\lambda}{k}\right) - 1\right)\right]^k\ .\end{aligned}$$

Unfortunately, this bound is not very useful, since calculating it requires the error we are trying to estimate. Substituting $p = 1$ into the expression provides a trivial bound. We will need more sophisticated ways to get upper bounds on the expression. □

4.3 Subgaussian and subexponential distributions

In the Cramér-Chernoff method, it is clearly beneficial to obtain the smallest m.g.f. possible. One way of broadly categorizing variables in this case are the classes of subgaussian and subexponential distributions.

Generally, the idea is that distributions whose m.g.f.'s are bounded by a function similar in shape to the m.g.f. of a normal (exponential) distribution, are said to have a subgaussian (subexponential) distribution.

Consider the m.g.f. of a normal distribution with parameters μ and σ^2, $\exp\left(\mu\lambda + \frac{\sigma^2\lambda^2}{2}\right)$. Based on this, we say a r.v. has a subgaussian distribution if its m.g.f. does not exceed $\exp\left(\frac{c\lambda^2}{2}\right)$ for some constant c.

An exponential distribution with parameter v has m.g.f. $\frac{1}{1-\frac{\lambda}{v}}$. Thus, we say a r.v. has a subexponential distribution if its m.g.f. does not exceed $\frac{1}{1-c\lambda}$ for some constant c.

Thus, our interpretation is that a variable with a subgaussian distribution is more concentrated than some normal distribution, while one with a subexponential distribution is more concentrated than some exponential distribution. Of course, it is beneficial if the constant c is small.

Much of what follows will be aimed at obtaining subgaussian distributions for statistics with small values of c. This is because, if a r.v. V has a subgaussian distribution with constant c, it follows from the exponential moment method that

$$\mathbb{P}\{V > \epsilon\} \leq \exp\left(\frac{-\epsilon^2}{2c}\right) \quad .$$

For further properties of such variables, see Lugosi (2004, Exercises 2.3–2.6).

A centred normal distribution is subgaussian with c equal to its variance, and we know that a sum of i.i.d. variables converges to a normal distribution with variance $n\sigma^2$, where σ^2 is the variance of each variable. Since the risk or error of a decision rule w is an average of i.i.d. variables, the best value for c we can hope to obtain is $c = \frac{\sigma^2}{n}$, which would yield a test sample bound of

$$\mathbb{P}_{T\sim D^k}\left\{r_T(w) > r_D(w) + \epsilon\right\} \leq \exp\left(\frac{-k\epsilon^2}{2\sigma^2}\right) \quad .$$

In the case of error of a decision rule, $\sigma^2 = e_D(w)[1 - e_D(w)]$, and we have that $\sigma^2 \leq \frac{1}{4}$. The following section discusses a bound where $c = \frac{1}{4n}$.

4.4 Additive Hoeffding bounds

Hoeffding (1963) implicitly provides a bound for bounded loss functions, now commonly known as *Hoeffding's lemma*[43]. We provide a more general form we shall need later.

Theorem 4.1 (Lemma 2.3 of Devroye and Lugosi, 2001). *Let W_1 and W_2 be r.v.'s with $\mathbb{E}(W_1|W_2) = 0$ and $\phi(W_2) \leq W_1 \leq \phi(W_2) + c$ for some function ϕ. Then, for $\lambda \geq 0$,*

$$\mathbb{E}\left(e^{\lambda W_1}|W_2\right) \leq \exp\left(\frac{\lambda^2 c^2}{8}\right) \quad .$$

The proof of this theorem rests on results we shall introduce later, but Hoeffding's original result rested on Jensen's inequality.

We can apply this result for a bounded loss function: we set $W_1 = V_i = \frac{1}{k}[r_D(w) - L(w(x_i^*), y_i^*)]$. Then, for any W_2 independent of W_1[44], it is clear that we can set $\phi(W_2) = \frac{1}{k}(r_D(w) - 1)$ and $c = \frac{1}{k}$. This yields a bound on the m.g.f. showing that the test risk, $r_T(w)$, has a subgaussian distribution. Using this bound with the exponential moment method, and optimizing over all $\lambda \geq 0$ yields

$$\mathbb{P}_{T\sim D^k}\left\{r_D(w) - r_T(w) \geq \epsilon\right\} \leq e^{-2k\epsilon^2} \quad , \tag{4.2}$$

for $\epsilon > 0$. We shall refer to this result as *Hoeffding's tail inequality.* An identical result holds for $r_T(w) - r_D(w)$. These are applications of a more general inequality commonly known as Hoeffding's inequality:

[43] See his Equation 4.16.

[44] The usefulness of W_2 will become apparent in later applications.

Theorem 4.2 (Hoeffding's inequality: Theorem 2 of Hoeffding, 1963). *Let $V_1, V_2, \cdots, V_n$ be independent r.v.'s with $\mathscr{L}_i \leq V_i \leq \mathscr{U}_i$ for $i \in [1:n]$. Then for $\epsilon > 0$,*

$$\mathbb{P}\left\{\sum_{i=1}^{n}(V_i - \mathbb{E}\, V_i) \geq n\epsilon\right\} \leq \exp\left(\frac{-2n^2\epsilon^2}{\sum_{i=1}^{n}(\mathscr{U}_i - \mathscr{L}_i)^2}\right) \ .$$

Setting the right hand side in Hoeffding's tail inequality equal to δ, and solving for ϵ, yields the one-sided Hoeffding's tail interval

$$\mathrm{Conf}_{1-\delta}(r_D(w)) = \left[0, r_T(w) + \sqrt{\frac{\ln\frac{1}{\delta}}{2k}}\right] \ . \tag{4.3}$$

Note that this result constitutes a bound on upper deviation with $\epsilon(T, w) = \sqrt{\frac{-\ln\delta}{2k}}$. Note that by applying the same analysis to $-W_1$ one can obtain a corresponding lower interval.

These results were derived earlier in the case of zero-one loss, implicitly by Herman Chernoff (Chernoff, 1952), and explicitly by Masashi Okamoto (Okamoto, 1958) — see Hoeffding (1963).

4.5 Relative entropy Hoeffding bounds

The proof of Hoeffding's inequality in Hoeffding (1963) actually yields a stronger result than Hoeffding's tail inequality in our situation. Hoeffding's inequality only requires independence, not identical distribution, of the values being summed. The identical distribution of the losses in our case allow a strengthening of the inequality. Modifying the statement slightly for our purposes, Hoeffding points out (in his Theorem 1) that

$$\mathbb{P}_{T\sim D^k}\left\{r_D(w) - r_T(w) \geq \epsilon\right\} \leq \exp(-k\,\mathrm{KL}(r_D(w) - \epsilon || r_D(w))) \ ,$$

for $0 < \epsilon \leq r_D(w)$. We shall refer to this result as *Hoeffding's relative entropy (r.e.) inequality.* Again, Chernoff and Okamoto had derived this result for error rates earlier. Unfortunately, the KL divergence is not analytically invertible, leaving us with two options. One approach is to use an invertible upper bound on the KL divergence to obtain bounds. Using the common upper bound $\mathrm{KL}(v-\epsilon||v) \leq 2\epsilon^2$ (Langford, 2003), one recovers Hoeffding's tail inequality above. Thus, for test sample risk intervals, this improvement is not very useful. However, the improvement will be vital for some of the training sample bounds discussed later.

The second approach is a numerical inversion of the bound. Let us write $\epsilon(t) = \epsilon(t, k, \delta)$ for the value of $\epsilon \leq t$ satisfying

$$\exp(-k\,\mathrm{KL}(t-\epsilon||t)) = \delta \ . \tag{4.4}$$

It follows that

$$\mathbb{P}_{T\sim D^k}\left\{r_D(w) - \epsilon(r_D(w)) \geq r_T(w)\right\} \leq \delta \ .$$

For the sample T, a $100(1-\delta)\%$ confidence region thus consists of those t for which

$$t - \epsilon(t) < r_T(w) \ .$$

If the left hand side is increasing, the resulting region will be an upper interval. Now,

$$\frac{\partial}{\partial t}(t-\epsilon(t)) = 1 - \frac{\partial}{\partial t}\epsilon(t)$$

so that we need $\frac{\partial}{\partial t}\epsilon(t) \leq 1$ to obtain an interval. Rewriting (4.4) as

$$\mathrm{KL}(t-\epsilon||t) = \frac{-\ln\delta}{k} \ ,$$

expanding the KL divergence, and employing implicit differentiation allows one to obtain

$$\frac{\partial}{\partial t}\epsilon(t) = 1 - \frac{\left[\frac{t-\epsilon(t)}{t} - \frac{1-(t-\epsilon(t))}{1-t}\right]}{\left[\ln\frac{t-\epsilon(t)}{t} - \ln\frac{1-(t-\epsilon(t))}{1-t}\right]} \ .$$

Since $\epsilon(t) \leq t$, the numerator and denominator of the fraction are both negative, which means that $\frac{\partial}{\partial t}(t-\epsilon(t)) \geq 0$, so that inverting the bound yields an interval.

It follows that to invert the bound, we merely need to find the value of t, say $t^\star$, for which $t - \epsilon(t) = r_T(w)$. The resulting confidence interval, $[0, t^\star]$, is the upper Hoeffding's r.e. interval, and we denote $t^\star$ by $\mathrm{URE}(r_T(w), k, \delta)$.

A lower interval can be obtained similarly, and the intervals can be combined to obtain a two-sided interval if desired. It is important to note that inverting this bound requires two embedded numerical inversions: to obtain $\mathrm{URE}(r_T(w), k, \delta)$, $t - \epsilon(t)$ must be repeatedly evaluated in order to find the point where it equals $r_T(w)$. To do this, for each t, $\epsilon(t)$ must be found by inverting the KL divergence numerically.

Since Hoeffding's r.e. inequality is tighter than Hoeffding's tail inequality, and the inversion described here is exact (up to the limits of numerical accuracy), it follows that the Hoeffding's r.e. intervals are strict improvements on the corresponding Hoeffding's tail inequality intervals.

4.6 Multiplicative Hoeffding bounds

Another popular set of bounds flow from Hoeffding's r.e. inequality. These bounds are known as *multiplicative* Chernoff bounds, and bound the relative, rather than absolute, deviation between empirical and true means. One complication, however, is that the bounds are again expressed in terms of the true mean. The resulting intervals are not as tight as Hoeffding's r.e. intervals, since they employ upper bounds on portions of the KL divergence before inverting the resulting probability statement.

Let $V_1, \cdots, V_n$ be i.i.d variables with $\mathbb{E}\, V_i = \mu$. Then, for $\mu < \epsilon \leq 1$,

$$\begin{aligned}\mathbb{P}\left\{\sum_{i=1}^{n} V_i \geq n\epsilon\right\} &= \mathbb{P}\left\{\sum_{i=1}^{n} V_i - n\mu \geq n\epsilon - n\mu\right\} \\ &\leq \exp\left(-n\,\mathrm{KL}(\mu + (\epsilon - \mu)||\mu)\right) \;,\end{aligned}$$

with the final inequality an application of Hoeffding's inequality. From the definition of the KL divergence, we obtain

$$\left(\frac{\mu}{\epsilon}\right)^{n\epsilon}\left(\frac{1-\mu}{1-\epsilon}\right)^{n(1-\epsilon)} \quad .$$

We can bound the second factor here by noting that

$$\left(\frac{1-\mu}{1-\epsilon}\right)^{n(1-\epsilon)} = \left(1 + \frac{n(\epsilon - \mu))}{n(1-\epsilon)}\right)^{n(1-\epsilon)}$$

and observing that the second term is strictly less than $e^{n(\epsilon-\mu)}$. Since, for $v > 0$,

$$\left(1+\frac{\epsilon}{v}\right)^{v} \leq e^{\epsilon} \ ,$$

we obtain the bound

$$\left(\frac{\mu}{\epsilon}\right)^{n\epsilon} e^{n(\epsilon-\mu)} \ .$$

To obtain a relative deviation inequality, we set $\epsilon = \mu(1+\kappa)$ (i.e. a factor multiplied by the mean) into the result above (for $0 < \kappa < \frac{1-\mu}{\mu}$). This yields

$$\begin{aligned}\mathbb{P}\left\{\sum_{i=1}^{n} V_i \geq n\mu(1+\kappa)\right\} &\leq \left(\frac{\mu}{\mu(1+\kappa)}\right)^{n\mu(1+\kappa)} \exp(n(\mu(1+\kappa)-\mu)) \\ &= (1+\kappa)^{-n\mu(1+\kappa)} e^{n\mu\kappa} \\ &= \exp\left(-n\mu((1+\kappa)\ln(1+\kappa)-\kappa)\right) \\ &= \exp(-n\mu\kappa\Psi(\kappa)) \ ,\end{aligned}$$

where

$$\Psi(v) = (1+\frac{1}{v})\ln(1+v) - 1$$

is a function which shall (not incidentally) crop up again when we later discuss Bennett's inequality. We shall refer to this result as the lower multiplicative Hoeffding inequality. When $\kappa \geq \frac{1-\mu}{\mu}$ the associated probability is 0, so the inequality still holds. By a similar route, one obtains the upper multiplicative Hoeffding inequality: for $\kappa \in (0,1]$,

$$\mathbb{P}\left\{\sum_{i=1}^{n} V_i \leq n\mu(1-\kappa)\right\} \leq \exp(n\mu\kappa\Psi(-\kappa)) \ .$$

These two bounds are special cases of Theorem 1 in Boucheron et al. (1999), where $h(v) = v\Psi(v)$.

One can obtain test set interval estimators from these results by setting $n = k$, and $V_i = L(w(x_i^*), y_i^*)$, so that $\mu = r_D(w)$.[45] Solving for κ when setting the right hand side to δ would generally need to be done numerically,

[45] It is important to note that these choices yield significantly tighter bounds than using our default (until now), $V_i = \frac{L(w(x_i^*),y_i^*)}{k}$. This is because the multiplicative bound is sensitive to the scale used. Since Hoeffding's inequality only applies to variables $V_i \in [0,1]$, we would like to use the largest scaling constant possible on our default V_i. Assuming L maps into $[0,1]$ means that the largest scaling constant we can use is k.

though. We are not interested in inverting this probability inequality numerically, since it will be looser than Hoeffding's r.e. intervals. Our reason for investigating the multiplicative Hoeffding bounds is to obtain bounds with closed forms, which are easier to analyse theoretically. Appropriate upper bounds on $\varPsi$ may allow the probability statement to be inverted in this manner. This is the focus of Angluin-Valiant (AV) bounds.

4.6.1 Angluin-Valiant bounds

Two useful inequalities involving $\varPsi$ are:

- for $v \in (0, 1]$,
$$\varPsi(-v) \leq \frac{-v}{2} \quad ;$$
- for $v > 0$,
$$\varPsi(v) \geq \frac{v}{2 + \frac{2v}{3}} \quad . \tag{4.5}$$

These follow from the equivalent formulae for h in Section 2 of Boucheron et al. (1999).

Applying the first inequality along with the upper multiplicative Hoeffding inequality, yields for $0 < \kappa < 1$,

$$\mathbb{P}\left\{\sum_{i=1}^{n} V_i \leq n\mu(1-\kappa)\right\} < \exp\left(-\frac{n\mu\kappa^2}{2}\right) \quad ,$$

which we shall call the upper Chernoff inequality.

For a test sample interval estimator, we obtain

$$\mathbb{P}_{T\sim D^k}\left\{r_T(w) \leq r_D(w)(1-\kappa)\right\} < \exp\left(-\frac{kr_D(w)\kappa^2}{2}\right) \quad .$$

Setting the right hand side to δ and solving for κ, one obtains

$$\kappa = \sqrt{\frac{-2\ln\delta}{kr_D(w)}} \quad .$$

The expression inside the probability is then

$$r_T(w) \leq r_D(w)\left(1 - \sqrt{\frac{-2\ln\delta}{kr_D(w)}}\right) \ ,$$

which can be rewritten as

$$\psi(r_D(w), r_T(w)) \geq \sqrt{\frac{-2\ln\delta}{k}} \ ,$$

where ψ is the upper relative deviation.

Inverting this deviation measure yields the upper AV interval,

$$\left(0, r_T(w) - \frac{\ln\delta}{k}\left(1 + \sqrt{1 - \frac{2kr_T(w)}{\ln\delta}}\right)\right] \ .$$

Machine learning Chernoff inequalities

One common formulation of the upper Chernoff inequality for risk in the machine learning literature is (using $\kappa = \frac{\epsilon}{\sqrt{r_D(w)}}$)

$$\mathbb{P}_{T\sim D^k}\left\{\frac{r_D(w) - r_T(w)}{\sqrt{r_D(w)}} > \epsilon\right\} < \exp\left(\frac{-k\epsilon^2}{2}\right) \ .$$

A number of related results generally hold for exponential bounds on relative deviations such as this one. We present a number of them below, and note that the arguments generally hold for similar exponential bounds on relative deviations. The results here are based on the similar results in Bartlett (1998, Corollary 7) and Shawe-Taylor et al. (1998, Theorems 3.1 and 3.2).

Given $r_D(w) > \epsilon$, it can be seen that $r_T(w) = 0$ implies

$$\frac{r_D(w) - r_T(w)}{\sqrt{r_D(w)}} > \sqrt{\epsilon} \ .$$

Thus,

$$\mathbb{P}_{T\sim D^k}\left\{r_T(w) = 0 | r_D(w) > \epsilon\right\} \leq \exp\left(\frac{-k\epsilon}{2}\right) \ .$$

This probability statement yields an interval estimator as follows: we use the statistic $r_T(w)$. If this is non-zero, the statement gives us no information, and the interval is $[0, 1]$. If $r_T(w) = 0$, we obtain the interval

$$\left[0, \frac{-2\ln\delta}{k}\right]$$

from the equation $\delta = \exp\left(\frac{-k\epsilon}{2}\right)$. We call the resulting confidence interval the *realizable risk interval.* The probability statement above is as tight as the upper Chernoff inequality, but it's use is restricted to cases where we observe $r_T(w) = 0$. This scenario is more common for training sample bounds, when it may be known a priori that a training error of zero will be achieved by the training algorithm. This situation is known as the *realizable case* in machine learning — see Section 3.2.12 for a further discussion.

More generally, given $r_D(w) > \epsilon$, if

$$r_T(w) \leq (1-\kappa)r_D(w)$$

for some κ, then

$$r_D(w) - r_T(w) \geq \kappa r_D(w) \ .$$

Dividing by $\sqrt{r_D(w)}$ and using $r_D(w) > \epsilon$ on the right hand side, we obtain that

$$\frac{r_D(w) - r_T(w)}{\sqrt{r_D(w)}} \geq \kappa\sqrt{\epsilon} \ .$$

It follows that

$$\begin{aligned} & \mathbb{P}_{T\sim D^k}\left\{r_T(w) \leq (1-\kappa)r_D(w) | r_D(w) > \epsilon\right\} \\ \leq \ & \mathbb{P}_{T\sim D^k}\left\{\frac{r_D(w) - r_T(w)}{\sqrt{r_D(w)}} > \kappa\sqrt{\epsilon}\right\} \\ \leq \ & \exp\left(-\frac{k\kappa^2\epsilon}{2}\right) \ . \end{aligned}$$

When $\kappa = 1$, the realizable case result above is obtained. Inverting this probability statement is slightly more tricky. If $r_T(w) \leq (1-\kappa)\epsilon$, we can set

$$\delta = \exp\left(-\frac{k\kappa^2\epsilon}{2}\right)$$

to obtain the interval

$$\left[0, \frac{-2\ln\delta}{k\kappa^2}\right] \ .$$

When $r_T(w) > 1-\kappa$, the probability statement provides no information, so we obtain the interval $[0,1]$. For intermediate values of $r_T(w)$, we do not generally know whether the probability statement provides information or not, since we do not know whether $r_T(w) \leq (1-\kappa)r_D(w)$ or not. As a result, we use the bound $[0,1]$ in this case as well. We call the resulting interval the *realistic risk interval* — this is because the bounds apply to what we shall call the realistic case in training sample bounds: where we expect a low training error on the selected decision rule.

Another result is that, for all $\kappa > 0$,[46]

$$\mathbb{P}_{T\sim D^k}\left\{r_D(w) > (1+\kappa)r_T(w) + \epsilon\right\} \leq \exp\left(-\frac{k\epsilon\kappa^2}{2(1+\kappa)^2}\right) \ .$$

We provide a brief proof for this.

Proof. First, we suppose that

$$r_D(w) - r_T(w) \leq \epsilon_0\sqrt{r_D(w)} \ .$$

Select some $\kappa > 0$. Then, if $\kappa r_T(w) \geq \epsilon_0\sqrt{r_D(w)}$,

$$\begin{array}{rcl} r_D(w) & \leq & r_T(w) + \epsilon_0\sqrt{r_D(w)} \\ & \leq & r_T(w) + \kappa r_T(w) \\ & \leq & (1+\kappa)r_T(w) + \left(\dfrac{1+\kappa}{\kappa}\right)^2 {\epsilon_0}^2 \ . \end{array}$$

On the other hand, if $\kappa r_T(w) < \epsilon_0\sqrt{r_D(w)}$,

$$\begin{array}{rcl} r_D(w) & \leq & r_T(w) + \epsilon_0\sqrt{r_D(w)} \\ & < & \dfrac{\epsilon_0\sqrt{r_D(w)}}{\kappa} + \epsilon_0\sqrt{r_D(w)} \\ & = & \dfrac{1+\kappa}{\kappa}\epsilon_0\sqrt{r_D(w)} \ . \end{array}$$

Dividing throughout by $\sqrt{r_D(w)}$ and squaring yields,

$$\begin{array}{rcl} r_D(w) & < & \left(\dfrac{1+\kappa}{\kappa}\right)^2 {\epsilon_0}^2 \\ & \leq & (1+\kappa)r_T(w) + \left(\dfrac{1+\kappa}{\kappa}\right)^2 {\epsilon_0}^2 \ . \end{array}$$

[46] Note that Bartlett (1998) mistakenly uses $\kappa < 0$.

It follows that

$$\begin{aligned}\mathbb{P}_{T\sim D^k}&\left\{r_D(w) > (1+\kappa)r_T(w) + \left(\frac{1+\kappa}{\kappa}\right)^2 {\epsilon_0}^2\right\} \\ \leq\ & \mathbb{P}_{T\sim D^k}\left\{\frac{r_D(w) - r_T(w)}{\sqrt{r_D(w)}} > \epsilon_0\right\} \ .\end{aligned}$$

Using $\epsilon_0 = \frac{\kappa\sqrt{\epsilon}}{1+\kappa}$, and applying the upper Chernoff inequality yields the result. □

Setting $\kappa = 1$ in this result, we obtain an upper confidence interval for $r_D(w)$ of $[0, 2r_T(w) - \frac{8\ln\delta}{k}]$. Compared to, for example, the Hoeffding's tail interval, we see that the term involving k and δ decreases at a faster rate, but at the expense of the factor 2 before $r_T(w)$. Since this interval is based on a slackening of the upper Chernoff inequality, this interval is inferior to the upper AV interval, which is also easily obtained analytically.

The lower multiplicative Hoeffding inequality with the second inequality for Ψ yields for $\kappa > 0$,

$$\mathbb{P}\left\{\sum_{i=1}^n V_i \geq n\mu(1+\kappa)\right\} < \exp\left(-\frac{n\mu\kappa^2}{2+\frac{2\kappa}{3}}\right) \ .$$

Setting the right hand side to δ and solving yields a quadratic equation, but the resulting expression for κ is very difficult to handle when inverting the resulting probability statement.

Clearly the denominator in the right hand side expression is always less than $2+\kappa$, and for $\kappa \leq 1 < 1.5$ the denominator is less than 3. Using these facts, we obtain the lower multiplicative Chernoff inequalities, due to Angluin and Valiant (1979) [47]: for $0 < \kappa < 1$,

$$\mathbb{P}\left\{\sum_{i=1}^n V_i \geq n\mu(1+\kappa)\right\} < \exp\left(-\frac{n\mu\kappa^2}{3}\right) \ ;$$

and, for $\kappa \geq 1$,

$$\mathbb{P}\left\{\sum_{i=1}^n V_i \geq n\mu(1+\kappa)\right\} < \exp\left(-\frac{n\mu\kappa^2}{2+\kappa}\right) \ .$$

[47] We note in passing that Vidyasagar (2002) presents an alternative bound when $\kappa > 1$, in the case of i.i.d. Bernoulli r.v.'s.

Applying this to the risk of a decision rule yields, for $0 < \kappa < 1$,

$$\mathbb{P}_{T\sim D^k}\left\{r_T(w) \geq r_D(w)(1+\kappa)\right\} < \exp\left(-kr_D(w)\frac{\kappa^2}{3}\right) \quad ;$$

and, for $\kappa \geq 1$,

$$\mathbb{P}_{T\sim D^k}\left\{r_T(w) \geq r_D(w)(1+\kappa)\right\} < \exp\left(-kr_D(w)\frac{\kappa^2}{2+\kappa}\right) \quad .$$

Setting the right hand side of the first probability statement to δ yields that

$$\mathbb{P}_{T\sim D^k}\left\{\psi(r_D(w), r_T(w)) < \sqrt{\frac{-3\ln\delta}{k}}\right\} \geq 1-\delta$$

when $r_D(w) > \frac{-3\ln\delta}{k}$ (otherwise, the value of κ exceeds 1 and the bound does not apply), and where ψ now denotes lower relative deviation.

Inverting this deviation yields the interval

$$\begin{aligned} & \left(r_T(w) + \frac{\epsilon^2\left[1-\sqrt{1+\frac{4r_T(w)}{\epsilon^2}}\right]}{2}, 1\right] \\ = & \left(r_T(w) - \frac{\frac{3\ln\delta}{k}\left[1-\sqrt{1-\frac{4kr_T(w)}{3\ln\delta}}\right]}{2}, 1\right] \quad . \end{aligned}$$

However, this result does not, strictly speaking, provide a confidence interval: this interval only applies when $r_D(w) > \frac{-3\ln\delta}{k}$, so that instead we have a confidence region

$$\left[0, \frac{-3\ln\delta}{k}\right] \cup \left(r_T(w) - \frac{\frac{3\ln\delta}{k}\left[1-\sqrt{1-\frac{4kr_T(w)}{3\ln\delta}}\right]}{2}, 1\right] \quad .$$

Using the second probability statement to solve for κ again runs into a quadratic equation, and the resulting expression inside the probability statement must be inverted numerically. Thus, although the lower Chernoff inequality may be a useful source of bounds, obtaining a general lower (and hence a two-sided) Angluin-Valiant bound is not an easy task, and we shall not pursue it further here. The difficulties we face result from the method employed to upper bound the KL divergence in Hoeffding's r.e. inequality, which is asymmetric in its arguments. To obtain Hoeffding's tail inequality, a symmetric upper bound on the KL divergence was employed. For the multiplicative Hoeffding and Chernoff inequalities, the upper bound employed is asymmetric (notably $v\Psi(v)$ is not symmetric in v).

We note that the upper Angluin-Valiant interval is narrower than the upper Hoeffding's tail interval when the underlying risk $r_D(w)$ is small, even though they both rely on the same result, due to the different natures of the approximations used (Vidyasagar, 2002). This is because Hoeffding's tail inequality trades symmetry of the upper bound on the KL divergence for tightness. As a result, it performs poorly when the distribution of the sum of the V_i around $n\mu$ is not symmetric (in our case, when $r_D(w)$ is not near 0.5).

4.7 Bennett's and Bernstein's inequalities

We know that the empirical risk of a decision rule w on an i.i.d. sample P of size n is asymptotically normally distributed, with mean $r_D(w)$. Let the variance of the loss of w be σ^2. Then the variance of $r_P(w)$ is $\frac{\sigma^2}{n}$, so that asymptotically

$$\mathbb{P}_{P\sim D^n}\left\{r_D(w) - r_P(w) \geq \epsilon\right\} \approx \mathbb{P}_{Z\sim N(0,1)}\left\{Z \geq \phi(\epsilon)\right\} \;\;,$$

where

$$\phi(\epsilon) = \frac{\epsilon\sqrt{n}}{\sigma} \;\;.$$

The area under this normal tail equals[48]

$$\begin{aligned} 1 - \Phi(\phi(\epsilon)) &< \frac{\varphi(\phi(\epsilon))}{\phi(\epsilon)} \\ &= \frac{1}{\sqrt{2\pi}} \frac{\exp\left(-\frac{\phi(\epsilon)^2}{2}\right)}{\phi(\epsilon)} \;\;. \end{aligned}$$

[48] The first inequality, now known as *Mills' inequality*, is based on a bound on the Mills' ratio in Gordon (1941).

It is instructive to compare the form of this asymptotic expression with Hoeffding's tail inequality (Devroye and Lugosi, 2001, Hoeffding, 1963). In particular, we are interested in the rate of decay of the probability as the sample size increases, i.e. the form of the exponent. For our asympotic expression, this is

$$\frac{-\phi(\epsilon)^2}{2} = \frac{-n\epsilon^2}{2\sigma^2} \ .$$

Hoeffding's tail inequality for error rates has an exponent of $-2n\epsilon^2$. Inspection reveals that this is simply an upper bound on the exponent in the asymptotic expression, achieved when $\sigma^2 = 0.25$ (such as for a zero-one loss, with $e_D(w) = 0.5$). Thus, Hoeffding's tail inequality appears to use a uniform upper bound on the variance of $r_D(w)$ — this is once again a result of its symmetry. We shall say that the *effective variance* of Hoeffding's tail inequality is 0.25.

An investigation of the exponent in the upper Chernoff inequality shows that the expression corresponding to σ^2 there, the effective variance of the inequality, is $r_D(w)$, a simple upper bound on σ^2. Furthermore, consideration of $r_D(w) \leq 0.25$ shows that the upper Chernoff inequality is a stronger result than Hoeffding's tail inequality for $r_D(w) \leq \frac{1}{4}$.

One inequality incorporating the variance of sums of independent random variables directly is the following:

Theorem 4.3 (Theorem 3 of Hoeffding, 1963). *Let $V_1, \cdots, V_n$ be independent real-valued r.v.'s with zero mean, and assume that $V_i \leq c$ for $i = 1, \cdots, n$. Let*

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} \mathbb{V}\, V_i \ .$$

Then, for $0 < \epsilon < c$,

$$\begin{aligned} \mathbb{P}\left\{\sum_{i=1}^{n} V_i > n\epsilon\right\} &\leq \left(\left(1 + \frac{c\epsilon}{\sigma^2}\right)^{-\frac{\sigma^2 + c\epsilon}{\sigma^2 + c^2}} \left(1 - \frac{\epsilon}{c}\right)^{-\frac{c^2 - c\epsilon}{\sigma^2 + c^2}}\right)^n \\ &\leq \exp\left(-\frac{n\epsilon}{c}\Psi\left(\frac{c\epsilon}{\sigma^2}\right)\right) \ , \end{aligned}$$

where $\Psi(v) = (1 + \frac{1}{v})\ln(1 + v) - 1$, for $v \geq 0$.

The second bound above was developed in 1962 by George Bennett (Bennett, 1962), and is now known as *Bennett's inequality.* The derivation of Bennett's inequality is closely related to the derivation of the multiplicative Hoeffding inequalities from Hoeffding's r.e. inequality.

Hoeffding's article further discusses the relationship of Bennett's inequality to previous bounds. Particularly, Bennett's inequality is a strengthening of Bernstein's inequality, which was derived in the 1920's by Sergei Bernstein (Bernstein, 1924, 1927). The most simple form of Bernstein's result shows more clearly how the variance of the loss is being employed:

Theorem 4.4 (p.34 of Bennett, 1962). *Under the conditions of Theorem 4.3, for any $\epsilon > 0$,*

$$\mathbb{P}\left\{\sum_{i=1}^{n} V_i > n\epsilon\right\} \leq \exp\left(-\frac{n\epsilon^2}{2\sigma^2 + \frac{2c\epsilon}{3}}\right) \ .$$

This result follows by employing the lower bound on Ψ in (4.5). The interested reader is referred to Hoeffding (1963) and Bennett (1962) for a more extensive discussion of the relationships between these bounds and other bounds, including earlier inequalities by Prohorov, Kolmogorov, Loève, and Berry.

In fact, stronger forms of Bernstein's inequality are available, although perhaps not as widely recognised. For a collection of the major forms of Bernstein's inequality for independent variables, the interested reader is referred to Theorem 2.1 of Bousquet (2002b). The most basic result there is a slight, but useful, strengthening of Bernstein's inequality above, which we reproduce here.

Theorem 4.5 (Theorem 2.1 of Bousquet, 2002b). *Under the conditions of Theorem 4.3, for any $\delta > 0$,*

$$\mathbb{P}\left\{\sum_{i=1}^{n} V_i \geq \sigma\sqrt{-2n\ln\delta} - \frac{c\ln\delta}{3}\right\} \leq \delta \ .$$

We would like to use Bernstein's inequality to obtain a test sample interval estimator for $r_D(w)$. To show that Theorem 4.5 provides an improvement over Theorem 4.4, we shall illustrate both of Bernstein's bounds above in this scenario. From Theorem 4.4 one obtains (using $V_i = r_D(w) - L(w(x_i^*), y_i^*)$)

$$\mathbb{P}_{T\sim D^k}\{r_D(w) - r_T(w) > \epsilon\} \leq \exp\left(-\frac{k\epsilon^2}{2\sigma^2 + \frac{2\epsilon}{3}}\right) \ .$$

Setting the right hand side equal to δ yields a quadratic equation in ϵ. Solving for ϵ yields the one-sided interval

$$\left[0, r_T(w) + \frac{(-\ln\delta) + \sqrt{(-\ln\delta)^2 - 18k\sigma^2\ln\delta}}{3k}\right] \quad , \tag{4.6}$$

with the complication being that σ is typically dependent on $r_D(w)$.

On the other hand, Theorem 4.5 leads to the confidence interval

$$\left[0, r_T(w) + \frac{(-\ln\delta) + \sqrt{-18k\sigma^2\ln\delta}}{3k}\right] \quad , \tag{4.7}$$

so that the $(-\ln\delta)^2$ term has been eliminated from the square root term.

The effective variance of Bernstein's inequality in Theorem 4.4 is $\sigma^2 + \frac{\epsilon}{3}$, very similar to the desired σ^2 (Lugosi, 2004). Bernstein's bound thus incorporates information on the variance, but at a premium of $\frac{\epsilon}{3}$, which may be undesirably large for some ϵ: when σ^2 is close to 0.25, this premium is typically not worthwhile compared to Hoeffding's tail inequality, but for smaller and larger values of $r_D(w)$ it becomes worthwhile. A further complication is that σ^2 is generally unknown, so a bound on σ^2 must actually be used. To get an improvement in bounding the error rate over Hoeffding's tail inequality from Bernstein's inequality, we need a bound, ς^2, on σ^2 to satisfy $\varsigma^2 + \frac{\epsilon}{3} \leq 0.25$. This is easily verified for a specific ϵ. However, when constructing confidence intervals, we consider a range of values for ϵ, so that this bound will generally not hold uniformly.

We now consider a simple bound on the variance and the resulting test sample interval estimators. The corresponding confidence regions must be obtained numerically, but they turn out to be intervals.

Bernstein interval based on $r_D(1 - r_D)$

The maximum variance of a r.v. taking on values in $[0, 1]$, with mean μ is $\mu(1 - \mu)$. This is achieved when the r.v. is Bernoulli, and it is easy to show that it is the maximum by expanding the formula defining variance.

This leads us to consider the bound on variance

$$\varsigma^2 = r_D(w)[1 - r_D(w)] \geq \sigma^2 \ .$$

With this choice, Theorem 4.5 yields

$$\mathbb{P}_{T \sim D^k}\left\{ r_D(w) - r_T(w) \geq \frac{(-\ln\delta) + \sqrt{-18kr_D(w)[1 - r_D(w)]\ln\delta}}{3k} \right\} \leq \delta \ .$$

We can thus construct a confidence region by including the values of t for which $t - r_T(w) < \epsilon(t)$, where

$$\epsilon(t) = \frac{(-\ln\delta) + \sqrt{-18kt(1-t)\ln\delta}}{3k} \ .$$

So, if $t - \epsilon(t)$ is a nondecreasing function, the confidence region will be an upper interval. Now

$$\begin{aligned} \frac{d}{dt}[t - \epsilon(t)] &= 1 - \frac{1}{3k}\left[\frac{1}{2}(-18kt(1-t)\ln\delta)^{-\frac{1}{2}}(-18k\ln\delta)(1-2t)\right] \\ &= 1 - \frac{(-3\ln\delta)(1-2t)}{\sqrt{-18kt(1-t)\ln\delta}} \ . \end{aligned} \tag{4.8}$$

It is easy to see that the derivative above exceeds 1 for $t \geq \frac{1}{2}$. For $t < \frac{1}{2}$, the inequality

$$\begin{aligned} 0 &\leq \frac{d}{dt}[t - \epsilon(t)] \\ &= 1 - \frac{(-3\ln\delta)(1-2t)}{\sqrt{-18kt(1-t)\ln\delta}} \end{aligned}$$

can be solved by multiplying throughout by the square root expression, placing the resulting free-standing square root alone on one side of the inequality, and squaring both sides. The lower root of the resulting quadratic equation satisfies $t < \frac{1}{2}$. It follows that $t - \epsilon(t)$ is increasing for

$$t \geq \frac{1 - \sqrt{1 - \frac{(-\ln\delta)}{(-\ln\delta) + \frac{k}{2}}}}{2} > 0 \ ,$$

and decreasing otherwise, so that it seems we cannot always strictly be guaranteed an upper interval by numerical inversion in this case.

However, since $t - \epsilon(t)$ is decreasing over the interval

$$\left[0, \frac{1 - \sqrt{1 - \frac{(-\ln\delta)}{(-\ln\delta) + \frac{k}{2}}}}{2}\right) ,$$

it follows that

$$\begin{aligned} t - \epsilon(t) &\leq -\epsilon(0) \\ &= \frac{\ln\delta}{3k} \\ &< 0 \end{aligned}$$

over that range. Thus, no

$$t \leq \frac{1 - \sqrt{1 - \frac{(-\ln\delta)}{(-\ln\delta) + \frac{k}{2}}}}{2}$$

will have $t - \epsilon(t) > r_T(w)$, so that it turns out inverting the inequality does indeed always yield an upper interval. We call the resulting confidence interval the *upper Bernstein interval based on* $r_D(1 - r_D)$. When applied to a zero-one loss, we shall call it the *upper exact-variance Bernstein interval for error.*

Note that both Bennett and Bernstein's bounds are applied to centred variables, so that lower and two-sided intervals can potentially be obtained by applying the inequalities to the negated variables.

Bounding Pollard-Haussler deviation

The simple form of Bernstein's inequality in Theorem 4.4 can be used to obtain an interval estimator from the upper P-H ν-deviation between the empirical and true risk of a decision rule w,

$$\psi_\nu\left(r_D(w), r_T(w)\right) = \frac{r_D(w) - r_T(w)}{\nu + r_D(w) + r_T(w)} .$$

We have

$$\begin{aligned}
& \mathbb{P}_{T\sim D^k}\left\{\psi_\nu(r_D(w), r_T(w)) > \epsilon\right\} \\
= \; & \mathbb{P}_{T\sim D^k}\left\{\sum_{i=1}^k L(x_i^*, y_i^*) - kr_D(w) > \epsilon\left[k(\nu + r_D(w)) + \sum_{i=1}^k L(x_i^*, y_i^*)\right]\right\} \\
\leq \; & \mathbb{P}_{T\sim D^k}\left\{\sum_{i=1}^k L(x_i^*, y_i^*) - kr_D(w) > \epsilon k(\nu + r_D(w))\right\} \quad ,
\end{aligned}$$

since the losses are all positive.

We wish to apply Bernstein's inequality here to $V_i = L(x_i^*, y_i^*) - r_D(w)$ and $n = k$, so $c = 1$. To do this we again need an upper bound for the variance of V_i. Employing $\varsigma^2 = r_D(w)$, we obtain

$$\mathbb{P}_{T\sim D^k}\left\{\sum_{i=1}^k L(x_i^*, y_i^*) - kr_D(w) > k\epsilon(\nu + r_D(w))\right\} \leq \exp\left(-\frac{k[\epsilon(\nu + r_D(w))]^2}{2r_D(w) + \frac{2\epsilon(\nu + r_D(w))}{3}}\right) \quad .$$

The above result is dependent on $r_D(w)$, so that we can not evaluate it. We can remove this dependence by maximizing the bound over potential values of $r_D(w)$. This maximum occurs when

$$\frac{(\nu + r_D(w))^2}{r_D(w) + \frac{\epsilon(\nu + r_D(w))}{3}}$$

is minimized. We thus solve

$$\begin{aligned}
0 &= \frac{d}{dt}\frac{(\nu + t)^2}{t + \frac{\epsilon(\nu+t)}{3}} \\
&= \frac{2(\nu + t)}{t + \frac{\epsilon(\nu+t)}{3}} - \frac{(1 + \frac{\epsilon}{3})(\nu + t)^2}{\left[t + \frac{\epsilon(\nu+t)}{3}\right]^2} \quad .
\end{aligned}$$

This reduces to

$$2[(3 + \epsilon)t + \epsilon\nu] - (3 + \epsilon)(\nu + t) = 0 \quad ,$$

so that the bound is maximized when $r_D(w) = \frac{3-\epsilon}{3+\epsilon}\nu$. This yields a bound of

$$\exp\left(-\frac{k\left[\epsilon(\nu + \frac{3-\epsilon}{3+\epsilon}\nu)\right]^2}{2\frac{3-\epsilon}{3+\epsilon}\nu + \frac{2\epsilon\left(\nu + \frac{3-\epsilon}{3+\epsilon}\nu\right)}{3}}\right) \quad ,$$

which simplifies to

$$\exp\left(\frac{-18k\epsilon^2\nu}{(3+\epsilon)^2}\right) \quad . \tag{4.9}$$

This can in turn be bounded by

$$\exp\left(-\frac{9}{8}k\epsilon^2\nu\right) < \exp(-k\epsilon^2\nu) \ ,$$

for $\epsilon < 1$, the only case of interest. The exposition up to here roughly follows that in Haussler (1992).

Setting (4.9) to δ yields a quadratic equation in ϵ. This equation has no solution for $18k\nu \leq -\ln\delta$, so that for a specific k and δ, ν needs to exceed a certain size to obtain meaningful bounds from this analysis. When $18k\nu > -\ln\delta$, we obtain

$$\epsilon(\delta) = \frac{-3\left(1+\sqrt{\frac{18k\nu}{-\ln\delta}}\right)}{1-\frac{18k\nu}{-\ln\delta}} \quad . \tag{4.10}$$

This bound on the P-H deviation can then be inverted to obtain a corresponding test-set upper confidence interval for $r_D(w)$,

$$\left(0, \frac{[1+\epsilon(\delta)]r_T(w)+\epsilon(\delta)\nu}{1-\epsilon(\delta)}\right] \quad .$$

It is also possible to invert the P-H deviation by employing other inequalities. Haussler (1992, Lemma 9) presents a result for the Chebyshev inequality, for example, but the resulting bound is less tight than that employing Bernstein's inequality as above.

It is natural to wonder whether we can improve this result by relying on the more direct estimate of the variance, $\varsigma^2 = r_D(w)(1-r_D(w))$. In a similar manner to before, we obtain

$$\begin{aligned}\mathbb{P}_{T\sim D^k}&\left\{\sum_{i=1}^{k} L(x_i^*, y_i^*) - kr_D(w) > k\epsilon(\nu + r_D(w))\right\} \\ &\leq \exp\left(-\frac{k\left[\epsilon(\nu+[r_D(w)(1-r_D(w))])\right]^2}{2\left[r_D(w)(1-r_D(w))\right]+\frac{2\epsilon(\nu+[r_D(w)(1-r_D(w))])}{3}}\right) \quad .\end{aligned}$$

To eliminate the dependence on $r_D(w)$, we maximize the bound on the right over possible values of $r_D(w)$. A maximum on $(0,1)$ can only occur if

$$0 = \frac{d}{dt}\frac{(\nu + t(1-t))^2}{t(1-t) + \frac{\epsilon(\nu+t(1-t))}{3}} \quad .$$

This equation can be shown to have three solutions: $t = \frac{1}{2}$, and

$$t = \frac{1 \pm \sqrt{1 - 4\nu\frac{3-\epsilon}{3+\epsilon}}}{2}$$

(when $\nu\frac{3-\epsilon}{3+\epsilon} \leq \frac{1}{4}$). Substitution of the roots above into the original formula shows that the last two roots yield a larger bound than that for $t = \frac{1}{2}$ when they exist. The last two choices of t lead to $t(1-t) = \frac{3-\epsilon}{3+\epsilon}\nu$, so that the same bound is obtained as with the more naïve upper bound on the variance. When $\nu\frac{3-\epsilon}{3+\epsilon} > \frac{1}{4}$, the only root $t = \frac{1}{2}$ can be shown to yield a minimum for the bound, rather than a maximum. It follows that we can make the bound uniform by substituting $r_D(w) = 0$ or $r_D(w) = 1$ into the bound (both choices yield the same result): for $\nu\frac{3-\epsilon}{3+\epsilon} > \frac{1}{4}$, we can improve the bound on P-H deviation to

$$\exp\left(-\frac{3}{2}k\epsilon\nu\right) \quad .$$

4.8 The martingale method

More generally, Hoeffding (1963) points out that Bennett's inequality is the best possible bound obtainable by directly employing the exponential moment method, under general assumptions. Thus, the previous section brings our consideration of bounds on the sums of independent random variables to an end.

Until now, we have restricted our attention to those simple concentration inequalities which can be applied directly to test sample bounds. In what follows, we will discuss more sophisticated results, allowing us to construct training sample bounds. The appeal of these results is that they can directly provide bounds without resorting to the union bound, which we shall discuss later.

Lugosi (2004) presents an informal argument demonstrating that the sum of independent random variables can be seen as the least concentrated of any measurable function of those variables. The idea of the results we shall study now will be to obtain stronger concentration results for other functions of these variables.

The next four sections will briefly consider the major approaches to obtaining concentration inequalities, viz. the martingale, transportation, induction, and entropy methods. Central to all of these approaches is the exponential moment method, so all the approaches effectively consider sophisticated methods for obtaining bounds on m.g.f.'s.

The first approach we shall study employs the exponential moment method, but uses martingales to obtain the m.g.f. of the relevant statistic. This will be necessary to decompose the m.g.f. of a function into a product of individual m.g.f.'s. We will also need to employ some advanced techniques in order to obtain bounds on m.g.f.'s of certain statistics. The concentration inequalities we discuss next will make use of the full power of Theorem 4.1.

4.8.1 Bounded differences and McDiarmid's inequality

To obtain bounds on a function ϑ of independent variables $E_1, \cdots, E_n$ from some space $\mathcal{E}$, we will need to make some assumptions about $\vartheta : \mathcal{E}^n \to \mathrm{I\!R}$. The most popular and well-known such assumption is that of *bounded differences.* This assumption basically states that each variable E_i has a limited influence on the realization of the statistic $V = \vartheta(E_1, \cdots, E_n)$. We say that ϑ satisfies the *bounded difference assumption* with constants $c_1, \cdots, c_n$ if, for each $i \in [1 : n]$,

$$\sup_{\eta_1,\ldots,\eta_n,\eta' \in \mathcal{E}} |\vartheta(\eta_1, \cdots, \eta_n) - \vartheta(\eta_1, \cdots, \eta_{i-1}, \eta', \eta_{i+1}, \cdots, \eta_n)| \leq c_i \ .$$

We bound $W = V - \mathbb{E}\, V$ by expanding it into a telescoping series, with each term the difference of two conditional expectations of V: defining

$$V_i = \mathbb{E}(V|E_1, \ldots, E_i)$$

and

$$W_i = V_i - V_{i-1}$$

we have

$$W = \sum_{i=1}^{n} W_i \ .$$

We note that the sequence of V_i's are the *Doob martingale* derived from ϑ and $E_1, \cdots, E_n$.

Now, it is clear that

$$\begin{aligned} \inf_{\eta \in \mathcal{E}} \mathbb{E}(V|E_1, \ldots, E_{i-1}, \eta) - V_{i-1} &\leq W_i \\ &\leq \sup_{\eta \in \mathcal{E}} \mathbb{E}(V|E_1, \ldots, E_{i-1}, \eta) - V_{i-1} \ . \end{aligned}$$

The difference between these upper and lower bounds on W_i is

$$\sup_{\eta \in \mathcal{E}} \mathbb{E}(V|E_1, \ldots, E_{i-1}, \eta) - \inf_{\eta \in \mathcal{E}} \mathbb{E}(V|E_1, \ldots, E_{i-1}, \eta)$$

which can be seen to be upper bounded by the value of c_i in the bounded differences assumption.

Thus, we can apply Theorem 4.1 to W_i, using

$$\phi(E_1, \cdots, E_{i-1}) = \inf_{\eta \in \mathcal{E}} \mathbb{E}(V|E_1, \ldots, E_{i-1}, \eta) - V_{i-1}$$

and $c = c_i$ to obtain

$$\mathbb{E}\left(e^{\lambda W_i}|E_1, \ldots, E_{i-1}\right) \leq \exp\left(\frac{\lambda^2 c_i^2}{8}\right) \ .$$

The next step is to apply Chernoff's bounding method to W, but we can not decompose the m.g.f. by using independence directly. Instead, we decompose the m.g.f. into the product of m.g.f.'s of *conditional* variables:

$$\begin{aligned} \mathbb{E}\, e^{\lambda W} &= \mathbb{E} \exp\left(\lambda \sum_{i=1}^{n} W_i\right) \\ &= \mathbb{E}\left(\exp\left(\lambda \sum_{i=1}^{n-1} W_i\right) \mathbb{E}\left(e^{\lambda W_n}|E_1, \ldots, E_{n-1}\right)\right) \\ &\leq \exp\left(\frac{\lambda^2 c_n^2}{8}\right) \mathbb{E} \exp\left(\lambda \sum_{i=1}^{n-1} W_i\right) \ . \end{aligned}$$

Repeating this process n times[49], and optimizing the resulting bound over λ provides *the bounded difference inequality* developed by McDiarmid in 1989[50]:

Theorem 4.6 (Theorem 2.2 of Devroye and Lugosi, 2001). *If a function ϑ of r.v.'s $E_1, \cdots, E_n$, defining a random variable V, satisfies the bounded difference assumption with constants $c_1, \cdots, c_n$, then, for all $\epsilon > 0$,*

$$\mathbb{P}\{V - \mathbb{E}\, V \geq \epsilon\} \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right) \ .$$

This result is easily shown to be a generalization of Hoeffding's inequality: Consider V defined by $\vartheta(E_1, \cdots, E_n) = \frac{1}{n}\sum_{i=1}^n E_i$, with $\mathscr{L}_i \leq E_i \leq \mathscr{U}_i$ for $i \in [1:n]$, and the E_i independent. Note that ϑ satisfies the bounded difference assumption with constants $\frac{\mathscr{U}_i - \mathscr{L}_i}{n}$, allowing us to apply the bounded difference inequality to V. The result is Hoeffding's inequality.

Another inequality which yields a similar result is *Azuma's inequality* (Azuma, 1967). Application of Azuma's inequality to the Doob martingale corresponding to $\vartheta(E_1, \cdots, E_n)$ in the previous paragraph yields a generalization of Hoeffding's inequality with the independence assumption replaced by a bounded difference type assumption. A result of the generalization is a factor four weakening in the exponent of Hoeffding's inequality. In contrast, the bounded difference inequality here strictly extends Hoeffding's inequality, but retains the independence requirement and the bounded difference requirement.

Example 4.2. In our case, it will be useful to apply the bounded difference inequality to $V = \vartheta(S) = \sup_{w \in \mathcal{W}}(r_D(w) - r_S(w))$, where S is the training sample. Since $r_S(w)$ cannot change by more than $\frac{1}{m}$ when one element of the training sample S is modified, regardless of w, it follows that V satisfies the bounded difference assumption with $c_i = \frac{1}{m}$ for $i \in [1:m]$. Using the first bounded difference inequality, it follows that

$$\mathbb{P}\{V \geq \mathbb{E}\, V + \epsilon\} \leq e^{-2m\epsilon^2} \ .$$

Thus the maximal deviation between the empirical and true mean of a decision rule for a given sample is, with high probability, close to the mean maximal deviation. □

[49] In this process, we show that W has a subgaussian distribution.

[50] A two-sided result can be obtained by applying the theorem to $-\vartheta$.

The approach used to derive the bounded difference inequality is based on decomposing a function into a sum of martingale differences. This approach is known as *Yurinski's method* (named so, it seems, by Michel Talagrand (Talagrand, 1996c), with reference to the work in Yurinskii, 1974). Next, we consider an alternative approach to deriving concentration inequalities, known as the *information-theoretic* or *transportation* method.

4.9 The transportation method

This approach is again based on bounding an m.g.f. and applying the Cramér-Chernoff method. This time, bounds on the m.g.f. of a statistic follow from statements which relate the difference in expectation of the statistic over two distributions to the Kullback-Leibler divergence between the two distributions.

Massart (1998) uses this approach to derive an N-dimensional generalization of Hoeffding's inequality. One of the main ingredients was the following result (stated in more generality in the reference):

Theorem 4.7 (Lemma 2.2 of Massart, 1998). *Let $(\mathcal{E}, \Sigma, \tau)$ be a probability space. The following two statements are equivalent for any r.v. $E \in \mathcal{E}$ and $v > 0$:*

- $\mathbb{E}_{E\sim\tau} \exp\left(\lambda[E - \mathbb{E}\, E]\right) \leq \exp\left(\frac{\lambda^2 v}{2}\right)$, *and*
- *For any probability measure τ' on $\mathcal{E}$ which is absolutely continuous*[51] *with respect to τ,*

$$\mathbb{E}_{E\sim\tau'} E - \mathbb{E}_{E\sim\tau} E \leq \sqrt{2v\,\mathrm{KL}(\tau'||\tau)} \ .$$

Note that the first statement of the pair is similar to Hoeffding's lemma in that it posits a subgaussian distribution of $E - \mathbb{E}\, E$: if we have a result of this form, the Cramér-Chernoff method will yield a bound as desired. Our approach is then to obtain such a bound by seeking a statement of the second form. A source of results of this form are so-called *transportation*

[51] This is required for the Kullback-Leibler divergence to be defined.

cost inequalities derived in the study of the *transportation problem* (Massart, 2006, p.36).

The most well-known such inequality for deriving concentration of measure results is Marton's inequality, named for Katalin Marton, who discovered the basic form (Marton, 1986). We present a refined form here.

Theorem 4.8 (Marton's inequality: Proposition 2.5 of Massart, 1998). *Let $(\mathcal{E}, \Sigma, \tau)$ be an N-dimensional product probability space, and let τ' be absolutely continous with respect to τ. Denote by $\mathcal{P}(\tau, \tau')$ the collection of* couplings *of τ and τ', i.e. probability measures over $\mathcal{E}^2$ such that the marginal distribution over the first N coordinates is τ, and over the second N coordinates is τ'. Then, for $E \in \mathcal{E}^2$,*

$$\inf_{\tau^\star \in \mathcal{P}(\tau,\tau')} \sum_{i=1}^{N} \left(\mathbb{P}_{E\sim\tau^\star} \left\{ E^{(i)} \neq E^{(N+i)} \right\} \right)^2 \leq \frac{1}{2} \mathrm{KL}(\tau' || \tau) \ .$$

Combining these two ingredients leads to bounds on the m.g.f. of functions exhibiting a generalized Lipschitz continuity, leading to a concentration of measure results for such functions.

Theorem 4.9 (Part of Theorem 3.2 of Massart, 1998). *Let $(\mathcal{E}, \Sigma, \tau)$ be an N-dimensional product probability space, with $\mathcal{E} = \prod_{i=1}^{N} \mathcal{E}_i$, where $(\mathcal{E}_i, d_{\mathcal{E}_i})$ is a metric space of diameter[52] $\mathscr{D}_i$ for each $i \in [1:N]$. Let $\vartheta : \mathcal{E} \to \mathbb{R}$ be a functional defining a r.v. $V = \vartheta(E)$.*

If ϑ is Lipschitz in the sense that for any $E_1, E_2 \in \mathcal{E}$,

$$|\vartheta(E_1) - \vartheta(E_2)| \leq \sum_{i=1}^{N} c_i d_{\mathcal{E}_i} \left(E_1^{(i)}, E_2^{(i)} \right) \ ,$$

then, for $\epsilon > 0$,

$$\mathbb{P}_{E\sim\tau} \left\{ V - \mathbb{E}\, V \geq \epsilon \right\} \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{N} c_i^2 \mathscr{D}_i^2} \right) \ .$$

This result is slightly generalized from that in the article: the original had $c_i = 1$ for every i. We call $(c_1, \cdots, c_N)$ the Lipschitz vector of ϑ, and call ϑ a

[52] The diameter of a metric space is the supremum of the distance between any two points in the space.

$(c_1, \cdots, c_N)$-Lipschitz functional. If all the components of the Lipschitz vector of ϑ are equal to some constant c, we shall call ϑ a c-Lipshitz functional, and c the Lipschitz constant of ϑ. This generalizes the standard meaning of the term to product spaces by combining the distance measures using a Manhattan metric.

Proof. We begin by considering any distribution τ' absolutely continuous with respect to τ, and the collection of couplings $\mathcal{P}(\tau, \tau')$.

Our intent is to establish a relationship between $\mathbb{E}_{E\sim\tau'} V - \mathbb{E}_{E\sim\tau} V$ and

$$\min_{\tau^\star \in \mathcal{P}(\tau,\tau')} \sum_{i=1}^{N} \left(\mathbb{P}_{E\sim\tau^\star} \left\{ E^{(i)} \neq E^{(N+i)} \right\} \right)^2 \quad ,$$

and then applying Marton's inequality. Thereafter, Theorem 4.7, constituting the first ingredient of the transportation cost method, will deliver a bound on the m.g.f. of V.

Let $\tau^\star \in \mathcal{P}(\tau, \tau')$. Then,

$$\begin{aligned}
\mathbb{E}_{E\sim\tau'} V - \mathbb{E}_{E\sim\tau} V &= \mathbb{E}_{E\sim\tau^\star} \left(\vartheta \left(E^{(N+1)}, \cdots, E^{(2N)} \right) - \vartheta \left(E^{(1)}, \cdots, E^{(N)} \right) \right) \\
&\leq \mathbb{E}_{E\sim\tau^\star} \left(\sum_{i=1}^{N} c_i d_{\mathcal{E}_i} \left(E^{(i)}, E^{(N+i)} \right) \right) \\
&\leq \mathbb{E}_{E\sim\tau^\star} \left(\sum_{i=1}^{N} c_i \mathscr{D}_i I \left(E^{(i)} \neq E^{(N+i)} \right) \right) \\
&= \sum_{i=1}^{N} (c_i \mathscr{D}_i) \, \mathbb{E}_{E\sim\tau^\star} I \left(E^{(i)} \neq E^{(N+i)} \right) \\
&= \sum_{i=1}^{N} (c_i \mathscr{D}_i) \, \mathbb{P}_{E\sim\tau^\star} \left\{ E^{(i)} \neq E^{(N+i)} \right\} \\
&\leq \left(\sum_{i=1}^{N} c_i^2 \mathscr{D}_i^2 \right)^{\frac{1}{2}} \left(\sum_{i=1}^{N} \mathbb{P}_{E\sim\tau^\star} \left\{ E^{(i)} \neq E^{(N+i)} \right\}^2 \right)^{\frac{1}{2}} \quad ,
\end{aligned}$$

where we have used the definition of diameter, the fact that the expectation of an indicator function reduces to probability, and the Cauchy-Schwartz inequality. Noting that this argument holds for all couplings of τ and τ', we have that

$$\begin{aligned}
\mathbb{E}_{E\sim\tau'} V - \mathbb{E}_{E\sim\tau} V &\leq \left(\sum_{i=1}^{N} c_i^2 \mathscr{D}_i^2\right)^{\frac{1}{2}} \inf_{\tau^\star\in\mathcal{P}(\tau,\tau')} \left(\sum_{i=1}^{N} \mathbb{P}_{E\sim\tau^\star}\left\{E^{(i)} \neq E^{(N+i)}\right\}^2\right)^{\frac{1}{2}} \\
&\leq \left(\sum_{i=1}^{N} c_i^2 \mathscr{D}_i^2\right)^{\frac{1}{2}} \left(\frac{1}{2}\,\mathrm{KL}(\tau'||\tau)\right)^{\frac{1}{2}} \\
&= \sqrt{2\left[\frac{1}{4}(\sum_{i=1}^{N} c_i^2 \mathscr{D}_i^2)\right]\mathrm{KL}(\tau'||\tau)} \ ,
\end{aligned}$$

with the second step following from Theorem 4.8 after taking the infimum into the (continuous) square root.

The result follows by applying Theorem 4.7 (noting that the above argument holds for any τ' absolutely continuous with respect to τ), and applying the Cramér-Chernoff method. □

Example 4.3. We shall use the results above to derive the multi-dimensional version of Hoeffding's inequality in Massart (1998).

Let $\mathcal{E}_i$ be an axis-parallel rectangle in $\mathbb{R}^M$ for each $i \in [1:N]$. Specifically,

$$\mathcal{E}_i = \left\{\eta \in \mathbb{R}^M : \mathscr{L}_{i,j} \leq \eta^{(j)} \leq \mathscr{U}_{i,j} \forall j \in [1:M]\right\} \ .$$

Consider the statistic V generated by the function

$$\vartheta(E) = \sup_{j\in[1:M]} \sum_{i=1}^{N} E^{(i,j)} \ ,$$

where $E^{(i,j)}$ denotes the j-th coordinate of $E^{(i)}$. Then

$$\begin{aligned}
|\vartheta(E_1) - \vartheta(E_2)| &= \left|\sup_{j\in[1:M]} \sum_{i=1}^{N} E_1^{(i,j)} - \sup_{j\in[1:M]} \sum_{i=1}^{N} E_2^{(i,j)}\right| \\
&\leq \left|\sum_{i=1}^{N}\left(\sup_{j\in[1:M]} E_1^{(i,j)} - \sup_{j\in[1:M]} E_2^{(i,j)}\right)\right| \\
&\leq \left|\sum_{i=1}^{N} \sup_{j\in[1:M]} \left(E_1^{(i,j)} - E_2^{(i,j)}\right)\right| \\
&\leq \sum_{i=1}^{N} \sup_{j\in[1:M]} |E_1^{(i,j)} - E_2^{(i,j)}| \ .
\end{aligned}$$

Thus, with respect to the ℓ^∞ metric over each $\mathcal{E}_i$, we have that ϑ is a 1-Lipschitz functional. Furthermore, the diameter of $\mathcal{E}_i$ in this metric is $\sup_{j\in[1:M]}(\mathscr{U}_{i,j} - \mathscr{L}_{i,j})$. Applying the above theorem thus yields, for $\epsilon > 0$,

$$\mathbb{P}\{V - \mathbb{E}\,V \geq \epsilon\} \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^N \sup_{j\in[1:M]}(\mathscr{U}_{i,j} - \mathscr{L}_{i,j})^2}\right) \ .$$

This is an M-dimensional generalization of Hoeffding's inequality, and hence the Hoeffding tail inequality. □

4.10 Isoperimetric inequalities and the induction method

We now outline a third approach to deriving concentration of measure results, pioneered by Michel Talagrand. We begin by sketching the framework in which we shall discuss the approach, known as the *induction method.* We shall see that the distinguishing feature of this approach is the use of *isoperimetric inequalities*, despite the name of the method. The name refers to a technique for proving such inequalities employing mathematical induction, pioneered by Michel Talagrand in the late 1980s and early 1990s. For a review, see Talagrand (1995).

This approach also employs the Cramér-Chernoff bounding method. However, the form of the bounds on m.g.f.'s derived with this technique means that the results of the exponential moment method naturally bound the deviation of a statistic from its median, rather than its mean, as is the case with the previous two methods. For highly concentrated variables, the mean and median are typically very close, so results bounding deviation from the mean can typically be obtained from the natural results at a small premium. We will discuss this conversion at the end of this section.

Definition 4.1. For a prametric space[53] $(\mathcal{E}, d)$, we define the *point-set extension* of d, as the function $\bar{d}$ mapping $\eta \in \mathcal{E}$ and a set $A \subseteq \mathcal{E}$ to

$$\bar{d}(\eta, A) = \inf_{\eta' \in A} d(\eta, \eta') \ .$$

[53] A prametric space is a generalization of a metric space, with the metric being replaced by a prametric. A prametric is a non-negative function satisfying $d(\eta,\eta) = 0$ for all η in the space.

Furthermore, we define the ε-*blowup* of A as

$$B_\varepsilon(A) = \left\{\eta \in \mathcal{E} : \bar{d}(\eta, A) \leq \varepsilon\right\} \ ,$$

i.e. the closure of the set of points in $\mathcal{E}$ within "distance" ε of A.

We now consider functions of independent r.v.'s $E_1, \cdots, E_n$ each from a probability space $(\mathcal{E}, \varSigma, \tau)$. It is convenient to view these r.v.'s a single r.v. E in the product probability space[54] $(\mathcal{E}^n, \varSigma_n, \tau^n)$.

For some function ϑ of E defining a statistic V, consider the set

$$A(V) = \{\eta \in \mathcal{E}^n : \vartheta(\eta) \leq M(V)\} \ ,$$

where $M(V)$ is the median of V. By definition, $\tau^n(A(V)) \geq \frac{1}{2}$.

For certain functions ϑ, the measure of $B_\varepsilon(A(V))$ can be related to the probability that $\vartheta(\eta) \leq M(V) + \epsilon$, for an appropriate metric. Such relationships can be established by so-called *isoperimetric inequalities*, and the use of isoperimetric inequalities is the distinguishing feature of Talagrand's approach. Such inequalities provide upper bounds on the measure of the complement of an *enlargement* (such as the ε-blowup) of some set A in a probability space $(\mathcal{E}', \varSigma', \tau')$. A typical *blowup inequality* (an isoperimetric inequality with respect to a blowup enlargement) might state that for all $A \subseteq \mathcal{E}'$ with $\tau'(A) \geq \frac{1}{2}$,

$$\tau(B_\varepsilon(A)) \geq 1 - \phi(\tau, \varepsilon) \ ,$$

for some function ϕ. Ideally, ϕ should exhibit exponential decay.

Isoperimetric results developed when studying the question of which curve of a fixed length encloses the largest area.[55] (The term "isoperimetric" refers to this fixed length.) Later the problem developed into higher dimensions (e.g. which body of fixed area encloses the largest volume), and also into more abstract spaces. Ironically, in this process, the problem statement converted to the equivalent dual problem of finding, for a fixed area, the set of that area

[54] As is customary, $\varSigma_n$ is the σ-algebra generated by the n-fold Cartesian product of measurables sets.

[55] Of course, some mathematical sophistication was also necessary.

with the shortest perimeter, or boundary. In Euclidean space with a Borel measure, the solution is, intuitively, the sphere. Other classical results (see, for example, Ledoux, 2001) are the spherical isoperimetric inequality of Paul Lévy, which shows that geodesic balls are the solution to the problem on a sphere using its geodesic metric and the normalized Haar measure; and the closely related Gaussian isoperimetric inequality, which states that if Euclidean space is endowed with a Gaussian measure, the solution becomes halfspaces, instead of balls. In this case, it seems that the length of the perimeter has become infinite. However, for generalization of the problem to general measure spaces, the volume enclosed by and perimeter of a set are defined with respect to the measure: the volume of the set is considered to be its measure, while the perimeter is the so-called *boundary measure* (or *Minkowski content*) of the set. For a set A in the space $(\mathcal{E}', \Sigma', \tau')$, the boundary measure of A can be defined as

$$\liminf_{\varepsilon \to 0^+} \frac{1}{\varepsilon}[\tau'(B_\varepsilon(A)) - \tau'(A)] \ .$$

This derivative-type quantity can be seen as the rate of growth of the volume of R if the set is expanded along it's boundary. The value inside the limit inferior appears in most isoperimetric inequalities, which explains their name. However, our applications of these inequalities will consider the non-limiting behaviour of this quantity (i.e. when ε is a little distance away from zero). This use of these inequalities was pioneered by Vitali Milman in his simplified proof of Dvoretzky's theorem in 1971.

We obtain our first isoperimetric inequality from the following bound on the m.g.f. of the Hamming distance of a point from a set, based on Talagrand (1995, Sections 2.1 and 2.2).

Theorem 4.10. *Consider the product probability space* $(\mathcal{E}^n, \Sigma_n, \tau^n)$*, equipped with the* Hamming distance

$$d(\eta_1, \eta_2) = \left| \left\{ i \in [1:n] : \eta_1^{(i)} \neq \eta_2^{(i)} \right\} \right| \ .$$

and its point-set extension $\bar{d}$*.*

Then, for $A \subseteq \mathcal{E}^n$, a r.v. $E \in \mathcal{E}^n$, and $\lambda, v > 0$,

$$\begin{aligned}\mathbb{E}\, e^{\lambda \bar{d}(A,E)} &\leq \frac{\phi(v,\lambda)^n}{(\tau^n(A))^v} \\ &\leq (\tau^n(A))^{-v} \exp\left(\frac{\lambda^2 n(v+1)}{8v}\right) \quad ,\end{aligned}$$

where

$$\phi(v,\lambda) = \frac{v^v \left(\exp(\lambda) - \exp\left(\frac{-\lambda}{v}\right)\right)^{1+v}}{(v+1)^{v+1}\left(1 - \exp\left(\frac{-\lambda}{v}\right)\right)(\exp(\lambda)-1)^v} \quad .$$

Thus $\bar{d}(A,E)$ has a subgaussian distribution, and it follows from Cramér-Chernoff bounding after setting $v = 1$ that

$$\mathbb{P}\left\{\bar{d}(A,E) \geq \epsilon\right\} \leq \frac{1}{\tau^n(A)} \exp\left(\frac{-\epsilon^2}{n}\right) \quad .$$

By optimizing over the choice of λ, one obtains the stronger result,

$$\mathbb{P}\left\{\bar{d}(A,E) \geq \epsilon\right\} \leq \exp\left(-\frac{2}{n}\left(\epsilon - \sqrt{-\frac{n \ln \tau^n(A)}{2}}\right)^2\right) \quad ,$$

which, however, only applies for $\epsilon \geq \sqrt{-\frac{n \ln \tau^n(A)}{2}}$.

The proof of the first part is obtained by mathematical induction on n. Early isoperimetric inequalities were proved using a sequence of transformations known as *rearrangements* (Talagrand, 1995). Michel Talagrand was the first to use induction on the number of dimensions to obtain isoperimetric inequalities (Talagrand, 1988), and discovered that this approach was very useful. This approach has now been applied widely, and the proof technique is commonly known as *Talagrand's induction method*. Interestingly, the induction on the number of coordinates can be seen as a generalization of the martingale approach, explaining why the approach leads to more powerful results (Talagrand, 1995).

We note that a theoretical improvement on this bound when $\mathcal{E} = \{0,1\}$ is also provided in Section 2.3 of Talagrand (1995). However, the resulting bounds on error are expressed in terms of the error itself (in a similar fashion to Example 4.1), and the bound depends on an unspecified constant. As such, we shall not reproduce the result here.

The following example shows that the sample error of a decision rule is close to the median sample error with high probability by applying Theorem 4.10.[56]

Example 4.4. Consider a decision rule w. The error of w on a point $(x, y) \in \mathcal{Z}$ drawn according to D then has a Bernoulli distribution with parameter $e_D(w)$. Thus, a sample $P = [\![(x_1, y_1), \cdots, (x_n, y_n)]\!]$ generates a sequence of Bernoulli r.v.'s, $E_i(P) = L(w(x_i), y_i), i \in [1 : n]$. We shall apply the concentration result to $V = \vartheta(E_1(P), \cdots, E_n(P)) = \sum_{i=1}^{n} E_i(P)$ to obtain an error bound.

To apply the bound, we set $\mathcal{E} = \{0, 1\}$ and $\tau(\{1\}) = 1 - \tau(\{0\}) = e_D(w)$. Let $M(V)$ be the median of V with respect to the product measure τ^n. Now consider the set $R(V) = \{\eta \in \mathcal{E}^n : \vartheta(\eta) \leq M(V)\}$. Then, by the definition of median, $\tau^n(R(V)) \geq \frac{1}{2}$.

It is apparent, but will be convenient to note, that the function ϑ is 1-Lipschitz with respect to the Hamming distance d:

$$|\vartheta(\eta_1) - \vartheta(eta_2)| \leq d(\eta_1, \eta_2) \ .$$

Consider a point $\eta_1 \notin R(V)$ with $\vartheta(\eta_1) = j \geq M(V)$. Then, for any $\eta_2 \in R(V)$, we have

$$\begin{aligned} d(\eta_1, \eta_2) &\geq |\vartheta(\eta_1) - \vartheta(\eta_2)| \\ &= j - \vartheta(\eta_2) \\ &\geq j - M(V) \ , \end{aligned}$$

so that $\bar{d}(R(V), \eta_1) \geq j - M(V)$. It follows that

$$\{\eta \in \mathcal{E}^n : \vartheta(\eta) \geq j \geq M(W)\} \subseteq \left\{\eta \in \mathcal{E}^n : \bar{d}(R(V), \eta) \geq j - M(V)\right\} \ .$$

The set on the left corresponds to those loss sequences with j or more components equal to 1. For a sample P generating such a loss sequence, we thus have $ne_P(w) \geq j$. We are thus interested in the measure of this set for $j = M(W) + n\epsilon$. But this is upper bounded by the measure of the set on the right hand side, which can in turn be bounded by the isoperimetric inequality presented above.

[56] It is interesting to note that the result obtained in this example can also be obtained by an application of the transportation method results outlined earlier — see Massart (1998, p.17).

Applying the isoperimetric inequality above yields

$$\mathbb{P}_{E\sim\tau^n}\left\{\bar{d}(R(V),E)\geq n\epsilon)\right\}\leq \exp\left(-2n\left(\epsilon-\sqrt{\frac{\ln 2}{2n}}\right)^2\right)$$

for $\epsilon \geq \sqrt{\frac{\ln 2}{2n}}$.

Thus (for $\epsilon > \sqrt{\frac{\ln 2}{2n}}$)

$$\begin{aligned}\mathbb{P}_{E\sim\tau^n}\left\{V\geq M(V)+n\epsilon\right\} &= \mathbb{P}_{P\sim D^m}\left\{e_P(w)\geq M(e_P(w))+\epsilon\right\}\\ &\leq \exp\left(-2n\left(\epsilon-\sqrt{\frac{\ln 2}{2n}}\right)^2\right) \ .\end{aligned}$$

A similar treatment of a reversal of the loss function (which has value 1 when the original loss function has value 0, and vice-versa), yields the same result with $e_P(w)$ and $M(e_P(w))$ exchanged, as we desire (i.e. this approach yields a two-sided result).

Note that applying this result to the test sample T allows us to obtain a test sample interval estimator for the median error. □

We now consider the problem of obtaining a bound on the regular deviation from the mean, rather than the median. Suppose we have a function $\phi(\epsilon)$ such that for a r.v. V

$$\mathbb{P}\left\{|V-M(V)|\geq\epsilon\right\}\leq\phi(\epsilon) \ .$$

Then

$$\begin{aligned}|M(V)-\mathbb{E}\,V| &\leq \mathbb{E}\,|M(V)-V|\\ &= \int_0^\infty \mathbb{P}\left\{|M(V)-V|\geq\epsilon\right\}\,d\epsilon\\ &\leq \int_0^\infty \phi(\epsilon)\,d\epsilon \ .\end{aligned}$$

Now

$$|V-\mathbb{E}\,V|\leq|V-M(V)|+|M(V)-\mathbb{E}\,V| \ ,$$

so that

$$\begin{aligned}
\mathbb{P}\{|V - \mathbb{E}\,V| \geq \epsilon\} &\leq \mathbb{P}\{|V - M(V)| \geq \epsilon - |M(V) - \mathbb{E}\,V|\} \\
&\leq \mathbb{P}\left\{|V - M(V)| \geq \epsilon - \int_0^\infty \phi(\epsilon)\,d\epsilon\right\} \\
&\leq \phi\left(\epsilon - \int_0^\infty \phi(\epsilon)\,d\epsilon\right) \quad .
\end{aligned}$$

Finally, we present Talagrand's *convex distance inequality*, which was originally proved using (a sophisticated form of) the approach outlined in this section.

Theorem 4.11 (Theorem 3.10 of Philips, 2005). *For any probability measure τ on $[0,1]^n$, any convex function ϑ on $\mathbb{R}^n$ which is c-Lipschitz (w.r.t. the ℓ_n^2 metric), and any $\epsilon \geq 0$,*

$$\mathbb{P}_{E\sim\tau}\{|V - M(V)| \geq \epsilon\} \leq 4\exp\left(\frac{-\epsilon^2}{4c^2}\right) \quad ,$$

where the statistic V is defined by $V = \vartheta(E)$.

4.11 The entropy method

Another approach, known as the *entropy method*, can also be used to obtain concentration inequalities. The entropy method was popularised by Michel Ledoux with his work in Ledoux (1996). This work attempted to obtain better insight into the dramatically improved concentration inequalities derived by Talagrand in the mid-1990's with his induction method, by simplifying the proofs. In addition, the entropy method has provided bounds with optimal constants in important cases, whereas bounds from the induction method typically employ unspecified constants. We shall see an example of this later in the section.

Central to the basic entropy method is the so-called *Herbst argument* (Massart, 2006), which shows that a centred r.v. has a subgaussian distribution if a certain relationship between the entropy and the m.g.f. of the underlying r.v. holds (hence the name of the method).

Theorem 4.12 (Herbst argument: Proposition 2.14 of Massart, 2006). *Let E be an integrable r.v. on a probability space $(\mathcal{E}, \Sigma, \tau)$.*

If, for some $v > 0$, the inequality

$$\operatorname{Ent}_{E\sim\tau}\left(e^{\lambda E}\right) \leq \frac{\lambda^2 v}{2}\, \mathbb{E}_{E\sim\tau}\, e^{\lambda E}$$

holds for all $\lambda > 0$, then, for all $\lambda > 0$,

$$\mathbb{E}_{E\sim\tau}\, e^{\lambda(E - \mathbb{E}\, E)} \leq \exp\left(\frac{\lambda^2 v}{2}\right) \quad . \tag{4.11}$$

Proof. Suppose

$$\operatorname{Ent}_{E\sim\tau}\left(e^{\lambda E}\right) \leq \frac{\lambda^2 v}{2}\, \mathbb{E}\, e^{\lambda E} \quad ,$$

where $\mathbb{E}_{E\sim\tau}\, E = 0$. Expanding the formula for entropy and dividing by $\lambda^2\, \mathbb{E}_{E\sim\tau}\, e^{\lambda E}$ yields

$$\frac{\mathbb{E}_{E\sim\tau}(E e^{\lambda E})}{\lambda\, \mathbb{E}_{E\sim\tau}\, e^{\lambda E}} - \frac{\ln \mathbb{E}_{E\sim\tau}\, e^{\lambda E}}{\lambda^2} \leq \frac{v}{2} \quad .$$

Writing $\phi(\lambda) = \mathbb{E}_{E\sim\tau}\, e^{\lambda E}$, and noting that

$$\phi' \equiv \frac{d\phi}{d\lambda} = \mathbb{E}_{E\sim\tau}(E e^{\lambda E})$$

(assuming the relevant integrals exist), we obtain a differential inequality:

$$\frac{\phi'(\lambda)}{\lambda \phi(\lambda)} - \frac{\ln \phi(\lambda)}{\lambda^2} \leq \frac{v}{2} \quad .$$

This inequality is solved by noting that the left hand side equals

$$\frac{d}{d\lambda}\left[\frac{\ln \phi(\lambda)}{\lambda}\right] \quad ,$$

so that we obtain

$$\frac{\ln \phi(\lambda)}{\lambda} \leq \frac{\lambda v}{2} \quad .$$

Multiplying both sides by λ and exponentiating both sides yields the desired result.

The proof is concluded by noting that if the condition in the theorem holds with $\mathbb{E}_{E\sim\tau}\, E \neq 0$, it also holds for $E - \mathbb{E}_{E\sim\tau}\, E$. □

The second major ingredient of the entropy method is an appropriate *logarithmic Sobolev inequality*[57]. Probably the best-known such logarithmic Sobolev inequality is Gross' inequality (Gross, 1975).

Theorem 4.13 (Gross' inequality: Theorem 3.9 of Massart, 2006). *Let τ be the standard Gaussian measure on the Euclidean space $\mathbb{R}^N$, and ϕ be any continuously differentiable function on $\mathbb{R}^N$. Then, if $E \sim \tau$,*

$$\mathrm{Ent}_{E\sim\tau}([\phi(E)]^2) \leq 2\,\mathbb{E}_{E\sim\tau}\,\|[\nabla\phi](E)\|^2 \ ,$$

where $\nabla\phi$ denotes the gradient of ϕ.

Comparing this result to the condition for the Herbst argument, it is natural to consider $\phi = \exp(\frac{\lambda\vartheta}{2})$, where ϑ defines a statistic V on $\mathbb{R}^N$.[58] In that case,

$$\begin{aligned}\|[\nabla\phi](E)\|^2 &= \left\|\exp\left(\frac{\lambda V}{2}\right)\frac{\lambda}{2}[\nabla\vartheta](E)\right\|^2 \\ &= \frac{\lambda^2 e^{\lambda V}}{4}\|[\nabla\vartheta](E)\|^2 \ ,\end{aligned}$$

so that

$$\begin{aligned}\mathrm{Ent}_{E\sim\tau}([\phi(E)]^2) &= \mathrm{Ent}_{E\sim\tau}(e^{\lambda V}) \\ &\leq \frac{\lambda^2}{2}\,\mathbb{E}_{E\sim\tau}\left(\|[\nabla\vartheta](E)\|^2 e^{\lambda V}\right) \ .\end{aligned}$$

We note this is similar, but not yet the same, as the condition used in the Herbst argument. To employ the Herbst argument, we need some further condition on ϑ. The simplest assumption is that ϑ is c-Lipschitz. It follows that $\|[\nabla\vartheta](E)\| \leq c$, so that one can apply the Herbst argument to V with $v = c^2$. The roots of this idea have been traced back to work in Davies and Simon (1984), but it seems to have first been stated in roughly this form in Ledoux (1996).

[57] A Sobolev space is a subset of an L^p space meeting certain requirements. A Sobolev inequality is an inequality showing that a Sobolev space is also a subset of other L^p spaces. A logarithmic Sobolev inequality is an inequality showing that a Sobolev space can be embedded into an Orlicz space (Ledoux, 2001, Section 5.1)

[58] Technically, ϑ needs to be continuously differentiable, but this restriction can be removed after the theorem has been proved (Massart, 2006, p.62).

A major restriction of Gross' inequality is the restriction to the Gaussian measure. The major development presented in Ledoux (1996) was the idea of using logarithmic Sobolev inequalities over other distributions to obtain concentration results. A major contribution in this direction was Ledoux (1996, Theorem 1.2) which showed that Gross' inequality also holds for smooth functions with respect to any product probability on $[0,1]^N$.

Example 4.5. Consider a decision rule w. The risk of w on a point $(x,y) \in \mathcal{Z}$ drawn according to D then has some distribution τ on $[0,1]$ with mean $r_D(w)$. An i.i.d. sample $P = [\![(x_1,y_1),\cdots,(x_n,y_n)]\!]$ generates a sequence of r.v.'s each with distribution τ, namely $E_i(P) = L(w(x_i),y_i), i \in [1:n]$. Thus the sequence $(E_1(P),\cdots,E_n(P)) \in [0,1]^n$ has the product measure τ^n. Defining the function $\vartheta(\eta) = \frac{1}{n}\sum_{i=1}^n \eta^{(i)}$ for $\eta \in [0,1]^n$, we note that ϑ is $\sqrt{\frac{1}{n}}$-Lipschitz (w.r.t. the Euclidean metric) and smooth. Thus we have that Gross' inequality holds for $\phi(\eta) = \exp(\frac{\lambda\vartheta(\eta)}{2})$, which in turn yields, by the Herbst argument (using $v = \frac{1}{n}$), that

$$\mathbb{E}_{E\sim\tau}\, e^{\lambda(V-\mathbb{E}\,V)} \leq \exp\left(\frac{\lambda^2}{2n}\right) \quad ,$$

where V is the statistic generated by ϑ. Clearly, for a sample P, $V = r_P(w)$. Applying the exponential moment method, and optimising for λ yields

$$\mathbb{P}_{P\sim D^n}\left\{r_P(w) \geq r_D(w) + \epsilon\right\} \leq \exp\left(\frac{-n\epsilon^2}{2}\right) \quad .$$

□

In the mid-1990's, Talagrand (1996b) derived a strong concentration inequality for the supremum of an empirical process, using the induction method. The bound corresponds to a result of the form

$$\mathbb{P}_{E\sim\tau^n}\{V \geq \mathbb{E}\,V + \epsilon\} \leq K \exp\left(\frac{-\epsilon^2}{2(c_1\varsigma^2 + c_2 b\epsilon)}\right) \quad , \tag{4.12}$$

where $V = \vartheta(E)$, with

$$\vartheta(E) = \sup_{\phi\in\mathcal{V}} \sum_{i=1}^n \phi(E^{(i)}) \quad ,$$

the supremum of an empirical process indexed by a countable function class $\mathcal{V}$, the components of E are independent, b is a uniform bound on the norm of the elements of $\mathcal{V}$, and ς^2 is a bound on the variance of V. However, while the result has good asymptotic properties, the constants K, c_1 and c_2 implied by the proof were rather poor. When Ledoux (1996) proposed the entropy method, he showed that the same concentration inequality could be obtained using the approach, and that the resulting constants were reasonable: $K = 2$, $c_1 = 42$ and $c_2 = 8$. Later work using the entropy method in Boucheron et al. (1999), Massart (2000) refined the constants even further, obtaining $K = 1$, $c_1 = 8$, and $c_2 = 2.5$.

If we consider (4.12) when $\mathcal{V}$ is restricted to a single function, we see that the result corresponds to the simple form of Bernstein's inequality in Theorem 4.4, with $V - \mathbb{E}\, V$ here corresponding to $\sum_{i=1}^{n} V_i$ there, and ϵ here equal to $n\epsilon$ there. Furthermore, the typical definitions of ς and b ensure that Bernstein's inequality is a specialization of (4.12) with tighter constants: $K = 1$, $c_1 = 1$ and $c_2 = \frac{1}{3}$. Based on results on concentration of the Gaussian measure in high dimensions, it was conjectured that it may be possible to obtain the same constants in the general case (for further discussion, see Massart, 2000).

The work in Boucheron et al. (1999) made a significant step towards confirming this conjecture. However, they presented a concentration inequality not directly for suprema of empirical processes, but instead for *self-bounding functions.*

Definition 4.2 (Self-bounding function). A function $\vartheta : \mathcal{E}^n \to \mathbb{R}$ of r.v.'s $E_1, \cdots, E_n$ defining $V = \vartheta(E_1, \cdots, E_n)$ is self-bounding if there are functions $\vartheta_i : \mathcal{E}^{n-1} \to \mathbb{R}$ for $i \in [1 : n]$ defining

$$V_i = \vartheta_i(E_1, \cdots, E_{i-1}, E_{i+1}, \cdots, E_n) \ ,$$

such that

$$0 \leq V - V_i \leq 1$$

for $i = 1, \cdots, n$ and

$$\sum_{i=1}^{n} V_i \leq V \ .$$

It turns out that for self-bounding functions, $\mathbb{V}\, V \leq \mathbb{E}\, V$, which explains their name. Details of this result, which is related to the *Efron-Stein inequal-*

ity, can be found in Section 4.2 of Lugosi (2004). It turns out that the supremum of a non-negative empirical process is self-bounding, but sumprema of general empirical processes are not. However, the result still yielded bounds on such suprema by considering squared variables (for details, see Massart, 2000).

Further work in Rio (2000, 2002) and Bousquet (2002a) resulted in confirmation of the conjecture by obtaining a functional equivalent of Bennett's inequality for the suprema of general empirical processes. In order to achieve this, the concept of self-bounding functions needed to be generalized to introduce new parameters. We present the general result below.[59]

Theorem 4.14 (Theorem 2.1 of Bousquet, 2002a). *Let $E, E'_1, E'_2, \cdots, E'_n$ be r.v.'s in $\mathcal{E}^n$, and $E_1, E_2, \cdots, E_n$ be r.v.'s in $\mathcal{E}^{n-1}$. Assume there is a $c > 0$ such that*

$$E'_i \leq E - E_i \leq 1$$

and

$$E'_i \leq c$$

almost surely, and furthermore that

$$\mathbb{E}\, E'_i \geq 0 \ \ .$$

Let ς almost surely satisfy $\varsigma^2 \geq \frac{1}{n}\sum_{i=1}^n \mathbb{E}(E'_i)^2$. Define $\nu = (1+c)\,\mathbb{E}\,E + n\varsigma^2$, $h(v) = (1+v)\ln(1+v) - v$, and $\psi(v) = e^{-v} - 1 + v$.

If

$$\sum_{i=1}^n (E - E_i) \leq E$$

almost surely, then

$$\ln \mathbb{E}\, e^{\lambda(E - \mathbb{E}\,E)} \leq \psi(-\lambda)\nu \tag{4.13}$$

for all $\lambda \geq 0$.

It follows that for all $\epsilon > 0$,

$$\mathbb{P}\{E \geq \mathbb{E}\,E + \epsilon\} \leq \exp\left(-\nu h\left(\frac{\epsilon}{\nu}\right)\right) \ \ , \tag{4.14}$$

and for all $v \geq 0$,

$$\mathbb{P}\left\{E \geq \mathbb{E}\,E + \sqrt{2\nu v} + \frac{v}{3}\right\} \leq e^{-v} \ \ . \tag{4.15}$$

[59] A weaker, but more generally applicable result is provided in Theorem 2.2 of Bousquet (2002a).

The key to understanding the improvement in this result is to compare the control on the m.g.f. of the centred variable in (4.13) to that for the basic Herbst argument in (4.11). The improved formulation captures the non-symmetric behaviour of the variables, with subexponential decay in certain regions instead of subgaussian decay everywhere.

Consider the case where all the random variables E'_i are identically zero, and $E = \vartheta(W_1, \cdots, W_n)$ for a self-bounding ϑ, and r.v.'s $W_1, \cdots, W_n$. If we set

$$E_i = \vartheta_i(W_1, \cdots, W_{i-1}, W_{i+1}, \cdots, W_n)$$

for each i (where the ϑ_i are specified by the definition of a self-bounding function), these variables meet the requirements of the theorem with $c = 0$ and $\varsigma^2 = 0$. Applying the bound in this case yields Boucheron et al. (1999, Theorem 1), so that this result is a generalization of that result to a wider class of functions. Furthermore the result in Boucheron et al. (1999) also provides a lower bound:

$$\mathbb{P}\{E \leq \mathbb{E}\, E - \epsilon\} \leq \exp\left(-\,\mathbb{E}\, E h\left(\frac{\epsilon}{\mathbb{E}\, E}\right)\right) \quad . \tag{4.16}$$

Bousquet also showed that if $E = \sup_{\phi \in \mathcal{V}} \sum_{i=1}^{n} \phi(W_i)$, where the functions in $\mathcal{V}$ have zero mean, are square-integrable, and are upper bounded by one[60], the theorem above can be applied with $c = 1$ and any $\varsigma^2 \geq \sup_{\phi \in \mathcal{V}} \mathbb{V}\, \phi(W_1)$, yielding bounds of

$$\mathbb{P}\{E \geq \mathbb{E}\, E + \epsilon\} \leq \exp\left(-[n\varsigma^2 + 2\,\mathbb{E}\, E] h\left(\frac{\epsilon}{n\varsigma^2 + 2\,\mathbb{E}\, E}\right)\right) \quad , \tag{4.17}$$

and

$$\mathbb{P}\left\{E \geq \mathbb{E}\, E + \sqrt{2x[n\varsigma^2 + 2\,\mathbb{E}\, E]} + \frac{x}{3}\right\} \leq e^{-x} \quad . \tag{4.18}$$

When $\mathcal{V}$ consists of a single function, the relaxed form of Bennett's inequality (Theorem 4.3), and the refined form of Bernstein's inequality (Theorem 4.5) are recovered from these results respectively. As a result, these concentration

[60] This assumption can be relaxed slightly. Again, the form of these results mean they are sensitive to scaling — see footnote 45.

inequalities are sometimes called the functional Bennett's or Bernstein's inequality respectively. The same bound also holds for

$$E = \sup_{\phi \in \mathcal{V}} | \sum_{i=1}^{n} \phi(W_i) |$$

if the functions are uniformly bounded by one.

Chapter 5

Training sample bounds

The dilemma with the test sample approach can be summarized in the following 3 points:

- more data for training generally means that a better hypothesis can be selected[61]; but
- more data for training means less data for testing; and
- less data for testing implies higher variance of point estimates of, and wider confidence intervals for, error rates.

There is thus a trade-off: one can, in general, either be more sure of worse performance, or less sure of better performance. This balancing act is the focus of the training sample bounds under investigation in the following chapters.

As mentioned before, it is in many cases possible to use the training sample to predict the performance of the hypothesis selected using that training sample. There is, of course, a dependence between the training sample and the chosen hypothesis, so one must work carefully. We shall once again begin with point estimators, and later discuss interval estimators.

[61] A larger training sample typically provides better information about the distribution generating data points.

Section 5.1 discusses the optimism of estimating true risk using apparent risk, and considers various classical techniques for taking this optimism into account, including bootstrap point estimators of varying degrees of sophistication. An overview of the related CV approach is given in Section 5.2.

Our consideration of interval estimators based on the training sample begins in Section 5.3, and Section 5.4 presents the most basic tool for constructing such interval estimators, the Occam's razor method.

Section 5.5 contains the core of the chapter: we begin by introducing the concept of covering numbers (Section 5.5.1), followed by the presentation of symmetrization lemmas (Section 5.5.3) and dual sample bounds (Section 5.5.4) for various measures of deviation. With these tools in place, we construct covering number bounds based on the various measures of deviation (Section 5.5.5). Next we present the random subsample lemma and bound (Section 5.5.6), and use them to obtain a bound on regular deviation (Section 5.5.7). We conclude the section with some results involving thresholded classes, which relate general loss functions to zero-one loss functions (Section 5.5.8).

Section 5.6 investigates obtaining bounds by employing dominating loss functions. Our primary focus in the section is on various types of margin bounds for thresholding classifiers.

In Section 5.7, we consider the chaining and generic chaining methodology, and apply the methodology to present an alternative to the random subsample bound.

The last part of the chapter considers the problem of obtaining the covering numbers necessary for applying the results presented. We present a number of dimension-like quantities in Section 5.8, and illustrate how they can be used to bound covering numbers. Finally, we consider methods for obtaining covering numbers for complex classes in terms of simpler classes, and other methods of obtaining covering numbers directly in Section 5.9.

5.1 Training sample point estimates

First, note that one should not simply use the apparent risk, $r_S(w_S)$, as an estimate of the true risk of a decision rule w_S in our setting: since the decision rule is selected based on the input-output pairs in S, we expect the decision rule to perform better on them than on unseen examples — in general, the apparent risk would thus be a (typically optimistic) biased estimate of the true risk (see, for example, Hastie et al., 2001, Section 7.4). We thus need a more sophisticated approach than simply using the apparent risk.

5.1.1 Optimism

The most obvious approach to overcoming the objection just raised is to estimate the risk of a decision rule by adjusting the apparent risk for the bias introduced by the decision rule's dependence on the sample. The extent to which the apparent risk understates the true risk of a decision rule can be broken up into a number of components. To do this, we write $w_{S'}$ for the decision rule selected by the algorithm $\mathit{\Theta}$ under consideration when a sample S' is chosen. Then, for a training sample S, we are interested in the optimism of the algorithm-sample pair $(\mathit{\Theta}, S)$ (w.r.t. the distribution D):

$$\begin{aligned}
\mathrm{op}_D(\mathit{\Theta}, S) &= r_D(w_S) - r_S(w_S) \\
&= [r_D(w_S) - r_D(\mathit{\Theta})] \\
&\quad + [r_D(\mathit{\Theta}) - \mathbb{E}_{S' \sim D^m} r_{S'}(w_{S'})] \\
&\quad + [\mathbb{E}_{S' \sim D^m} r_{S'}(w_{S'}) - r_S(w_S)] \ .
\end{aligned}$$

The first and last terms in this expression have mean zero, so that the mean deviation between apparent risk and true risk when employing $\mathit{\Theta}$ is

$$\begin{aligned}
\mathrm{op}_D(\mathit{\Theta}) &= r_D(\mathit{\Theta}) - \mathbb{E}_{S' \sim D^m} r_{S'}(w_{S'}) \\
&= \mathbb{E}_{S' \sim D^m} [r_D(w_{S'}) - r_{S'}(w_{S'})] \ ,
\end{aligned}$$

which we call the optimism of $\mathit{\Theta}$. It follows that $r_S(w_S) + \mathrm{op}_D(\mathit{\Theta})$ is an unbiased estimator of both $r_D(w_S)$ and $r_D(\mathit{\Theta})$.

This leads us to the problem of estimating $\mathrm{op}_D(\Theta)$. Let us write D_X for the marginal distribution of X, and S'_X and S'_Y for the inputs and outputs of S' (as well as the corresponding empirical distributions over $\mathcal{X}$ and $\mathcal{Y}$). Then we can expand $\mathrm{op}_D(\Theta)$ into

$$\mathbb{E}_{S'_X \sim D_X{}^m}\,\mathbb{E}_{S'_Y \sim D_{Y^m|S'_X}}\left[\begin{array}{r}\Big(r_D(w_{S'}) - r_{S'_X \times D_{Y^m|S'_X}}(w_{S'})\Big)\\ +\Big(r_{S'_X \times D_{Y^m|S'_X}}(w_{S'}) - r_{S'_X \times S'_Y}(w_{S'})\Big)\end{array}\right]$$

where $D_{Y^m|S'_X}$ denotes the distribution obtained by independently selecting each Y_i in S'_Y conditionally on $X = X_i$, the i-th component of S'_X.

The two terms in the expectation represent two distinct phenomena: consider a fixed sample S', on which Θ yields a decision rule $w_{S'}$. Then the second term represents the amount by which the apparent error rate understated the average error which would be obtained if $w_{S'}$ were tested on representative outputs for the given inputs (i.e. how much of the optimism of (Θ, S') is due to the specific values of the responses). We shall call this the response-driven optimism of (Θ, S'), and we shall call its mean the response-driven optimism of Θ. The first term indicates how much the "representative output" error rate for the given inputs understates the true error of $w_{S'}$ (i.e. how much of the optimism of (Θ, S') is due to the specific values of the predictors). We shall call this the predictor-driven optimism of (Θ, S'), and its mean the predictor-driven optimism of Θ.

5.1.2 In-sample optimism

It is important to note that optimism in the sense above is measured relative to the true risk of the decision rule (or the algorithm). It is common in other texts to use the term optimism for the conditional mean (with respect to $S'_Y \sim D_{Y^m|S'_X}$) of the response driven optimism of (Θ, S'). We shall call this concept the in-sample optimism of Θ at S'_X. The mean in-sample optimism is the response-driven optimism of Θ.

This distinction is important to note, since many techniques which adjust for optimism actually only make provision for in-sample optimism. Although these techniques are valuable in other fields (notably model selection in regression problems[62]), they are not directly relevant to our problem, so we shall restrict ourselves to mentioning a few popular techniques. With their roots in model selection, these techniques generally assume some form of model for the predictor-response relationship. These approaches have in turn fathered a number of variants with benefits in certain situations.

- Adjusting the mean *residual squared error*[63] (RSE) of a linear model, $\frac{\text{RSE}}{m}$, to take in-sample optimism into account, is common. The adjusted value employed is usually $\frac{\text{RSE}}{m-2v}$, where v is the number of regressors in the linear model (Efron and Tibshirani, 1993, Section 17.4). The resulting estimate of in-sample optimism is
$$\frac{\text{RSE}}{m-2v} - \frac{\text{RSE}}{m} = \frac{2v\,\text{RSE}}{m(m-2v)} \quad .$$
- Mallows' C_P, originally proposed in the 1960's as a visual aid for model selection, is now often used as an automated model selection criterion. The resulting estimate of in-sample optimism here is $\frac{2v\varsigma^2}{m}$, where ς^2 is an estimate of the variance of the residuals (losses). Using the unbiased estimate $\varsigma^2 = \hat{\sigma^2} = \frac{\text{RSE}}{m-v}$, one obtains a result similar to that from the adjusted mean RSE. Some relevant reading is Allen (1974), Hastie et al. (2001), Mallows (1973).
- The Akaike information criterion (AIC) (Akaike, 1974), named for Hirotsugu Akaike, is a generalized version of the C_P criterion, based on a log-likelihood instead of an estimate of the residual variance (Hastie et al., 2001).
- The Bayes (or Schwartz) information criterion (BIC) (Schwartz, 1979) is also popular, and as the name implies, arises from a Bayesian approach to model selection — see Hastie et al. (2001, Section 7.7). The estimates of in-sample optimism from the BIC grow more quickly with model complexity than those of AIC. Hastie et al. (2001, Section 7.8)

[62] In fact, most of these techniques are primarily for model selection, not adjusting for optimism to obtain a risk estimate, since any risk estimates obtained would not apply to the hypothesis or model selected on the basis of the criterion.

[63] This corresponds to the apparent risk under the squared error loss function.

> further discuss the relationship between the BIC and the minimum description length (MDL) approach to model selection. The MDL approach to estimating model complexity is closely related to the sample compression interval estimators we discuss in Section 6.5.

When we do not assume a model for relating predictor-response pairs, we are more interested in the optimism of Θ than in the in-sample optimism. Estimates of this optimism must generally be obtained numerically. The major technique employed for this purpose is the bootstrap.

5.1.3 The bootstrap

The bootstrap (Efron, 1979, 1983, 1986, 1992, Efron and Tibshirani, 1997) is based on the idea of inferring the statistical behaviour of a procedure in a population (underlying distribution or real world) from the procedure's behaviour on a sample (empirical distribution or *bootstrap world*). In some cases, the procedure's behaviour on the sample can be calculated directly, but usually the procedure's behaviour on the empirical distribution must be approximated by a kind of Monte Carlo approximation, involving taking a number of so-called *bootstrap samples*.

This Monte Carlo-style approach to the bootstrap is, of course, usually quite time-consuming.

The naïve bootstrap approach

Consider the following statistical procedure:

Procedure 5.1. *Select a training sample S of size m from a population (with underlying distribution D), and on the basis of S, select a decision rule w_S. Calculate the expected risk over D of w_S, $r_D(w_S)$.*

Note that we can not perform the last step of this procedure, since we do not know D.

The corresponding procedure in the bootstrap world would be:

Procedure 5.2. *Select a training sample (with replacement) $S^\star$ of size m from the bootstrap world S.*[64] *On the basis of $S^\star$, select an hypothesis $w_{S^\star}$. Calculate the expected risk, over the distribution S, of $w_{S^\star}$, i.e. $r_S(w_{S^\star})$.*

Clearly, we can perform this entire procedure.

The most naïve bootstrap approach to estimating the true risk $r_D(w_S)$ would be to simply use $r_S(w_{S^\star})$ as an estimator. However, this estimator depends on the specific bootstrap sample, which provides an undesirable source of variance. Since all bootstrap samples are equally likely, it makes sense to use the average bootstrap prediction as the estimate, i.e.

$$\widehat{r_D(w_S)} = \mathbb{E}_{S^\star \sim S^m}\, r_S(w_{S^\star}) \ .$$

Calculating the expectation of the right hand side exactly is usually not feasible. However, a Monte Carlo estimate of this expectation also helps reduce the variability in the estimate of the expected error, yielding the (still) naïve bootstrap estimator,

$$\widehat{r_D(w_S)} = \frac{1}{B} \sum_{b=1}^{B} r_S(w_{S^{\star b}}) \ ,$$

where the $S^{\star b}$ indicate the bootstrap samples for each of the B Monte Carlo replications.[65]

Unfortunately, this naïve approach does not work very well. There are a number of reasons for this, and we address them in turn, thus building a better bootstrap estimator of true risk.

Estimating optimism with the bootstrap

Note that the approach described above does not in fact attempt to adjust for optimism, but rather tries to estimate the error directly. Making use of a

[64] i.e. sampling is done from the empirical distribution generated by the points in S.

[65] For practical work, some more variation may often be eliminated by using *balanced* bootstrap samples, as proposed by Davison et al. (1986), resulting in a more efficient estimator for a fixed number of replications. Adjusting the process to utilise this idea is usually rather straightforward, so we shall not elaborate on the method further.

bootstrap estimate of the optimism is the next refinement for our bootstrap approach. This modification removes some of the variability between the Monte Carlo replications. Intuitively, the idea is that this approach will work better because we extract more information from the real world directly, and use bootstrapping on a smaller portion of the problem.[66]

Basically, the last step of the (real-world) Procedure 5.1 is modified to "Calculate the expected error over D of w_S as the sum of the apparent error $e_S(w_S)$ and the optimism of (Θ, S), $\mathrm{op}_D(\Theta, S) = r_D(w_S) - r_S(w_S)$". We still can not perform this step directly, though.

However, in the bootstrap world, we can calculate the corresponding optimism:

$$\mathrm{op}_S(\Theta, S^\star) = r_S(w_{S^\star}) - r_{S^\star}(w_{S^\star}) \ , \tag{5.1}$$

and the two terms on the right are the bootstrap world true error and apparent error of an hypothesis chosen from a bootstrap sample.

Once again, we could use this value as the estimate of the optimism of (Θ, S), but we can eliminate variance by using the expected value over all possible bootstrap samples. Again, this value usually can not be obtained directly, but must be approximated using a Monte Carlo estimate. This results in the estimate of optimism

$$\widehat{\mathrm{op}_D(\Theta, S)} = \frac{1}{B}\sum_{b=1}^{B} \mathrm{op}_S\left(\Theta, S^{\star b}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(r_S\left(w_{S^{\star b}}\right) - r_{S^{\star b}}\left(w_{S^{\star b}}\right)\right) \ ,$$

which leads to the true error estimate of w_S,

$$\widehat{r_D(w_S)} = r_S(w_S) + \widehat{\mathrm{op}_D(\Theta, S)} \ .$$

The .632 bootstrap estimator

In a typical problem with an infinite, or very large, population, the set of training input points, $S_X = \cup_{i=1}^{m}\{x_i\}$, typically has measure zero, or close to it, relative to the marginal distribution over $\mathcal{X}$ defined by the distribution

[66] A full theoretical explanation can be found in Efron and Tibshirani (1993, Section 17.6).

D. On the other hand, in the bootstrap world, the set of input points in any bootstrap sample, $S^\star_X$, has a distinct non-zero measure relative to the marginal distribution over $\mathcal{X}$ defined by the distribution S. Specifically, the measure of $S^\star_X$ is $\frac{1}{m}|\{x \in S_X : x \in S^\star_X\}|$. It turns out that the mean measure of a bootstrap sample is $1-(1-\frac{1}{m})^m$. As m becomes large, this value converges to $1-\frac{1}{e}=0.632$.

This observation is the key to the next refinement of the bootstrap estimator. Our intuition tells us that, because of this higher degree of "overlap" between the bootstrap sample and the bootstrap population (the original sample), compared to that of the original sample with the population, overfitting in the bootstrap world is likely to be more severe, thus yielding too-high estimates of optimism, and thus of true error.

The solution proposed as the .632 bootstrap estimator is to use an estimate of the risk on points not in the training sample to obtain an estimate of optimism. However, since the points not in the training sample *in the bootstrap world* are typically further away from the points in the training sample, than is the case in the real world, this approach, if applied naïvely, will typically overestimate the optimism. A compromise is reached by scaling this estimate of optimism to account for the difference in average distance between the training sample and test points in the bootstrap and real world. The scaling factor turns out to be 0.632, due to the argument at the beginning of this section.[67]

More precisely, for a bootstrap sample $S^\star$, we consider the average risk of $w_{S^\star}$ on the points in the bootstrap world not in $S^\star$, i.e.[68]

$$r_{S\setminus S^\star}(w_{S^\star}) = \frac{1}{|S \setminus S^\star|} \sum_{(x,y)\in S\setminus S^\star} L(w_{S^\star}(x), y)) \ .$$

This can be estimated by Monte Carlo replications. However, even using balanced samples, another undesirable source of variation is present here: the size of $S \setminus S^\star$ can change for each bootstrap sample, giving the terms

[67] A full derivation of this estimator can be found in Efron (1983).

[68] We ignore the potential problem arising with this and later formulae if S contains duplicate points, as it can easily be circumvented with a slightly more cumbersome notation.

in the Monte Carlo average different variances. This problem is resolved by instead calculating[69]

$$\epsilon_0(\Theta, S) = \frac{1}{\sum_{b=1}^{B} |S \setminus S^{\star b}|} \sum_{b=1}^{B} \left| S \setminus S^{\star b} \right| r_{S \setminus S^{\star b}}(w_{S^{\star b}}) \ .$$

Now, $\epsilon_0(\Theta, S) - e_S(w_S)$ can be seen as a kind of out-of-sample optimism of Θ in the bootstrap world. As mentioned before, this optimism needs to be scaled to take into account the difference between testing in a bootstrap world and testing in the real world. With the scaling factor, we obtain the .632 bootstrap estimator,

$$\widehat{r_D(w_S)} = r_S(w_S) + 0.632(\epsilon_0(\Theta, S) - r_S(w_S)) = 0.368 r_S(w_S) + 0.632\epsilon_0(\Theta, S) \ .$$

A more thorough treatment of the .632 bootstrap estimator is in Section 17.6 of Efron and Tibshirani (1993)

The .632+ bootstrap estimator

In Efron and Tibshirani (1993), it is reported that the .632 estimator had the best performance among a number of alternatives, including CV, on the studies which had been performed until then. Although it is generally a good estimator, it was later found that the performance of this estimator behaves undesirably when overfitting is liable to occur. In cases like these, an adjustment must be made to prevent overoptimistic error estimates. This is the focus of the .632+ bootstrap estimator, introduced in Efron and Tibshirani (1997).

To understand this estimator we consider a modification of the distribution D: denote by D' the distribution over $\mathcal{Z}$ with the same marginal distributions as D for X and Y, but where X and Y are independent.

Consider now the *no-information risk* of w_S, $r_{D'}(w_S)$. The name derives from the fact that this is the error rate of w_S when the predictor X contains

[69] Note that this formula assumes no bootstrap sample contains all the points in the training set. If there are such samples, they are discarded, and B is reduced accordingly.

no information about the response Y.[70] Suppose $\epsilon_0(\Theta, S)$ is very close to $r_{D'}(w_S)$ (relative to $r_S(w_S)$). It would then seem reasonable to assume that X does not contain "much" information about Y, so that the relatively low value of $r_S(w_S)$ is mainly due to overfitting.

On the other hand, if ϵ_0 is instead closer to $r_S(w_S)$ (relative to $r_{D'}(w_S)$), it seems to imply that the value of $r_S(w_S)$ is mainly due to it encoding the actual information about Y present in X, indicating a low level of overfitting.

The intuition behind the 0.632+ estimator, then, is that the apparent risk, $r_S(w_S)$, becomes less reliable the more overfitting is present, so we should rely more on $\epsilon_0(\Theta, S)$ in such cases. This is accomplished by a measure of relative overfitting[71],

$$\hat{R} = \frac{\epsilon_0(\Theta, S) - r_S(w_S)}{r_{D'}(w_S) - r_S(w_S)} \ .$$

This relative overfitting rate is then used in replacing the constant 0.632 by $\omega = \frac{0.632}{1-0.368\hat{R}}$, resulting in the estimator

$$\widehat{r_D(w_S)} = r_S(w_S) + \omega(\epsilon_0(\Theta, S) - r_S(w_S)) = (1-\omega) r_S(w_S) + \omega\epsilon_0(\Theta, S) \ .$$

We now turn our attention to CV, which seemingly attempts to estimate the true error directly, rather than adjusting for optimism.

5.2 Cross-validation

The concept of CV originally referred to the idea of validating a model built from a training sample on an independent test sample, based on the realization that apparent error was optimistic (Stone, 1974). Modern usage of the term however, refers to partitioning a data set (either the full set for model assessment, or the training sample for model selection) repeatedly in order to make more efficient use of the data.

[70] This is not just an intuitive idea — the term is derived from information theory.

[71] In practice, some minor modifications are made to this formula to ensure $\hat{R}$ lies in $[0, 1]$. See Efron and Tibshirani (1997, Section 3) for details.

The first form of the modern approach to CV was introduced to statistics in the 1950's by Maurice Quenouille, and christened as *the jackknife* by Tukey. Central to this idea is the concept of a *jackknife sample*: a jackknife sample of a data set is simply the data set with one data point removed. The idea is to model the behaviour of a statistical procedure on the full data set by its behaviour on the jackknife samples. The idea is clearly similar to that of the bootstrap[72].

More specifically, considering our problem of estimating the risk of a decision rule selected based on a training sample, we denote the i-th jackknife sample (and the corresponding empirical distribution) by $S_{\setminus i}$ — the training sample with the i-th data point removed. The jackknife approach thus estimates the risk of w_S as the average of the risks of each $w_{S_{\setminus i}}$ on the corresponding deleted data point (x_i, y_i), i.e.

$$\widehat{r_D(w_S)} = \frac{1}{m} \sum_{i=1}^{m} L(w_{S_{\setminus i}}(x_i), y_i) \ .$$

The jackknife has further applications beyond the scope of this work, but this approach to risk estimation forms the foundation of modern CV. This approach is also called, for obvious reasons, leave-one-out CV (LOO-CV), and the risk estimate is sometimes called the *deleted estimate* (Devroye et al., 1996).

Since the jackknife methodology involves *m jackknife iterations*, i.e. performing the entire statistical procedure under consideration for each jackknife sample, it can be very time-consuming. It is often totally infeasible to calculate jackknife estimates for complex statistical procedures. In our scenario, this typically includes cases where the process for selecting a decision rule from a training sample is already quite computationally intensive. This problem can sometimes be solved by clever tricks which allow one to modify the decision rule selected for one sample to obtain the decision rule which would be selected for a slightly modified sample. An example of this is fitting a least squares regression model (Efron and Tibshirani, 1993). Generally,

[72] However, for simple applications of the jackknife, exact results can usually be obtained, rather than using Monte Carlo approximation techniques.

however, an alternative approach, representing a trade-off between accuracy and computation time, must be used.

The most popular and widely-used such alternative is K-fold CV: when Efron and Tibshirani introduced the .632+ bootstrap estimator, they noted that until then, K-fold CV was the "traditional method of choice" for estimating generalization error (Efron and Tibshirani, 1997). In this setting, the training sample is partitioned into K (usually) equal-sized[73] subsets or folds S_{Fi}[74]. We define the i-th *CV sample* as $S_{\setminus Fi} = S \setminus S_{Fi}$. Then the modified approach involves estimating $r_D(w_S)$ by the average of the performances of the decision rules selected on the basis of the CV samples (as evaluated on the omitted fold). The main advantage of this approach is that there are only K possible CV samples, instead of m. Thus, in many cases performing K-fold CV is much more feasible than LOO-CV when $K \ll m$. For m equal-sized folds, we recover the jackknife estimate, so that LOO-CV is sometimes called m-fold CV.

The computational advantage of this alternative has been made clear. There is, however, a corresponding loss in accuracy in employing K-fold CV. This derives from the fact that the CV samples have average size $\frac{K-1}{K}m$. Thus, for smallish K, the hypotheses selected from CV samples are based on less data, and are thus typically less accurate than w_S. This results in pessimistic error estimates[75]. For LOO-CV, the jackknife samples are of size $m-1$, so for large m, this bias is very small, so that LOO-CV estimates are typically said to be "nearly unbiased".

The value of K also affects the quality of CV estimators in another way. For K near m, the CV samples are almost identical to each other and to the complete sample. This means that the CV iterations (analagous to jackknife iterations) have very similar outcomes for a given training sample. However, this similarity means that such estimators are subject to almost all

[73] In general practice and reporting, subsets are of equal size (or within one element of each other) unless otherwise noted.

[74] Here F denotes *fold*.

[75] This is not always so serious, depending on how quickly the error is decreasing as the sample size increases (see the conclusions of Efron and Tibshirani, 1997, and Hastie et al., 2001, Section 7.10).

the variance implied by using only the training sample to estimate the risk. The lower level of overlap between CV samples and the training sample for smaller values of K, may lead to more biased estimates, but the more diverse outcomes of the CV iterations result in a lower variance of the estimator. This idea was developed, based on empirical observations, in Efron and Tibshirani (1997) and Hastie et al. (2001).

Despite the apparent differences in approach between the bootstrap and CV, Efron and Tibshirani (1997) points out that the .632 bootstrap estimator can be seen as a smoothed version of the LOO-CV estimator. The result of this smoothing is a variance reduction at the cost of increased bias. The authors also show that the .632+ bootstrap estimator performs bias reduction on the .632 estimator, and show that this estimator outperforms cross-validation on a variety of experiments. These experiments are restricted to zero-one loss functions, however. In the case of zero-one and other discontinuous losses, LOO-CV may suffer from a high variance for certain algorithms, and thus a smoothed estimate may be preferable.

5.3 Training sample interval estimators

We now turn our attention to constructing interval estimators for $r_D(w_S)$ and $r_D(\Theta)$ from the training sample S. Our approach can be broadly stated as follows: on the basis of S, we hope to be able to construct a region $A(S, w)$ for each decision rule w in a subset $\mathcal{W}^\star(S)$ of the decision class $\mathcal{W}$ such that, with high confidence, every region contains the true risk of the corresponding decision rule simultaneously.

Thus, we seek a statement of the form

$$\mathbb{P}_{S\sim D^m}\left\{\forall w \in \mathcal{W}^\star(S) : r_D(w) \in A(S, w)\right\} \geq 1 - \delta \ \ .$$

If $w_S \in \mathcal{W}^\star(S)$, it follows that $A(S, w)$ is a confidence region for $r_D(w_S)$ with coverage at least $1-\delta$.[76] As in Section 3.2.1, the expression $r_D(w) \in A(S, w)$

[76] If $w_S \notin \mathcal{W}^\star(S)$, we will have to make do with the confidence interval $[0, 1]$.

will often be represented by $\psi\left(r_D(w), r_S(w)\right) \leq \epsilon(S, w)$, where ψ is some measure of deviation.

We distinguish two major groups of bound: those for which ϵ or the class $\mathcal{W}^\star(S)$ depends on S, and those for which they do not. In the second case, $A(S, w)$ and the resulting confidence interval only depend on the sample through $r_S(w)$. Otherwise, the resulting interval can depend on the sample in more complicated ways. In general, if $\mathcal{W}^\star(S)$ is not dependent on S, it is difficult to guarantee that $w_S \in \mathcal{W}^\star(S)$, unless one selects $\mathcal{W}^\star(S) = \mathcal{W}$. This is a common choice for many approaches to deriving bounds, including all the bounds we consider in this chapter.

Those bounds for which ϵ or $\mathcal{W}^\star(S)$ do depend on S are often called *data-dependent bounds* (even though in all cases $A(S, w)$ is data-dependent). As an example, $\mathcal{W}^\star(S)$ may be the set of decision rules which could be selected on a permutation of S by a specific algorithm Θ.[77] In this case, the set $\mathcal{W}^\star(S)$ is generally much smaller than $\mathcal{W}$, but the dependence of $\mathcal{W}^\star(S)$ on S makes deriving bounds for this case much more difficult. In general, it is preferable that $\mathcal{W}^\star(S)$ be as small as possible, as that may allow the regions $A(S, w)$ to be smaller.

Examples of data-dependent bounds are the margin bounds of Section 5.6 and the various bounds discussed in Chapter 6.

Much of the classical work on bounds has focused on the case where ϵ is not dependent on w, so-called *uniform bounds*. In this case, our statement above can be rewritten as

$$\mathbb{P}_{S \sim D^m}\left\{\forall w \in \mathcal{W}^\star(S) : \psi\left(r_D(w), r_S(w)\right) \leq \epsilon(S)\right\} \geq 1 - \delta \ .$$

5.4 The Occam's razor method

In this section, we consider the case $\mathcal{W}^\star(S) = \mathcal{W}$, with $\mathcal{W}$ countable. In that case, this method allows one to convert any collection of test set interval

[77] This may seem contrived, but such constructions are used in the algorithm-specific bounds presented in later chapters.

estimators for $r_D(w)$, $w \in \mathcal{W}$, into a training sample interval estimator for $r_D(w)$, $w \in \mathcal{W}$, and thus also for $r_D(w_S)$.

The Occam's razor method employs the Bonferroni inequality to generate a confidence region for the sequence of decision rule risks in a countable decision class. Core to the weighted union bound is the understanding that a confidence region represents a probability statement in the sample space, $\mathcal{Z}^m$: if $A(S)$[78] is a $100(1-\delta)\%$ confidence region for a parameter t, then it means that[79]

$$\mathbb{P}_{S\sim D^m}\left\{t \in A(S)\right\} \geq 1-\delta \ .$$

We can combine a number of such probability statements by means of the Bonferroni inequality to obtain a confidence region for all the decision rule risks simultaneously. Suppose the decision class $\mathcal{W}$ is countable. For each decision rule $w \in \mathcal{W}$, we can derive a $100(1-\delta(w))\%$ confidence interval for the risk of w, $A(S,w)$ using any of the methods outlined in the section on test sample interval estimators.

Suppose that we select the $\delta(w)$ such that $\sum_{w\in\mathcal{W}} \delta(w) = \delta$. Then, for any w,

$$\mathbb{P}_{S\sim D^m}\left\{r_D(w) \notin A(S,w)\right\} \leq \delta(w) \ ,$$

so the probability of any one of these events occurring,

$$\begin{aligned}\mathbb{P}_{S\sim D^m}&\left\{\exists w \in \mathcal{W} : r_D(w) \notin A(S,w)\right\} \\ &= \mathbb{P}_{S\sim D^m}\left\{\bigcup_{w\in\mathcal{W}} \left\{S : r_D(w) \notin A(S,w)\right\}\right\} \ , \qquad (5.2)\end{aligned}$$

by the Bonferroni inequality, does not exceed

$$\begin{aligned}\sum_{w\in\mathcal{W}} \mathbb{P}_{S\sim D^m}\left\{r_D(w) \notin A(S,w)\right\} &\leq \sum_{w\in\mathcal{W}} \delta(w) \\ &= \delta \ .\end{aligned}$$

we obtain that

$$\mathbb{P}_{S\sim D^m}\left\{\forall w \in \mathcal{W}^\star(S) : r_D(w) \in A(S,w)\right\} \geq 1-\delta \ ,$$

[78] Note that a confidence region is a function of the sample used to calculate it.

[79] Assuming that the interval does not exhibit undercoverage.

as desired.

It may seem at first that the weighted union bound approach won't be very useful: after all, if we have to increase the confidence level for the confidence interval we construct for each decision rule substantially, it would seem the resulting bounds on the risk will be too wide to be useful. This may be the case for very large and infinite decision classes, when we have no idea which decision rules are more likely to be correct. However, in the finite case, when the decision class is not too large, the resulting interval estimator is not unmanageably wide, if we base our interval on the sample mean. This is because the sample mean has the desirable property of being concentrated around its expectation, the true mean, as discussed earlier. This means that a sample mean is very likely to be close to the true mean, and the probability of large deviation between the two values is very small: in fact, it drops off at an exponential rate as the deviation increases. This can be readily verified by considering the test set interval estimators as well as the results on risk concentration discussed in earlier sections.

For extremely large and infinite decision classes, we use the technique of assigning prior weights outlined below to obtain useful bounds. We shall, however, need more sophisticated techniques to handle the uncountable decision classes, which are not covered by the Occam's razor method.

5.4.1 Assigning 'prior' weights

It is natural to consider the problem of selecting the values $\delta(w)$ above. Of course, they must be selected to ensure a certain pre-specified confidence for the simultaneous bound. In the case where $\mathcal{W}$ is finite, and one desires a final confidence level of $1 - \delta$, one can simply use seperate confidence intervals for each decision rule with all $\delta(w) = \frac{\delta}{|\mathcal{W}|}$. Of course, we are not restricted to this option, but it seems sensible when we have no prior indication that one of the decision rules is more likely than the others.

Note that, once we have selected a decision rule w by some algorithm, we are only concerned with the length of $A(S, w)$; the lengths of all the other confidence intervals used in designing the simultaneous probability statement generally becomes immaterial to us. In addition, the confidence level in the original interval, $1-\delta(w)$, no longer applies — instead, the confidence level is $1-\delta$, the value assigned to the complete probability statement. Effectively, $\delta-\delta(w)$ represents "lost confidence" — the price we pay to be sure that our bound will apply regardless of which decision rule an algorithm selects.

Let us, for a moment, consider the problem from a Bayesian perspective: suppose the probability (measure of belief) that w will be the decision rule selected by an algorithm is $\alpha(w)$. Let us consider the question: what division of confidence will result in the least "lost confidence" on average.

Given a combined confidence level $1-\delta$, we wish to find which choice of $\delta(w)$, summing to δ, minimizes

$$\sum_{w_1 \in \mathcal{W}} \left[\alpha(w_1) \sum_{w_2 \neq w_1} \delta(w_2) \right]$$

Now, since

$$\sum_{w_2 \neq w_1} \delta(w_2) = \delta - \delta(w_1) \ \ ,$$

we can use

$$\sum_{w_1 \in \mathcal{W}} \alpha(w_1) = 1$$

to see that this is equivalent to maximizing

$$\sum_{w_1 \in \mathcal{W}} \alpha(w_1)\delta(w_1) \ \ .$$

But this is simply the problem of maximizing an inner product. Since the maximum inner product is achieved when the sequences are in the same direction, we have that $\delta(w) = c\alpha(w)$. The restriction on the $\delta(w)$'s summing to δ imply that $c = \delta$. In this framework, it follows that the best way to allocate confidence levels to individual decision rules is in proportion to their prior probability of selection. We do not have to use this allocation of course, even though deviating from it may be suboptimal in the Bayesian sense above.[80]

Motivated by this discussion, we will refer to a distribution function corresponding to the assignment of weights to decision rules as a "prior". This is because it generally reflects our beliefs about the suitability of each decision rule. This idea of a "prior" will be central to a number of the bounds we will discuss later, notably the shell decomposition and PAC-Bayesian bounds presented in Chapter 8.

5.4.2 Applying the Occam's razor method

This section will present some applications of the Occam's razor method, including the most common forms available in the machine learning literature.

Example 5.1 Occam's razor binomial interval. For a zero-one loss function, the tightest upper interval on $e_D(w)$ with guaranteed coverage $1 - \delta$ for a given decision rule w is the upper binomial interval,

$$[0, \mathrm{UBT}\left(e_T(w), k, 1 - \delta\right)] \quad .$$

The Occam's razor method suggests using the regions

$$A(S, w) = [0, \mathrm{UBT}\left(e_S(w), m, 1 - \delta(w)\right)]$$

for each w.

In this case, if the algorithm selects the decision rule w_S, the resulting $100(1 - \delta)\%$ interval estimator for $e_D(w_S)$ is

$$[0, \mathrm{UBT}\left(e_S\left(w_S\right), m, 1 - \delta(w_S)\right)] \quad .$$

A similar result to the interval above can be obtained for lower intervals. Two-sided bounds can also be obtained, but they can be improved by using, for example, the two-sided exact LR interval.

When all the $\delta(w_S)$ are selected equal, the bound employed on the binomial tail deviation is independent of w, so that the result is a uniform interval.

□

[80] Further, optimizing the "lost confidence" is not necessarily the best approach. Another approach is to choose the "prior" $\delta(w)$ to minimize the expected value of the bound, subject to an assumption on the distribution of the selected decision rule. For a short synopsis of this approach, see Langford and Blum (1999, Section 2.1).

The scenario presented in the next example is known variously as the *optimistic*, *realizable* or *consistent* case in machine learning. This case deals with the situation when a decision rule with zero error on the training sample can always be found. The result here is based on generalizing the realizable interval of Section 3.2.12.

Example 5.2 Occam's razor realizable intervals. Suppose for a given problem, it is known that a finite decision class $\mathcal{W}$ contains at least one decision rule w_0 which has $e_D(w_0) = 0$. It follows that for any sample S, there is at least one decision rule with zero empirical error. Consider any algorithm $\varTheta$ which always selects some decision rule w_S with $e_S(w_S) = 0$. It follows from the uniform Occam's razor binomial interval that a $100(1-\delta)\%$ confidence interval for $e_D(w_S)$ (and also for $e_D(\varTheta)$) is

$$\begin{aligned}\left[0, \mathrm{UBT}\left(0, m, 1-\frac{\delta}{|\mathcal{W}|}\right)\right] &= \left[0, \sqrt[m]{\frac{\delta}{|\mathcal{W}|}}\right] \\ &\subseteq \left[0, \frac{1}{k}\ln\frac{|\mathcal{W}|}{\delta}\right] ,\end{aligned}$$

with the last line following from (3.7).

Non-uniform and lower intervals (when $r_S(w) = 1$) can be derived similarly. □

Note that the realizable bound for general loss functions derived from the upper Chernoff inequality in Section 4.6.1 could also be employed to obtain realizable bounds for general loss functions. In both cases, we are actually employing the Occam's razor method for the binomial interval (upper AV interval) to the entire decision class, but the stated result restricts our attention to the case where $r_D(w) = 0$.

In contrast to the realizable case, the *agnostic case* considers the more general situation where one does not know *a priori* whether an hypothesis with zero error exists, or can be found, in the decision class (hence their name).

It is common to differentiate between two scenarios in the *agnostic* case:

1. the *pessimistic* (worst case) scenario: we do not assume that we shall wish to evaluate a good decision rule in the decision class;

2. the *realistic*[81] case: we are unsure whether good performance can be achieved by a decision rule or not, but focus on good decision rules for our bounds/interval estimators.

However, in most cases, bounds for the realistic case are simply based on inverting special cases of bounds holding for the pessimistic case (see the discussion of machine learning Chernoff inequalities in Section 4.6.1). As such, we will focus almost entirely on bounds for the pessimistic case.

The following example presents a bound for the pessimistic case:

Example 5.3 Occam's razor Hoeffding's tail interval. Hoeffding's tail interval (4.3) applies to general loss functions, allowing us to construct confidence intervals for risk. This interval suggests using

$$A(S,w) = \left[0, r_S(w) + \sqrt{\frac{\ln \frac{1}{\delta(w)}}{2m}}\right]$$

in the application of the Occam's razor method.

The resulting $100(1-\delta)\%$ interval estimator for $r_D(w_S)$ is

$$\left[0, r_S\left(w_S\right) + \sqrt{\frac{\ln \frac{1}{\delta(w_S)}}{2m}}\right] .$$

Due to the symmetry of Hoeffding's tail inequality, lower and two-sided intervals of a similar form follow easily. Again, a constant choice for all $\delta(w_S)$ yields a uniform interval. □

As discussed in Section 4.7, the Hoeffding tail inequality does not incorporate knowledge of the variance of the loss of a decision rule. We can also obtain narrower intervals for good decision rules at the expense of wider intervals for poor decision rules by employing the Angluin-Valiant interval.

Example 5.4 Occam's razor Angluin-Valiant interval. In Section 4.7, we noted that the upper Angluin-Valiant interval generally outperforms the upper Hoeffding's tail interval for $r_D(w) \leq 0.25$, but is generally poor when $r_D(w)$ is larger.

[81] Traditionally, this has been called the *general* case - see, for example, Vapnik (1998). We have avoided this usage to allow the use of "general" in the text with its regular meaning.

To apply the Occam's razor method to this interval, we use

$$A(S,w) = \left(0, r_S(w) - \frac{\ln\delta(w)}{m}\left(1+\sqrt{1-\frac{mr_S(w)}{\ln\delta(w)}}\right)\right]$$

obtaining the $100(1-\delta)\%$ interval

$$\left(0, r_S(w_S) - \frac{\ln\delta(w_S)}{m}\left(1+\sqrt{1-\frac{mr_S(w_S)}{\ln\delta(w_S)}}\right)\right] \ .$$

□

Note that both bounds above can be improved by applying the Occam's razor method to the Hoeffding's r.e. interval. In the case of zero-one loss, the exact-variance Bernstein interval for error will also provide an improvement.

PAC bounds

Suppose $\mathcal{W}$ is finite, and we are working with a uniform interval, so $\delta(w) = \frac{\delta}{|\mathcal{W}|}$. Inherent in applying the Occam's razor method to Hoeffding's tail interval in this case is the probability statement

$$\mathbb{P}_{S\sim D^m}\left\{\forall w \in \mathcal{W} : r_D(w) \in \left[0, r_S(w) + \sqrt{\frac{\ln\frac{|\mathcal{W}|}{\delta}}{2m}}\right]\right\} \geq 1-\delta \ .$$

Another way of stating this result is: with probability at least $1-\delta$, for all $w \in \mathcal{W}$,

$$r_D(w) \leq r_S(w) + \sqrt{\frac{\ln|\mathcal{W}| - \ln\delta}{2m}} \ .$$

This formulation is popular in the machine learning community, where such statements are known as PAC bounds. This name was given by Dana Angluin to the bounds presented in Leslie Valiant's seminal paper (Valiant, 1984) in computational learning theory, which were of a similar nature.

The focus in machine learning is strongly on providing such probabilistic upper bounds (hence, upper intervals) on the risk of a specific decision rule selected by an algorithm Θ. In the corresponding work in classical statistics, the focus is often more on $e_D(\Theta)$ than $e_D(w)$, but the focus on upper one-sided results remains, as we shall see.

We note that uniform Occam's razor results actually provide more than we typically need: we typically only desire an interval for the risk of the specific decision rule selected, or the algorithm under consideration, while the Occam's razor method provides simultaneous bounds on every decision rule in $\mathcal{W}$.

Sample complexity

We next consider the question: for an algorithm Θ, how large must the sample size m be to be confident that $r_D(w_S) \leq r_S(w_S) + \epsilon$ for any hypothesis? Setting

$$\sqrt{\frac{\ln \frac{|\mathcal{W}|}{\delta}}{2m}} = \epsilon \ ,$$

we obtain that $r_D(w) \leq r_S(w) + \epsilon$ for all $w \in \mathcal{W}$ (and hence also for w_S) with probability at least $1 - \delta$, if

$$m(\epsilon, \delta) = \frac{\ln |\mathcal{W}| - \ln \delta}{2\epsilon^2} \ .$$

Computational learning theorists often refer to the function $m(\epsilon, \delta)$ as the *sample complexity* of Θ (with respect to the specific interval or bound). Generally, one can consider any such $m(\epsilon, \delta)$ as an upper bound on the minimum number of samples $m_0(\epsilon, \delta)$ needed to ensure $r_D(w_S) \leq r_S(w_S) + \epsilon$ with probability $1 - \delta$. This $m_0(\epsilon, \delta)$ is the *true sample complexity.*

Computational learning theorists are generally interested in the asymptotic behaviour, such as the growth rate, of sample complexities, and less interested in the precise constants in $m(\epsilon, \delta)$. As such, many results in the literature are presented in *order notation.* In this example, we would say the sample complexity is $O(\epsilon^{-2} \ln \frac{1}{\delta})$. From a sample complexity result with specified constants, one can in turn obtain an interval estimator for $r_D(w_S)$.

The variety of approaches to presenting results can make navigating the literature of machine learning and computational learning theory, as well as comparing results in different forms, rather confusing. In many cases, order results on sample complexity suppress constants which can be retrieved by studying the proofs of the results. This work will generally *not* present sample complexity results.

5.5 Bounds using covering numbers

The Occam's razor method can not be directly applied to uncountably infinite decision classes.[82] In order to allow bounds on such decision classes, we shall next apply the Occam's razor method to a finite class of decision rules which approximate all the decision rules in the complete decision class $\mathcal{W}$. This approach also allows us to improve the Occam's razor intervals for extremely large and countably infinite decision classes. The finite approximating class will be known as a *cover* of $\mathcal{W}$. The Occam's razor method is then applied uniformly over the elements of the cover, and some adjustments are made to the result to fit it to the original decision class.

In order to apply this approach, it is generally necessary to restrict ourselves to the modified learning problem described in Section 2.3.1. Thus, in what follows, we assume a bounded loss, so that $\mathcal{Y} = \mathcal{A} = [0, 1]$, g is the identity function on $[0, 1]$, $\mathcal{X}$ represents the space of predictor-response pairs, with a specific pair denoted by x. In addition $\mathcal{H} = \mathcal{W}$ is a class of real-valued functions mapping into $[0, 1]$ corresponding to the loss class for an underlying problem (so the mapping is into $\{0, 1\}$ for zero-one loss functions). Furthermore we have $L(w(x), v) = w(x)$ for all v. One effect of this approach is that when the resulting bounds are applied in practice, various quantities expressed in terms of the modified framework need to be transformed to the

[82] As pointed out in Langford and McAllester (2004), however, we note that all computer implementation of algorithms effectively employ finite decision classes, even if the classes are very large. This is because the continuous parameters are represented by a finite set of bits, effectively discretizing the parameter space. This has two consequences: first, bounds based on *exact* minimization of risk on the training sample may not theoretically hold for the selected decision rule (although this effect is almost always negligible); second, bounds such as shell decomposition bounds may be able to use exponential decay of the shell sizes to obtain reasonable results even for continuous hypothesis classes (although such results have not yet been reported, examples of using Occam's razor bounds based on 8-bit and 32-bit representations of the parameter space appear in Langford and McAllester, 2004).

original framework. The most dominant such example is the decision class $\mathcal{W}$, which corresponds to the loss class in the original problem.

5.5.1 Pseudometrics and covering numbers

In order to apply this approach, we need to specify when one decision rule approximates another. This is typically done by means of a pseudometric[83] d: if $d(w_1, w_2) < \gamma$, we say w_2 γ-approximates w_1 (with respect to d). If, on the other hand, $d(w_1, w_2) \geq \gamma$ we say that w_1 and w_2 are γ-distinguishable. Various choices of pseudometric are popular, but they are nearly always based on the norm of some Lebesgue space $L^p(Q)$, with respect to some distribution Q. We will generally specify such a norm as $d_{p,Q}$. Note that when Q is a discrete distribution (such as an empirical distribution corresponding to a sample P, functions which are equal Q-almost everywhere are strictly equal, so that $d_{p,Q}$ is a metric. We now give a few examples of such $d_{p,Q}$.

Example 5.5.

$$\begin{aligned} d_{1,D}(w_1, w_2) &= \int_{\mathcal{Z}} |w_1 - w_2| \, dD \\ &= \mathbb{E}_{(x,y)\sim D} \, |w_1(x) - w_2(x)| \ , \end{aligned}$$

the mean absolute deviation between w_1 and w_2.

$$d_{\infty,D}(w_1, w_2) = \operatorname{ess} \sup_{(x,y)\sim D} |w_1(x) - w_2(x)| \ ,$$

i.e. the smallest value almost surely bounding $|w_1(x) - w_2(x)|$ with respect to $x \sim D$. □

Example 5.6.

$$\begin{aligned} d_{1,P}(w_1, w_2) &= \int_{\mathcal{Z}} |w_1 - w_2| \, dP \\ &= \sum_{(x_i,y_i)\in P} \frac{1}{m} |w_1(x_i) - w_2(x_i)| \ , \end{aligned}$$

[83] A pseudometric space is a set equipped with a pseudometric. A pseudometric is similar to a metric: a function $d(v_1, v_2)$ is a pseudometric if d is positive, symmetric and obeys a triangle inequality. However, it is possible to have $d(v_1, v_2) = 0$ even when $v_1 \neq v_2$.

the average difference between w_1 and w_2 over the sample P. This gives us an idea of whether w_1 and w_2 behave similarly on the predictors in the sample P (but it is not yet clear they act similarly over the rest of $\mathcal{X}$). □

Example 5.7. More generally, for $1 \leq p < \infty$,

$$\begin{aligned} d_{p,P}(w_1, w_2) &= \left(\int_{\mathcal{Z}} |w_1 - w_2|^p \, dP \right)^{\frac{1}{p}} \\ &= \left(\sum_{(x_i, y_i) \in P} \frac{1}{m} |w_1(x_i) - w_2(x_i)|^p \right)^{\frac{1}{p}} \quad . \end{aligned}$$

For some of the work that follows, we will be particularly interested in $d_{2,P}$. □

Example 5.8. As final examples,

$$d_{\infty,P}(w_1, w_2) = \max_{(x_i, y_i) \in P} |w_1(x_i) - w_2(x_i)| \ ,$$

the maximum difference between w_1 and w_2 on the predictors in P, and

$$d_{0,P}(w_1, w_2) = \sum_{(x_i, y_i) \in P} |w_1(x_i) - w_2(x_i)|^0 \ ,$$

the number of predictors in P for which w_1 and w_2 differ,[84] are popular choices of pseudometric. □

It can easily be shown (using Jensen's inequality) that for any $p \geq 1$,

$$d_{1,Q} \leq d_{p,Q} \leq d_{\infty,Q} \ .$$

In addition, if P is an n-sample, we also have that

$$d_{\infty,P} \leq n d_{1,P} \ .$$

Another result that will be useful is that when w_1, w_2 map into $[0, 1]$, we have

$$[d_{p,Q}(w_1, w_2)]^p \leq d_{1,Q}(w_1, w_2) \tag{5.3}$$

[84] In this result we define $0^0 = 0$. We shall use this notation even though there is no corresponding Lebesgue space.

for $1 \leq p < \infty$. This follows by noting that for $v_1, v_2 \in [0,1]$, $|v_1 - v_2|^p \leq |v_1 - v_2|$ for such p.[85] If the functions map into $\{0,1\}$ (such as zero-one loss functions), equality results.

Up to now we have used absolute value to measure the distance between $w_1(x_i)$ and $w_2(x_i)$, since they are assumed to be real. However, this is just a special case of a more general formulation, which we will need on occasion later. Specifically, assume that v_1 and v_2 are functions mapping a set $\mathcal{X}$ into a pseudometric space $(\mathcal{E}, d)$, and that D is a distribution on $\mathcal{E}$. Then, we can define

$$d_{p,D}(v_1, v_2) = \left(\int_{x \in \mathcal{X}} [d(v_1(x), v_2(x))]^p \, dD \right)^{\frac{1}{p}}$$

for $1 \leq p < \infty$, and for $p = \infty$, the distance is the essential supremum of $d(v_1(x), v_2(x))$. In the specific case we have been considering, we have $\mathcal{E} = \mathrm{I\!R}$, with d defined by absolute difference.

Definition 5.1 (Packing and covering numbers). Consider a pseudometric space $(\mathcal{E}, d)$. We say that $\mathcal{J} \subseteq \mathcal{E}$ is an γ-cover of $\mathcal{K} \subseteq \mathcal{E}$ with respect to d if $\mathcal{J} \subseteq \mathcal{K}$ and for every $k \in \mathcal{K}$, there is a $j \in \mathcal{J}$ such that $d(j,k) < \gamma$ (i.e. j γ-approximates k w.r.t. d). If only the second condition holds, we call $\mathcal{J}$ an *external* γ-cover.[86] In both cases, we refer to γ as the resolution of the cover.

The (external) γ-covering number of $\mathcal{K}$ with respect to d, is the minimal cardinality of an (external) γ-cover of $\mathcal{K}$, if this is finite, and infinity, otherwise. We denote the γ-covering number by $\mathcal{N}(\gamma, \mathcal{K}, d)$, and the external γ-covering number by $\bar{\mathcal{N}}(\gamma, \mathcal{K}, d)$. Note that since a cover is an external cover, $\bar{\mathcal{N}}(\gamma, \mathcal{K}, d) \leq \mathcal{N}(\gamma, \mathcal{K}, d)$. A cover with minimal cardinality is said to be minimal, and similarly for external covers.

We say that $\mathcal{J} \subseteq \mathcal{E}$ is γ-separated if for any $j_1, j_2 \in \mathcal{J}$, $d(j_1, j_2) \geq \gamma$ (i.e. j_1 and j_2 are γ-distinguishable w.r.t. d). The γ-packing number of $\mathcal{K}$ with respect to d, $\mathcal{M}(\gamma, \mathcal{K}, d)$, is the maximal cardinality of an γ-separated subset of $\mathcal{K}$, if this is finite, and infinity, otherwise. We call an γ-separated subset $\mathcal{J}$ of $\mathcal{E}$ *maximal* if $\mathcal{J}$ is the only γ-separated subset of $\mathcal{E}$ containing $\mathcal{J}$.

[85] Note that by translation, this in fact holds when $[0,1]$ is replaced by any interval of width 1.

[86] The terminology here is subject to debate. Sometimes, as in Vapnik (1982), what is here named a cover is called a *proper cover*. In those cases, the term cover refers to external covers. As pointed out in Vidyasagar (2002, Chapter 2), in learning theory it is typically more convenient to work with the definitions we have chosen.

Specifically, this means that no point in $\mathcal{E}$ can be added to $\mathcal{J}$ to obtain an γ-separated subset of $\mathcal{E}$. It follows that all points in $\mathcal{E}$ are within distance γ of some point in $\mathcal{J}$. Clearly, any γ-separated subset with $\mathcal{M}(\gamma, \mathcal{K}, d)$ elements is maximal.

Note that the (external) covering and packing numbers of a set are decreasing functions of the resolution.

In our case, a collection $\mathcal{W}^\star$ of decision rules is thus a γ-*cover* of the decision class $\mathcal{W}$ with respect to the pseudometric d if, for each $w \in \mathcal{W}$ there is a $w^\star \in \mathcal{W}^\star$ such that $w^\star$ γ-approximates w. If $\mathcal{W}^\star$ is a (proper) γ-cover of minimum cardinality, we call it minimal, and refer to the cardinality as the (proper) γ-covering number of $\mathcal{W}$ with respect to d. If $d = d_{p,P}$, we shall shorten notation by writing the covering number as $\mathcal{N}_{p,P}(\gamma, \mathcal{W})$ and analogously for the proper covering number.

Some bounds and related results are derived or stated in terms of packing numbers or external covering numbers, rather than covering numbers. However, this is not a large problem, as all of these quantities are closely related for bounded sets, as can be seen from the fact that a maximal ϵ-packing of a set is effectively a cover of the set. The following results, from Alon et al. (1993, Lemma 2.3) and Vidyasagar (2002, Lemmas 2.1 and 2.2), show that one can convert between the various concepts with at worst a change of resolution of factor 2.

Theorem 5.1. *If $\mathcal{K}$ is a subset of $\mathcal{E}$, and $\gamma > 0$,*

$$\mathcal{M}(2\gamma, \mathcal{K}, d) \leq \bar{\mathcal{N}}(\gamma, \mathcal{K}, d) \leq \mathcal{N}(\gamma, \mathcal{K}, d) \leq \mathcal{M}(\gamma, \mathcal{K}, d) \ .$$

A further complication arises in that some definitions of covers employ the condition $d(j, k) \leq \gamma$ instead of $d(j, k) < \gamma$. In this case, adding an arbitrarily small constant to the resolution of the covering number usually yields a result which is valid for the concepts we employ.

If $d_1 \leq d_2$, for any distribution Q, we always have

$$\mathcal{N}(\gamma, \mathcal{K}, d_1) \leq \mathcal{N}(\gamma, \mathcal{K}, d_2) \ .$$

For any $w \in \mathcal{K}$,

$$\big\{w' : d_2(w, w') \leq \gamma\big\} \subseteq \big\{w' : d_1(w, w') \leq \gamma\big\}$$

due to the relationship between the metrics. Thus any γ-cover with respect to d_2 is also one with respect to d_1, so the relationship between the covering numbers follows. In general, for any distribution Q, we have that $d_{p,Q} \leq d_{q,Q}$ for $1 \leq p \leq q$, so that

$$\mathcal{N}_{p,Q}(\gamma, \mathcal{K}) \leq \mathcal{N}_{q,Q}(\gamma, \mathcal{K}) \ , \tag{5.4}$$

and from (5.3), we obtain that for $\gamma > 0$, we have that

$$\mathcal{N}_{p,Q}\left(\gamma^{\frac{1}{p}}, \mathcal{K}\right) \leq \mathcal{N}_{1,Q}(\gamma, \mathcal{K}) \tag{5.5}$$

for $1 \leq p < \infty$, when $\mathcal{K}$ consists of functions into $[0, 1]$. If the functions map into $\{0, 1\}$, equality holds.

Suppose $Q = [\![q_1, \cdots, q_n]\!]$ is an n-sample. Then we also have that

$$\mathcal{N}_{\infty,Q}(\gamma, \mathcal{K}) \leq \mathcal{N}_{1,Q}\left(\frac{\gamma}{n}, \mathcal{K}\right) \ , \tag{5.6}$$

Consider any γ, v with $\gamma \geq v > 0$.[87] If $\gamma - v > 1$, set $j = 0$. Otherwise, let $\phi = \phi(\gamma, v)$ be the smallest natural number such that $(2\phi + 1)(\gamma - v) > 1$. Define $w_{a_1,a_2,\cdots,a_n}$ on $\text{supp}(Q)$ by

$$w_{a_1,a_2,\cdots,a_n}(q_i) = \min\{(2a_i + 1)(\gamma - v), 1\} \ ,$$

where the integers a_i satisfy $0 \leq a_i \leq \phi$ for $i \in [1 : n]$. Extend this function in an arbitrary way to $\mathcal{X}$ to obtain a decision rule. Then the set of all such decision rules $w_{a_1,a_2,\cdots,a_n}$ is an external γ-cover of any decision class $\mathcal{W}$ w.r.t. $d_{\infty,Q}$ (and hence $d_{p,Q}$, for $p \geq 1$), with cardinality $(\phi + 1)^n$. It follows that $\bar{\mathcal{N}}_{p,Q}(\gamma, \mathcal{W}) \leq [\phi(\gamma, v) + 1]^n$ for an arbitrarily small v. Thus

$$\mathcal{N}_{p,Q}(\gamma, \mathcal{W}) \leq \mathcal{M}_{p,Q}(\gamma, \mathcal{W}) \leq \bar{\mathcal{N}}_{p,Q}\left(\frac{\gamma}{2}, \mathcal{W}\right) \leq \left[\phi\left(\frac{\gamma}{2}, v\right) + 1\right]^n \ .$$

[87] Here, v is introduced to cater for the fact that strict inequalities are used in the definition of covering numbers, and we will be interested in the case where v is arbitrarily close to zero.

It should be clear the behaviour of decision rules on points not in Q are irrelevant to the construction of covers w.r.t. $d_{p,Q}$. Put another way, the $d_{p,Q}$ metric only compares values of the decision rule at the points in Q. If we restrict each decision rule to $\mathrm{supp}(Q)$, we can represent w by a point $Q_w \in [0,1]^n$, where

$$Q_w = (w(q_1), \cdots, w(q_n)) \ ,$$

so that we can represent $\mathcal{W}$ by

$$Q_{\mathcal{W}} = \{Q_w : w \in \mathcal{W}\} \ .$$

It can easily be shown that any cover of $Q_{\mathcal{W}}$ w.r.t. the metric on ℓ_n^p implies an external cover on $\mathcal{W}$ w.r.t. $d_{p,Q}$ (by arbitrary extension, as above), and vice versa. However, it is important to note that it is common to have $|Q_{\mathcal{W}}|$ much less than $|\mathcal{W}|$.

We see that covering numbers w.r.t. an empirical distribution grow at worst exponentially in the cardinality of the sample. In practice, the decision class is restricted in some way, and we hope to obtain covering numbers which will grow more slowly than this. We shall later see that this slower growth over a decision class is what allows one to obtain nontrivial training sample interval estimators.

The points in the cover constructed above correspond to a regular grid of points placed over $[0,1]^n$, ensuring that some point in the grid will be close to Q_w for any conceivable decision rule w. However, for many decision classes $\mathcal{W}$, $Q_{\mathcal{W}}$ is unlikely to lie in all the portions of $[0,1]^n$. An obvious example is the case of zero-one loss functions: in that case $Q_{\mathcal{W}} \subseteq \{0,1\}^n$, so that when constructing a cover, we have no need of points in the middle of $[0,1]^n$, but only near the corners. This leads us to a simple bound on $\mathcal{N}_{p,Q}(\gamma, \mathcal{W})$ of 2^n, the number of vertices of the unit hypercube. Note that this still grows exponentially in n, though. A more important example: suppose the decision rules in $\mathcal{W}$ are Lipschitz continuous with Lipschitz constant 1. Then for any two points $q_i, q_j \in Q$, and any decision rule w, we have $|w(q_i) - w(q_j)| \leq d(q_i, q_j)$. A pairwise constraint on the coordinates of points in $Q_{\mathcal{W}}$ follows: for all $i, j \in [1 : n]$,

$$|(Q_w)^{(i)} - (Q_w)^{(j)}| \leq d(q_i, q_j) \ .$$

This result effectively constrains the location of the points in $Q_{\mathcal{W}}$ to a smaller region of $[0,1]^n$, allowing a smaller cover. This type of constraint typically restrains $Q_{\mathcal{W}}$ to a subclass of dimension less than n, unlike the previous example. Similarly, thresholded classifiers which threshold a Lipschitz function will tend to have $Q_{\mathcal{W}}$ restricted to certain portions of $\{0,1\}^n$.

Finally, we turn to covers of functions into $\{0,1\}$. Consider any two distinct points w_1, w_2 in $\{0,1\}^n$. Then we have that

$$d_{p,Q}(w_1, w_2) = d_{1,Q} \geq m^{\frac{-1}{p}}$$

for $1 \leq p < \infty$. Thus, we have that any two distinct points in $Q_{\mathcal{W}}$ are γ-distinguishable for all $\gamma \leq m^{\frac{-1}{p}}$. Thus, if $\gamma \leq m^{\frac{-1}{p}}$, the only point in $Q_{\mathcal{W}}$ which approximates a given $Q_w \in Q_{\mathcal{W}}$ is Q_w itself. It follows that the only γ-cover of $Q_{\mathcal{W}}$ is $Q_{\mathcal{W}}$, so that $\mathcal{N}_{p,Q}(\gamma, \mathcal{W}) \leq |Q_{\mathcal{W}}| \leq 2^n$ for all γ, with equality for $\gamma \leq m^{\frac{-1}{p}}$. A similar argument shows that $\mathcal{N}_{\infty,Q}(\gamma, \mathcal{W}) = |Q_{\mathcal{W}}|$ for $\gamma \leq 1$. Furthermore, if $\gamma > 1$, we have $\mathcal{N}_{\infty,Q}(\gamma, \mathcal{W}) = 1$ since $(0, \cdots, 0)$ is an appropriate $(1+v)$-cover for arbitrarily small v. By the relationships discussed above, there are appropriate scale thresholds for γ below which packing numbers or external covering numbers also equal $|Q_{\mathcal{W}}|$.

5.5.2 A naïve covering number bound

Given this background, we now outline an approach to constructing bounds employing covering numbers. Given a sample S of size m, we employ a bound on the deviation $\psi(r_S(w), r_D(w))$. We apply the uniform Occam's razor method to this result for each w in a minimal cover of $\mathcal{W}$ with respect to $d_{\infty,D}$. Thereafter, we extend the bound from all w in the cover to all w in $\mathcal{W}$.

As an example, we consider Hoeffding's tail inequality:

$$\mathbb{P}_{S \sim D^m} \{r_S(w) - r_D(w) > \epsilon\} < \exp(-2m\epsilon^2) \ .$$

Suppose $\mathcal{N}_{\infty,D}(\gamma, \mathcal{W})$ is finite, and that $\mathcal{W}^\star$ is a minimal γ-cover with respect to $d_{\infty,D}$. Applying the Occam's razor method over $\mathcal{W}^\star$ to the bound above yields

$$\mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}^\star}[r_S(w) - r_D(w)] > \epsilon\right\} < \mathcal{N}_{\infty,D}(\gamma, \mathcal{W})\exp(-2m\epsilon^2) \ .$$

One last step remains: extending this bound to all $w \in \mathcal{W}$. Now, for any $w \in \mathcal{W}$, we have

$$r_S(w) - r_D(w) = [r_S(w) - r_S(w^\star)] + [r_S(w^\star) - r_D(w^\star)] \ ,$$

where $w^\star$ is a member of $\mathcal{W}^\star$ which γ-approximates w. We have already bounded the second term, so we turn our attention to the first. Since

$$r_S(w) - r_S(w^\star) = \frac{1}{m}\sum_{i=1}^{m}[L(w(x_i), y_i) - L(w^\star(x_i), y_i)] \ ,$$

the fact that $w^\star$ γ-approximates w w.r.t. $d_{\infty,D}$ means that each term in this sum is less than $\frac{\gamma}{m}$.[88] Thus $r_S(w) - r_S(w^\star) < \gamma$. It follows that

$$\begin{aligned}&\mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}}[r_S(w) - r_D(w)] > \epsilon + \gamma\right\}\\ &\le \mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}^\star}[r_S(w) - r_D(w)] > \epsilon\right\} \ ,\end{aligned}$$

so that

$$\mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}}[r_S(w) - r_D(w)] > \epsilon\right\} < \mathcal{N}_{\infty,D}(\gamma, \mathcal{W})\exp(-2m(\epsilon-\gamma)^2) \ .$$

Setting the right hand side to δ and solving for ϵ, we obtain

$$\mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}}[r_S(w) - r_D(w)] > \gamma + \sqrt{\frac{\ln\mathcal{N}_{\infty,D}(\gamma, \mathcal{W}) - \ln\delta}{2m}}\right\} < \delta \ .$$

Note that the same proof can be applied with a minimal external γ-cover, so that the result above can be strengthened by replacing $\mathcal{N}_{\infty,D}(\gamma, \mathcal{W})$ by $\bar{\mathcal{N}}_{\infty,D}(\gamma, \mathcal{W})$.

[88] It complicates the argument to cater for the possibility that $|w_1(x) - w_2(x)| > d_{\infty,D}(w_1, w_2)$ for some $x \in S$. Since the probability of this occuring is zero by definition of the pseudometric, we ignore the possibility. A rigorous argument is longer, but not substantially different.

Since D is generally unknown, and only assumed to lie in some collection of distributions $\mathcal{S}$, to apply this bound we need to replace $\mathcal{N}_{\infty,D}(\gamma, \mathcal{W})$ by

$$\mathcal{N}_{\infty,\mathcal{S}}(\gamma, \mathcal{W}) = \sup_{Q \in \mathcal{S}} \mathcal{N}_{\infty,Q}(\gamma, \mathcal{W}) \ .$$

When the metric is based on $p = \infty$, as here, this supremum is equal to $\mathcal{N}_{\infty,Q}(\gamma, \mathcal{W})$ for any distribution Q with $\operatorname{supp}(Q) = \mathcal{Z}$.

When $\mathcal{S} = \mathcal{Q}_{\mathcal{Z}}$,

$$\begin{aligned} \sup_{Q \in \mathcal{S}} \mathcal{N}_{p,Q}(\gamma, \mathcal{W}) &= \mathcal{N}_{p,\mathcal{S}}(\gamma, \mathcal{W}) \\ &= \mathcal{N}_{\infty,\mathcal{S}}(\gamma, \mathcal{W}) \ , \end{aligned}$$

so that even if the bound above is constructed using covering numbers with respect to some metric $d_{p,D}$ where $p \neq \infty$, the need for this supremum makes the choice of p fairly arbitrary.

Furthermore, γ-approximation w.r.t. $d_{p,D}$ is generally not useful for bounding $r_S(w) - r_S(w^\star)$, except in the case $p = \infty$. These considerations motivate the choice of $p = \infty$ for the cover in the result above.

The bound presented in this section suffers from a major drawback: the need to find a bound on $\mathcal{N}_{\infty,D}(\gamma, \mathcal{W})$. Consider the simple case of a zero-one loss function, and any $\gamma \leq \frac{1}{2}$. Consider two decision rules w_1 and w_2 which differ at some point $x \in \operatorname{supp}(D)$, so that $d_{p,D}(w_1, w_2) > 0$. Then $|w_1(x) - w_2(x)| = 1$, so that for any decision rule $w^\star$, we can not have $|w_1(x) - w^\star(x)| < \gamma$ and $|w_2(x) - w^\star(x)| < \gamma$. Thus any two decision rules in $\mathcal{W}$ which differ on the support of D (i.e. essentially different decision rules), can not be γ-approximated by the same decision rule w.r.t. $d_{\infty,D}$. Thus, $\mathcal{N}_{\infty,D}(\gamma, \mathcal{W}) = |\mathcal{W}|$ (assuming that each decision rule in $\mathcal{W}$ is essentially different).

Since we are considering covering number bounds to address infinite decision classes, this bound will not be very helpful in practice. Its main value in the text is to introduce the main theme behind constructing bounds employing covering numbers. However, we shall need two refinements to the approach to obtain bounds which do not generally have infinite covering numbers.

The following two sections introduce these refinements for various measures of deviation.

5.5.3 Symmetrization lemmas

In this section, we investigate the possibility of obtaining bounds using covering numbers over empirical distributions. However, we would like our results to hold for any sample, which motivated our choice of using covering numbers based on D in the previous section. In what follows, we present an alternative solution: deriving bounds for a finite subset of $\mathrm{supp}(D)$ and then converting these bounds to apply to the whole support of D.

Key results which allow us to do this are the so-called *symmetrization lemmas*[89]. Such lemmas quantify how a probability statement w.r.t. a finite sample can be converted into a similar probability statement w.r.t. the entire sample space.

When the first symmetrization lemmas were originally derived by Vapnik and Chervonenkis in the late 1960's (for an account of these developments, see Vapnik, 1998), the results derived were tailored for their situation. As a result they were fairly tight, but they were restricted to zero-one loss functions. Later research considered obtaining bounds on general loss functions which were sensitive to the resolution of the covering numbers used. However, the basic idea behind the proofs of the original symmetrization lemmas used by Vapnik and Chervonenkis, and the later symmetrization lemmas were almost identical. We begin by presenting the general symmetrization lemma from Pollard (1984, Chapter II). An alternative formulation is in Dudley (1999, Lemma 11.2.4).

Theorem 5.2 (General symmetrization lemma). *Let $E(v)$ and $E'(v)$ be independent stochastic processes sharing an index set $\mathcal{V}$. Suppose there*

[89] The name of these lemmas come from the fact that applying such a lemma typically reduces the problem of bounding probabilities over an arbitrary distribution, to bounding a related probability over an empirical distribution corresponding to a finite sample. Such results can be obtained by considering the *symmetric group* of permutations of the finite sample.

exist constants $\alpha, \beta > 0$ *such that*

$$\mathbb{P}\left\{E'(v) \le \alpha\right\} \ge \beta$$

for every $v \in \mathcal{V}$. *Then*

$$\mathbb{P}\left\{\sup_{v\in\mathcal{V}} E(v) > \epsilon\right\} \le \beta^{-1}\,\mathbb{P}\left\{\sup_{v\in\mathcal{V}}(E(v) - E'(v)) > \epsilon - \alpha\right\} \ .$$

Proof. The proof sketch here follows that in Pollard (1984, Section II.3), except that the result stated here is one-sided. Application to the negated processes allows obtention of a two-sided result.[90]

Let v_E be a r.v. in $\mathcal{V}$ such that $E(v_E) > \epsilon$ when $\sup_{v\in\mathcal{V}} E(v) > \epsilon$.

Now, since E and E' are independent, v_E is independent of E'. Thus, by the assumptions of the Lemma,

$$\mathbb{P}\left\{E'(v_E) \le \alpha | E\right\} \ge \beta \ .$$

Now

$$\begin{aligned}
\beta\,\mathbb{P}\left\{\sup_{v\in\mathcal{V}} E(v) > \epsilon\right\} &\le \mathbb{P}\left\{E'(v_E) \le \alpha | E\right\}\mathbb{P}\left\{\sup_{v\in\mathcal{V}} E(v) > \epsilon\right\} \\
&= \mathbb{P}\left\{\big(E'(v_E) \le \alpha\big) \wedge \left(\sup_{v\in\mathcal{V}} E(v) > \epsilon\right)\right\} \\
&= \mathbb{P}\left\{\big(E'(v_E) \le \alpha\big) \wedge \left(E(v_E) > \epsilon\right)\right\} \\
&\le \mathbb{P}\left\{E(v_E) - E'(v_E) > \epsilon - \alpha\right\} \\
&\le \mathbb{P}\left\{\sup_{v\in\mathcal{V}}[E(v) - E'(v)] > \epsilon - \alpha\right\} \ .
\end{aligned}$$

Dividing throughout by β yields the result. □

The most common application of the general symmetrization lemma in our framework is when the stochastic processes E and E' represent the deviation of empirical risk from true risk for an m-sample S and a u-sample P respectively. In this case, $\mathcal{V} = \mathcal{W}$, $E(w) = r_D(w) - r_S(w)$, and $E'(w) = r_D(w) - r_P(w)$. To apply the symmetrization lemma, we then need a result of the form: for some $\alpha, \beta > 0$,

$$\mathbb{P}_{P\sim D^u}\left\{r_D(w) - r_P(w) \le \alpha\right\} \ge \beta$$

[90] If the conditions of the lemma are met for the negated process, of course.

for every $w \in \mathcal{W}$. But such bounds underlie all the exact upper test sample interval estimators we have derived. As an example, let us consider Hoeffding's tail inequality. For a fixed choice of β, the condition in the symmetrization lemma holds with $\alpha_H(u,\beta) = \sqrt{\frac{-\ln(1-\beta)}{2u}}$. Another bound which easily yields analytical results is Chebyshev's inequality. In this case, the condition seems to hold with $\alpha_C(u,\beta) = \sqrt{\frac{\mathbb{V}_{P\sim D^u}\, r_P(w)}{(1-\beta)}}$, but this is not the case, since this α_C is a function of w, not a constant, for a given β. Since $\mathbb{V}_{P\sim D^u}\, r_P(w) \leq \frac{r_D(w)[1-r_D(W)]}{u} \leq \frac{1}{4u}$, the condition also holds with $\alpha_{C'}(u,\beta) = \frac{1}{2\sqrt{u(1-\beta)}}$. Another proposal is the use of the Chebyshev-Cantelli inequality (4.1) to obtain a suitable α. This may be because the symmetrization lemma is usually considered for two-sided processes directly. However, when one is specifically considering one-sided results, the Chebyshev-Cantelli inequality yields an improved choice of α. Specifically, we obtain $\alpha_{CC'}(u,\beta) = \sqrt{\frac{\beta}{4u(1-\beta)}}$ by upper bounding an α_{CC} as with the Chebyshev inequality.

We thus obtain, for any $0 < \beta \leq 1$,

$$\mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}}\left[r_D(w) - r_S(w)\right] > \epsilon\right\} \\ \leq \beta^{-1}\,\mathbb{P}_{S\oplus P\sim D^{m+u}}\left\{\sup_{w\in\mathcal{W}}\left[r_P(w) - r_S(w)\right] > \epsilon - \alpha(u,\beta)\right\} \quad , \qquad (5.7)$$

where α is α_H, $\alpha_{C'}$, $\alpha_{CC'}$, or any other function $\alpha(u,\beta)$ satisfying the conditions of the lemma (note that we can not use $\alpha = \alpha_C$, since α_C is dependent on w). Here, and further, $S \oplus P$ denotes the sample obtained by concatenating S and P (as well as the associated empirical distribution).

It is common in the machine learning literature to set $\beta = \frac{1}{2}$ and then require $\alpha(u,\frac{1}{2}) \leq \frac{\epsilon}{2}$ when deriving results. The results are then stated along with a condition on the sample size in terms of ϵ. Since $\alpha_{C'}(u,\frac{1}{2}) = \sqrt{\frac{1}{2u}}$, we obtain the restriction $u \geq 2\epsilon^{-2}$, yielding:

Theorem 5.3 (Traditional symmetrization lemma for regular deviation).

$$\mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}}\left[r_D(w)-r_S(w)\right]>\epsilon\right\}$$
$$\leq 2\,\mathbb{P}_{S\oplus P\sim D^{m+u}}\left\{\sup_{w\in\mathcal{W}}\left[r_P(w)-r_S(w)\right]>\frac{\epsilon}{2}\right\}\quad ,$$

if $u\geq 2\epsilon^{-2}$.

Note that by using $\alpha_{CC'}$ instead, the sample size requirement is reduced by a factor of two, to $u\geq\epsilon^{-2}$.

To date, we are not aware of a useful symmetrization lemma for relative deviation which is applicable to general loss functions. This has led to the consideration of other deviations which attempt to take the variance of the decision rules into consideration.

Haussler (1992) employs a similar argument to the proof of the general symmetrization lemma to obtain a symmetrization lemma for the two-sided P-H ν-deviation. Below we present a slight generalization of this result, allowing $m\neq u$ and introducing α and β by analogy to the general symmetrization lemma. The proof is a similar to that of the original result in Haussler (1992, Lemma 12).

Theorem 5.4 (Haussler symmetrization lemma). *Let ψ_ν denote the two-sided P-H ν-deviation. Suppose there exist constants $\alpha,\beta>0$ such that*

$$\mathbb{P}_{P\sim D^u}\left\{\psi_\nu(r_D(w),r_P(w))\leq\alpha\right\}\geq\beta \tag{5.8}$$

for every $w\in\mathcal{W}$. Then

$$\mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}}\psi_\nu(r_D(w),r_S(w))>\epsilon\right\}$$
$$\leq\beta^{-1}\,\mathbb{P}_{S\oplus P\sim D^{m+u}}\left\{\sup_{w\in\mathcal{W}}\psi_\nu(r_P(w),r_S(w))>\epsilon-\alpha\right\}\quad .$$

Unfortunately, in this case, a one-sided result does not seem practical. The proof of the result relies on the fact that ψ_ν is a metric, and thus satisfies the triangle inequality. For the general symmetrization lemma, we could work around this issue, but we did not find a similar one-sided result for the P-H deviation.

In practice, we need a suitable way to determine the constants α and β. A useful tool for this is the bound on P-H deviation obtained by using Bernstein's inequality. Specifically, from the bound of (4.9) we have that the condition above holds for a specific α when

$$\beta_{PH}(\alpha) = 1 - \exp\left(\frac{-18u\alpha^2\nu}{(3+\alpha)^2}\right) \ .$$

It is usually more convenient to specify β and solve for α. This yields, for $18u\nu \leq -\ln(1-\beta)$,

$$\alpha_{PH}(\beta) = \epsilon(1-\beta) \ ,$$

where $\epsilon(\delta)$ is defined in (4.10).

Bartlett and Lugosi (1999) provide symmetrization lemmas for the upper and lower B-L deviation. The results provided here generalize the results provided there by permitting $m \neq u$, and allowing other choices of α and β.[91]

Theorem 5.5 (Bartlett-Lugosi symmetrization lemmas). *Let ψ_ν^U and ψ_ν^L denote the upper and lower B-L ν-deviations respectively. Suppose there exist constants $\alpha, \beta > 0$ such that*

$$\mathbb{P}_{P\sim D^u}\left\{r_D(w) - r_P(w) \leq \alpha\right\} \geq \beta \tag{5.9}$$

for every $w \in \mathcal{W}$. Then, for $\alpha < \nu$,

$$\mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}} \psi_\nu^U(r_D(w), r_S(w)) > \epsilon\right\} \leq \beta^{-1}\,\mathbb{P}_{S\oplus P\sim D^{m+u}}\left\{\sup_{w\in\mathcal{W}} \frac{r_P(w) - r_S(w) - (\nu-\alpha)}{\sqrt{\frac{1}{2}(r_P(w) + r_S(w))}} > \epsilon\right\} \ .$$

Suppose there exist constants $\alpha, \beta > 0$ such that

$$\mathbb{P}_{P\sim D^u}\left\{r_P(w) - r_D(w) \leq \alpha\right\} \geq \beta \tag{5.10}$$

[91] We also note a minor correction to their proof. For the upper deviation, we note that the function $\frac{x-a}{\sqrt{x+a}}$ is monotone increasing in $x > -a$ when $a \geq 0$, rather than just for $x > 0$. This additional information is necessary for the proofs provided there to hold. A similar change is necessary for the proof for the lower deviation.

for every $w \in \mathcal{W}$. Then, for $\alpha < \nu$,

$$\mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}} \psi_\nu^L(r_D(w), r_S(w)) > \epsilon\right\}$$
$$\leq \beta^{-1}\,\mathbb{P}_{S\oplus P\sim D^{m+u}}\left\{\sup_{w\in\mathcal{W}} \frac{r_S(w) - r_P(w) - (\nu-\alpha)}{\sqrt{\frac{1}{2}(r_P(w)+r_S(w))}} > \epsilon\right\} \;.$$

Note that the choice of α in these lemmas was originally determined using $\alpha_{CC'}$. The proof of this modified version is similar to the original proof.

Vapnik (1998) presents two symmetrization lemmas for zero-one loss functions, derived in a similar way to the general symmetrization lemma, for the case $m = u$. The first reduces the restriction on ϵ in the lemma on regular deviation above, and the second provides a symmetrization lemma for upper relative deviation.

Theorem 5.6 (Vapnik symmetrization lemma for regular deviation). *Suppose $m = u$. Then*

$$\mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}} [e_D(w) - e_S(w)] > \epsilon\right\}$$
$$\leq 2\,\mathbb{P}_{S\oplus P\sim D^{2m}}\left\{\sup_{w\in\mathcal{W}} [e_P(w) - e_S(w)] > \epsilon - \frac{1}{m}\right\} \;.$$

Thus, if $m \geq 2\epsilon^{-1}$,

$$\mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}} [e_D(w) - e_S(w)] > \epsilon\right\}$$
$$\leq 2\,\mathbb{P}_{S\oplus P\sim D^{2m}}\left\{\sup_{w\in\mathcal{W}} [e_P(w) - e_S(w)] > \frac{\epsilon}{2}\right\} \;.$$

The same results hold for lower and two-sided regular deviation.

Theorem 5.7 (Vapnik symmetrization lemma for upper relative deviation). *Let $p > 1$. For $u > \epsilon^{\frac{-p}{p-1}}$,*

$$\mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}} \frac{e_D(w) - e_S(w)}{\sqrt[p]{e_D(w)}} > \epsilon\right\}$$
$$< 4\,\mathbb{P}_{S\oplus P\sim D^{m+u}}\left\{\sup_{w\in\mathcal{W}} \frac{e_P(w) - e_S(w)}{\sqrt[p]{\frac{1}{2}[e_S(w) + e_P(w) + \frac{1}{u}]}} > \epsilon\right\} \;.$$

One can simplify, but loosen, this result by discarding the $\frac{1}{u}$ term in the denominator of the right hand side.

The lemma for regular deviation is based on Vapnik (1998, Section 4.5, Basic Lemma)[92]. The lemma for relative deviation is Vapnik (1998, Lemma 4.1). Note that the result in this theorem has been generalized to the case $m \neq u$. We omit the proof since it is highly similar to that provided in the original source. In general, we shall apply the lemma for relative deviation with $p = 2$.

5.5.4 Dual sample bounds

The symmetrization lemmas above provide us with a way of obtaining training sample interval estimators for risk by uniformly bounding the deviation between empirical risks on two samples. As an example, suppose

$$\mathbb{P}_{S\oplus P\sim D^{m+u}}\left\{\sup_{w\in\mathcal{W}}\left[r_P(w) - r_S(w)\right] > \epsilon\right\} < \delta(\epsilon) \ ,$$

for any $\epsilon > 0$. It follows from (5.7) that

$$\mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}}\left[r_D(w) - r_S(w)\right] > \epsilon\right\} \leq \beta^{-1}\delta(\epsilon - \alpha(u,\beta))$$

for any $\epsilon > \alpha(u,\beta)$.

In this context, we turn to the problem of finding a bound on

$$\sup_{w\in\mathcal{W}}\left[r_P(w) - r_S(w)\right] \ .$$

Consider the following naïve application of test set bounds. It is easy to see that

$$\begin{aligned}&\mathbb{P}_{S\oplus P\sim D^{m+u}}\left\{\left[r_P(w) - r_S(w)\right] > \epsilon\right\}\\ &\quad\leq \mathbb{P}_{P\sim D^u}\left\{\left[r_P(w) - r_D(w)\right] > \frac{\epsilon}{2}\right\} + \mathbb{P}_{S\sim D^m}\left\{\left[r_D(w) - r_S(w)\right] > \frac{\epsilon}{2}\right\} \ .\end{aligned}$$

Applying Hoeffding's tail inequality to each term, one obtains that this is less than

$$\exp\left(-\frac{\epsilon^2 u}{2}\right) + \exp\left(-\frac{\epsilon^2 m}{2}\right) \ . \tag{5.11}$$

[92] An easy proof that the result holds in the one-sided case is Kroon (2003, Theorem 66).

This approach is quite wasteful, however. A more refined approach allows us to obtain a much better bound, when $m = u$. In that case,

$$\mathbb{E}_{S \oplus P \sim D^{m+u}}[r_P(w) - r_S(w)] = 0 \ ,$$

and $r_P(w) - r_S(w)$ can be viewed as a sum of independent r.v.'s. Applying Hoeffding's inequality in this scenario yields the much better bound of

$$\exp(-m\epsilon^2) \ . \tag{5.12}$$

We are now faced with the problem of converting this bound on the deviation of a single decision rule w to a bound on the supremum of the deviation for all decision rules in $\mathcal{W}$. The approach which suggests itself is to apply the Occam's razor method to a γ-cover of $\mathcal{W}$. The choice of metric is troublesome however, because the bound above is with respect to samples drawn i.i.d. from D, seemingly necessitating a metric w.r.t. D.

Fortunately, we have an ingenious solution: work conditionally on the $m+u$ points in $S \oplus P$. Note that for any event $\mathscr{E}$ which can be written as a function of an $m+u$-sample, we can write

$$\mathbb{P}_{S \oplus P \sim D^{m+u}} \left\{ \mathscr{E}(S \oplus P) \right\}$$

as

$$\mathbb{E}_{Q \sim D^{m+u}} \, \mathbb{P}_{\tau \sim \mathrm{Unif}\, S_{m+u}} \left\{ \mathscr{E}(\tau(Q)) | Q \right\} \ , \tag{5.13}$$

where S_{m+u} denotes the symmetric group, i.e. the possible permutations of the sample Q. This can be seen by noting that each value of τ merely changes the order of integration in calculating the probabilities, because all the points in $S \oplus P$ are i.i.d. This technique for obtaining such probabilities is known as *symmetrization by permutation* (Herbrich and Williamson, 2002).

If $m = u$, we call the subgroup $S^\star_{2m}$ of S_{2m} generated by the transpositions $(i \leftrightarrow m+i)$ with $i \in [1:m]$ the *swapping subgroup* of S_{2m}. Note that the result above also holds when S_{2m} is replaced by $S^\star_{2m}$.

Writing $\mathscr{E}_w(S \oplus P, \epsilon)$ for the event $r_P(w) - r_S(w) > \epsilon$, our initial focus is then on bounding $\mathbb{P}_{\tau \sim \mathrm{Unif}\, S_{m+u}} \left\{ \mathscr{E}_w(\tau(Q), \epsilon) | Q \right\}$, for any decision rule w.

This bound can then be used later to obtain bounds on the supremum. Note that we have no guarantee that the bounds based on Hoeffding's inequality in (5.11) and (5.12) hold conditional on the combined sample Q: Hoeffding's inequality requires independence of each term in the sum, and given Q, the value of any one r.v. is completely fixed by knowledge of the other $m+u-1$ r.v.'s. Nevertheless, the bounds above serve as benchmarks: if we were to better them, taking an expectation over all possible Q would yield an improvement over the results above.

It turns out that we can very nearly achieve the benchmark for zero-one loss functions when $m = u$:

Theorem 5.8 (Vapnik double sample bound). *Let $\mathscr{E}_w(S \oplus P, \epsilon)$ be either of the events*

$$e_P(w) - e_S(w) > \epsilon$$

or

$$e_S(w) - e_P(w) > \epsilon \ .$$

Then

$$\mathbb{P}_{\tau \sim \mathrm{Unif}\, S_{2m}} \left\{ \mathscr{E}_w(\tau(Q), \epsilon) | Q \right\} < \exp\left(-\frac{\epsilon^2 m^2 - 1}{m+1} \right)$$

and for any decision rule w. A two-sided result follows by Bonferroni's inequality.

This result is based on an implicit result in the proof of Vapnik (1998, Theorem 4.1).[93] The proof of this result directly investigates the proportion of combinations satisfying $\mathscr{E}_w(\tau(Q))$. The interested reader is referred to Vapnik (1998, Sections 4.5.4 and 4.13) for the details.

Vapnik (1998) also presents a bound on regular deviation for general loss functions which effectively employs the double sample bound for zero-one loss functions $2m$ times, and applies Bonferroni's inequality. We present here a one-sided version of that result implicit in the proof of Vapnik (1998, Theorem 15.1).[94]

[93] There is a slight mistake in the derivation there. Equation 4.67 in the reference should replace $\epsilon^2 l^2$ by $(\epsilon^2 l^2 - 1)$ and the correction carried further.

[94] The proof there is subject to a few corrections. As mentioned earlier, Equation 4.67 needs a modification. Further, the expression $3 \exp\left(-\frac{\epsilon^2 l}{9}\right)$ on p.623 should instead have coefficient 2.

Theorem 5.9 (Double sample bound for regular deviation of risk). *Let $\mathscr{E}_w(S \oplus P, \epsilon)$ be either of the events*

$$r_P(w) - r_S(w) > \epsilon$$

or

$$r_S(w) - r_P(w) > \epsilon \ .$$

Then

$$\mathbb{P}_{\tau \sim \mathrm{Unif}\, S^{\star}_{2m}} \left\{ \mathscr{E}_w(\tau(Q), \epsilon) | Q \right\} < 2m \exp\left(-\frac{\epsilon^2 m^2 - 1}{m+1} \right)$$

for any decision rule $w \in \mathcal{W}$.

Combining these results with Bonferroni's inequality yields a two-sided result.

We will typically use this in conjunction with the general symmetrization lemma for regular deviation.

We now present a dual sample bound for regular deviation, based on ideas in Devroye (1982).[95] Note that given Q, sampling $\tau \sim \mathrm{Unif}\, S_{2m}$ corresponds to sampling from the finite sample Q without replacement. Hoeffding (1963, Section 6) shows that the bounds in that paper can be applied when sampling without replacement. The good idea in Devroye (1982) is to reformulate the expression

$$r_P(w) - r_S(w) > \epsilon$$

as

$$r_{S \oplus P}(w) - r_S(w) > \frac{\epsilon u}{m+u} \ .$$

This follows from noting that

$$r_P(w) = \frac{1}{u}((m+u) r_{S \oplus P}(w) - m r_S(w)) \ . \tag{5.14}$$

An equivalent result for $r_P(w) - r_S(w)$ follows identically. In the reformulated version, we can apply Hoeffding's inequality, yielding a dual sample bound.

[95] The cited article derives the result for errors with $m + u = m^2$.

Theorem 5.10 (Dual sample bound for regular deviation of risk). *Let $\mathscr{E}_w(S \oplus P, \epsilon)$ be either of the events*

$$r_P(w) - r_S(w) > \epsilon$$

or

$$r_S(w) - r_P(w) > \epsilon \ \ .$$

Then

$$\mathbb{P}_{\tau \sim \mathrm{Unif}\, S_{m+u}} \{\mathscr{E}_w(\tau(Q), \epsilon) | Q\} < \exp\left(-2m\left(\frac{\epsilon u}{m+u}\right)^2\right)$$

for any decision rule $w \in \mathcal{W}$.

Combining these results with Bonferroni's inequality yields a two-sided result.

Vapnik (1998) uses a similar argument to his double sample bound for regular deviation of errors to derive a double sample bound for a generalized relative deviation of errors. However, due to the flaw in his Equation 4.67, the derivation provided is not correct (Vapnik, 2007). We present here a modified result which works around the problem, but represents a slight weakening (by a factor of less than two) of the bound implicit in Vapnik (1998, Lemma 4.2).

Theorem 5.11 (Double sample bound for upper relative deviation of error). *Let $\mathscr{E}_w(S \oplus P, \epsilon)$ be the event*

$$\frac{e_P(w) - e_S(w)}{\sqrt[p]{e_{s \oplus P}(w) + \frac{1}{2m}}} > \epsilon \ \ .$$

Then

$$\mathbb{P}_{\tau \sim \mathrm{Unif}\, S^\star_{2m}} \{\mathscr{E}_w(\tau(Q), \epsilon) | Q\} < \exp\left(\frac{m+1}{2m+1} - \frac{\epsilon^2 m^{2-\frac{2}{p}}}{2^{1+\frac{2}{p}}}\right)$$

for any decision rule $w \in \mathcal{W}$.

Vapnik (1998, Theorem 4.2*) and all further results based on it need to be corrected for this extra factor of $\exp\left(\frac{m+1}{2m+1}\right) \approx \sqrt{e}$, notably Theorem 4.2, Theorem 5.2, Theorem 5.3* and Theorem 5.4 of Vapnik (1998), many of which are considered standard in the machine learning literature.[96]

Anthony and Shawe-Taylor (1993) derives a result closely related to this result in the case $p = 2$. Although their original result is stated for errors, nothing in their proof, provided below, employs the restriction to zero-one loss functions, so that the result also holds for general loss functions. The result is based on conditioning w.r.t. the swapping subgroup $S^\star_{2m}$ rather than the symmetric group S_{2m}, so it also only applies when $m = u$.

Theorem 5.12 (Double-sample bound for upper relative deviation of risk). *Let $\mathscr{E}_w(S \oplus P, \epsilon)$ be the event*

$$\frac{r_P(w) - r_S(w)}{\sqrt{r_{S\oplus P}(w)}} > \epsilon \ .$$

Then

$$\mathbb{P}_{\tau\sim\mathrm{Unif}\, S^\star_{2m}}\left\{\mathscr{E}_w(\tau(Q), \epsilon)|Q\right\} < \exp\left(\frac{-\epsilon^2 m}{4}\right)$$

for any decision rule w.

We present a proof of this result, as the concepts involved are very similar to those used later for the random subsample lemma and bound.

Proof.

$$\begin{aligned}&\mathbb{P}_{\tau\sim\mathrm{Unif}\, S^\star_{2m}}\left\{\mathscr{E}_w(\tau(Q), \epsilon)|Q\right\} \\ &\qquad = \mathbb{P}_{\tau\sim\mathrm{Unif}\, S^\star_{2m}}\left\{r_{P_{\tau(Q)}}(w) - r_{S_{\tau(Q)}}(w) > \epsilon\sqrt{r_{\tau(Q)}(w)}|Q\right\} \ ,\end{aligned}$$

where $S_{\tau(Q)}$ and $P_{\tau(Q)}$ denote the first and second half of $\tau(Q)$ respectively. Writing s_i for the i-th component of $S_{\tau(Q)}$ and analogously for p_i and $P_{\tau(Q)}$, the above equals

$$\mathbb{P}_{\tau\sim\mathrm{Unif}\, S^\star_{2m}}\left\{\frac{1}{m}\sum_{i=1}^{m}[w(p_i) - w(s_i)] > \epsilon\sqrt{\frac{1}{2m}\sum_{i=1}^{2m} w(q_i)|Q}\right\} \ ,$$

where q_i denotes the i-th component of Q. Now, for any $i \in [1 : m]$, either $s_i = q_i$ and $p_i = q_{m+i}$, or $s_i = q_{m+i}$ and $p_i = q_i$. Thus we can rewrite the

[96] However, in the case of $p = 2$, the following result makes the change unnecessary.

above probability as

$$\mathbb{P}_{\zeta\sim\text{Unif}(\{-1,1\}^m)}\left\{\frac{1}{m}\sum_{i=1}^{m}\zeta_i[w(q_{m+i})-w(q_i)]>\epsilon\sqrt{\frac{1}{2m}\sum_{i=1}^{2m}w(q_i)}|Q\right\} \quad ,$$

where the vector of independent Rademacher variables $\zeta = (\zeta_1,\cdots,\zeta_m)$ has replaced the role of τ. Since Q is an i.i.d. sample, we have that $V_1(w),\cdots,V_m(w)$ defined by

$$V_i(w)=\zeta_i[w(q_{m+i})-w(q_i)] \quad ,$$

are independent r.v.'s, each with mean zero.[97]

Applying Hoeffding's tail inequality, we obtain

$$\begin{aligned}
&\mathbb{P}_{\zeta\sim\text{Unif}(\{-1,1\}^{2m})}\left\{\frac{1}{m}\sum_{i=1}^{m}\zeta_i[w(q_{m+i})-w(q_i)]>\epsilon\sqrt{\frac{1}{2m}\sum_{i=1}^{2m}w(q_i)}|Q\right\}\\
=\;&\mathbb{P}_{\zeta\sim\text{Unif}(\{-1,1\}^{2m})}\left\{\frac{1}{m}\sum_{i=1}^{m}V_i(w)>\epsilon\sqrt{\frac{1}{2m}\sum_{i=1}^{2m}w(q_i)}|Q\right\}\\
\le\;&\exp\left(\frac{-2m^2\left(\epsilon\sqrt{\frac{1}{2m}\sum_{i=1}^{2m}w(q_i)}\right)^2}{\sum_{i=1}^{m}[2(w(q_{m+i})-w(q_i))]^2}\right)\\
=\;&\exp\left(\frac{-m\epsilon^2\sum_{i=1}^{2m}w(q_i)}{4\sum_{i=1}^{m}[w(q_{m+i})-w(q_i)]^2}\right) \quad .
\end{aligned}$$

Given Q and w, we have

$$\begin{aligned}
\sum_{i=1}^{m}[w(q_{m+i})-w(q_i)]^2 &= \sum_{i=1}^{m}\left([w(q_{m+i})]^2+[w(q_i)]^2-2w(q_{m+i})w(q_i)\right)\\
&\le \sum_{i=1}^{2m}[w(q_i)]^2\\
&\le \sum_{i=1}^{2m}w(q_i) \quad ,
\end{aligned}$$

since $0\le w(q_i)\le 1$, so that

$$\exp\left(\frac{-m\epsilon^2\sum_{i=1}^{2m}w(q_i)}{4\sum_{i=1}^{m}[w(q_{m+i})-w(q_i)]^2}\right)\le\exp\left(\frac{-m\epsilon^2}{4}\right) \quad ,$$

[97] It is interesting to note, however, that the V_i are not generally identically distributed, even in the case of zero-one loss functions.

concluding the proof. □

This relative deviation double sample bound can be used in conjunction with the Vapnik symmetrization lemma for relative deviations, with $p = 2$ in order to obtain error bounds. However, we do not know of an appropriate symmetrization lemma to employ to obtain bounds on relative deviation for general loss functions. On the other hand, we shall see that this result can be combined with the symmetrization lemmas for B-L deviation to obtain bounds.

We next present a double sample bound for P-H deviation of risk. This result is a one-sided version of Haussler (1992, Lemma 11).

Theorem 5.13 (Double sample bound for P-H deviation of risk). *Let the event*

$$\psi_\nu(r_P(w), r_S(w)) > \epsilon \ ,$$

where ψ_ν denotes the upper or lower P-H ν-deviation, be denoted by $\mathscr{E}_w(S \oplus P, \epsilon)$.

Then

$$\mathbb{P}_{\tau \sim \mathrm{Unif}\, S^\star_{2m}} \{\mathscr{E}_w(\tau(Q), \epsilon) | Q\} < \exp(-2\epsilon^2 \nu m)$$

for any decision rule w.

Applying Bonferroni's inequality yields a two-sided result.

The proof of this result once again employs Hoeffding's tail inequality and the swapping subgroup.

We can also obtain results for regular deviation for general loss functions by means of another symmetrization lemma-type result, known as the random subsample lemma. Before discussing that, however, we show how the double sample bounds above allow us to obtain training sample bounds.

5.5.5 Applying the cover to dual sample bounds

Regular deviation of error

We now apply the uniform Occam's razor method to the Vapnik double sample bound of Theorem 5.8 over a minimal γ-cover of $\mathcal{W}$ w.r.t. $d_{p,Q}$, $\mathcal{W}^\star(Q)$. Note that to apply the theorem, we need the elements of the cover also to map into $\{0,1\}$, so that we can not derive a result using external covers.[98] It follows that for zero-one loss functions, with $m = u$,

$$\mathbb{P}_{\tau \in \mathrm{Unif}\, S_{2m}} \left\{ \sup_{w \in \mathcal{W}^\star(Q)} \mathscr{E}_w(\tau(Q), \epsilon) | Q \right\} < \mathcal{N}_{p,Q}(\gamma, \mathcal{W}) \exp\left(- \frac{\epsilon^2 m^2 - 1}{m+1} \right) \quad ,$$

where $\mathscr{E}_w$ is the bound on upper regular deviation of error defined in Theorem 5.8.

In order to extend this result to $\mathcal{W}$, consider any $\tau \in S_{2m}$ and $w \in \mathcal{W}$. Suppose $w^\star \in \mathcal{W}^\star(Q)$ γ-approximates w w.r.t. $d_{p,Q}$. Let the first and second halves of $\tau(Q)$ be denoted by $S_{\tau(Q)}$ and $P_{\tau(Q)}$ respectively, so $\tau(Q) = S_{\tau(Q)} \oplus P_{\tau(Q)}$. Now

$$\begin{aligned} & r_{P_{\tau(Q)}}(w) - r_{S_{\tau(Q)}}(w) \\ = & \left[r_{P_{\tau(Q)}}(w) - r_{P_{\tau(Q)}}(w^\star) \right] + \left[r_{P_{\tau(Q)}}(w^\star) - r_{S_{\tau(Q)}}(w^\star) \right] + \left[r_{S_{\tau(Q)}}(w^\star) - r_{S_{\tau(Q)}}(w) \right] \quad . \end{aligned}$$

The middle term of the right hand side is subject to the bound above. We wish to use the fact that $w^\star$ γ-approximates w to bound the other terms.

[98] In theory, one could allow elements of the cover to lie outside $Q_{\mathcal{W}}$ but to lie in $\{0,1\}^{2m}$. We do not pursue this possibility further, though.

We have

$$\begin{aligned}
& \left[r_{P_{\tau(Q)}}(w) - r_{P_{\tau(Q)}}(w^\star)\right] + \left[r_{S_{\tau(Q)}}(w^\star) - r_{S_{\tau(Q)}}(w)\right] \\
= \; & \frac{1}{m}\sum_{i=1}^{m}[w(p_i) - w^\star(p_i)] + \frac{1}{m}\sum_{i=1}^{m}[w^\star(s_i) - w(s_i)] \\
\leq \; & \frac{1}{m}\sum_{i=1}^{m}|w(p_i) - w^\star(p_i)| + \frac{1}{m}\sum_{i=1}^{m}|w^\star(s_i) - w(s_i)| \\
= \; & 2\frac{1}{2m}\sum_{i=1}^{2m}|w(q_i) - w^\star(q_i)| \\
= \; & 2d_{1,Q}(w, w^\star) \\
\leq \; & 2d_{p,Q}(w, w^\star) \\
< \; & 2\gamma \ , \qquad (5.15)
\end{aligned}$$

where the s_i and p_i are the components of $S_{\tau(Q)}$ and $P_{\tau(Q)}$ respectively. This result holds for general loss functions. When, as here, w and $w^\star$ map into $\{0, 1\}$ we can tighten the last line slightly to

$$\begin{aligned}
2d_{1,Q}(w, w^\star) \; &= \; 2[d_{p,Q}(w, w^\star)]^p \\
&< \; 2\gamma^p \ .
\end{aligned}$$

It follows that

$$e_{P_{\tau(Q)}}(w) - e_{S_{\tau(Q)}}(w) < e_{P_{\tau(Q)}}(w^\star) - e_{S_{\tau(Q)}}(w^\star) + 2\gamma^p$$

so that we obtain the following result: for zero-one loss functions, with $m = u$, we have,

$$\begin{aligned}
& \mathbb{P}_{\tau \in \mathrm{Unif}\, S_{2m}}\left\{\sup_{w \in \mathcal{W}} \mathscr{E}_w(\tau(Q), \epsilon + 2\gamma^p)|Q\right\} \\
& \qquad\qquad < \mathcal{N}_{p,Q}(\gamma, \mathcal{W}) \exp\left(-\frac{\epsilon^2 m^2 - 1}{m+1}\right) \ .
\end{aligned}$$

This result is equivalent for all $1 \leq p < \infty$ due to the relationship between covering numbers for zero-one loss functions given in (5.5).

We have succeeded in obtaining a conditional bound employing a covering number over the finite set Q. We now take the expectation (over all possible

Q w.r.t. D^{2m}) of both sides to obtain a bound which is independent of a specific choice of Q. This yields

$$\begin{aligned} & \mathbb{P}_{S\oplus P\sim D^{2m}}\left\{\sup_{w\in\mathcal{W}}\left[e_P(w)-e_S(w)\right]>\epsilon+2\gamma\right\} \\ = \ & \mathbb{P}_{S\oplus P\sim D^{2m}}\left\{\sup_{w\in\mathcal{W}}\mathscr{E}_w(S\oplus P,\epsilon+2\gamma)\right\} \\ = \ & \mathbb{E}_{Q\sim D^{2m}}\,\mathbb{P}_{\tau\sim\mathrm{Unif}\,S_{m+u}}\left\{\sup_{w\in\mathcal{W}}\mathscr{E}_w(\tau(Q),\epsilon+2\gamma)|Q\right\} \\ < \ & \mathbb{E}_{Q\sim D^{2m}}\,\mathcal{N}_{1,Q}(\gamma,\mathcal{W})\exp\left(-\frac{\epsilon^2m^2-1}{m+1}\right) \ . \end{aligned}$$

Combining this result with the Vapnik symmetrization lemma for regular deviation (Theorem 5.6) yields a uniform bound on the regular deviation between the empirical and true risk of all the decision rules in $\mathcal{W}$:

$$\begin{aligned} & \mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}}\left[e_D(w)-e_S(w)\right]>\epsilon\right\} \\ \le \ & 2\,\mathbb{P}_{S\oplus P\sim D^{2m}}\left\{\sup_{w\in\mathcal{W}}\left[e_P(w)-e_S(w)\right]>\epsilon-\frac{1}{m}\right\} \\ < \ & 2\,\mathbb{E}_{Q\sim D^{2m}}\,\mathcal{N}_{1,Q}(\gamma,\mathcal{W})\exp\left(-\frac{\left(\epsilon-\frac{1}{m}-2\gamma\right)^2m^2-1}{m+1}\right) \ . \qquad (5.16) \end{aligned}$$

Setting the right hand side to δ, we obtain

$$\mathbb{P}_{S\sim D^m}\left\{\begin{array}{ll} & \sup_{w\in\mathcal{W}}\left[e_D(w)-e_S(w)\right] \\ > & 2\gamma+\frac{1+\sqrt{(m+1)\left[\ln 2\,\mathbb{E}_{Q\sim D^{2m}}\,\mathcal{N}_{1,Q}(\gamma,\mathcal{W})-\ln\delta\right]+1}}{m}\end{array}\right\}<\delta \ . \qquad (5.17)$$

As we discuss later, it is often difficult to obtain the covering number $\mathcal{N}_{1,Q}(\gamma,\mathcal{W}))$. If the intention is to bound it by $|Q_{\mathcal{W}}|$ for application, we note that

$$\begin{aligned} \mathcal{N}_{1,Q}(\gamma,\mathcal{W})) & \le \mathcal{N}_{\infty,Q}(\gamma,\mathcal{W})) \\ & \le \lim_{\gamma'\to0^+}\mathcal{N}_{\infty,Q}\left(\gamma',\mathcal{W}\right) \\ & = |Q_{\mathcal{W}}| \ . \end{aligned}$$

This leads to the bound

$$\mathbb{P}_{S\sim D^m}\left\{\begin{array}{ll} & \sup_{w\in\mathcal{W}}[e_D(w)-e_S(w)] \\ > & \frac{1+\sqrt{(m+1)\left[\ln 2\,\mathbb{E}_{Q\sim D^{2m}}\,|Q_{\mathcal{W}}|-\ln\delta\right]+1}}{m}\end{array}\right\}<\delta \ . \qquad (5.18)$$

Relative deviation of error

By following an analysis almost identical to that of the case for regular deviation above, but employing instead the double sample bound for relative deviation (Theorem 5.12) and the Vapnik symmetrization lemma for relative deviation (Theorem 5.7), we obtain a bound on relative deviation. There is one important difference in the derivation though. We restrict γ to be small enough that the γ-cover given Q is specified by $Q_{\mathcal{W}}$. This means that for any w, the $w^\star$ γ-approximating w satisfies $w^\star = w$, so that

$$\frac{e_P(w) - e_S(w)}{\sqrt{e_{S \oplus P}(w)}} - \frac{e_P(w^\star) - e_S(w^\star)}{\sqrt{e_{S \oplus P}(w^\star)}} = 0 \ .$$

If this was not the case, bounding this deviation would be tricky.

The resulting bound, first proved with these constants in Anthony and Shawe-Taylor (1993, Theorem 2.1), is

$$\mathbb{P}_{S \sim D^m} \left\{ \sup_{w \in \mathcal{W}} \frac{e_D(w) - e_S(w)}{\sqrt{e_D(w)}} > \epsilon \right\} < 4 \, \mathbb{E}_{Q \sim D^{2m}} \left| Q_{\mathcal{W}} \right| \exp\left(\frac{-m\epsilon^2}{4} \right) \ , \tag{5.19}$$

for $m > \epsilon^{-2}$.

Derived realistic and realizable bounds on error

This result provides an exponential bound on relative deviation for every decision rule in $\mathcal{W}$, so that we can use the techniques in Section 4.6.1 to obtain corresponding probability statements for the realizable and realistic cases.

In the realistic case, this yields the probability statement

$$\begin{aligned} &\mathbb{P}_{S \sim D^m} \left\{ \exists w \in \mathcal{W} : (e_S(w) \leq (1-\kappa) e_D(w)) \wedge (e_D(w) > \epsilon) \right\} \\ &\qquad\qquad < 4 \, \mathbb{E}_{Q \sim D^{2m}} \left| Q_{\mathcal{W}} \right| \exp\left(\frac{-m\kappa^2\epsilon}{4} \right) \ , \end{aligned} \tag{5.20}$$

for $m > \kappa^{-2}\epsilon^{-1}$.

Setting $\kappa = 1$, we obtain for the realizable case that

$$\mathbb{P}_{S\sim D^m}\left\{\exists w \in \mathcal{W} : (e_S(w) = 0) \wedge (e_D(w) > \epsilon)\right\} \\ < 4\,\mathbb{E}_{Q\sim D^{2m}}\left|Q_{\mathcal{W}}\right| \exp\left(\frac{-m\epsilon}{4}\right) \quad , \tag{5.21}$$

for $m > \epsilon^{-1}$.

We can improve this result by noting that when $e_T(w) = 0$, the upper relative deviation is simply $\sqrt{e_D(w)}$. In this case, the relative deviation exceeds ϵ exactly when the absolute deviation exceeds ϵ^2. Applying this reasoning along with the bound of (5.18), we obtain

$$\mathbb{P}_{S\sim D^m}\left\{\exists w \in \mathcal{W} : (e_S(w) = 0) \wedge (e_D(w) > \epsilon)\right\} \\ < 2\,\mathbb{E}_{Q\sim D^{2m}}\mathcal{N}_{1,Q}(\gamma, \mathcal{W}) \exp\left(-\frac{\left(\sqrt{\epsilon} - \frac{1}{m} - 2\gamma\right)^2 m^2 - 1}{m+1}\right) \quad . \tag{5.22}$$

A direct realizable bound on error

It is possible to improve this further by utilising the fact that $e_T(w) = 0$ and constructing a more efficient dual sample bound for this case by a direct permutation argument.

Theorem 5.14 (Realizable dual sample bound). *For zero-one loss functions,*

$$\mathbb{P}_{\tau\sim \mathrm{Unif}\, S_{m+u}}\left\{\left(e_{S_{\tau(Q)}}(w) = 0\right) \wedge \left(e_{P_{\tau(Q)}}(w) > \epsilon\right) | Q\right\} \leq \frac{\binom{u}{\lceil u\epsilon\rceil}}{\binom{m+u}{\lceil u\epsilon\rceil}} \\ \leq \left(\frac{u}{m+u}\right)^{\lceil u\epsilon\rceil} \\ \leq \left(\frac{u}{m+u}\right)^{u\epsilon} \quad ,$$

for any decision rule w.

This result is a generalization of a result implicit in Blumer et al. (1986, Lemma 5), to allow $m \neq u$.[99] The proof is very similar to the original, so we omit it here.

Following the same techniques as for earlier bounds in this section, we obtain

$$\begin{aligned}&\mathbb{P}_{S\oplus P\sim D^{m+u}}\left\{\exists w \in \mathcal{W} : (e_S(w) = 0) \wedge (e_P(w) > \epsilon)\right\}\\ &\qquad< \mathbb{E}_{Q\sim D^{m+u}}\mathcal{N}_{1,Q}(\gamma, \mathcal{W})\left(\frac{u}{m+u}\right)^{\left\lceil u\left(\epsilon - \frac{m+u}{u}\gamma\right)\right\rceil} \quad . \qquad (5.23)\end{aligned}$$

One difference between this bound and others presented earlier is the expression $\epsilon - \frac{m+u}{u}\gamma$ instead of $\epsilon - 2\gamma$. This occurs because one can not obtain a reduction in risk from zero on the first half of the sample, and because the two samples are no longer of equal size. To obtain a final result, we need to combine this result with a result like a symmetrization lemma. However, in its current form, we can not combine this result with the general symmetrization lemma of (5.7). Fortunately, we have derived an analogous symmetrization lemma in this case.

Theorem 5.15 (Realizable symmetrization lemma for risk). *Consider $\epsilon \in [0, 1)$. Suppose there exist constants $\alpha, \beta > 0$ such that*

$$\mathbb{P}_{P\sim D^u}\left\{r_P(w) > \alpha r_D(w)\right\} > \beta$$

for every $w \in \mathcal{W}$ satisfying $r_D(w) > \epsilon$. Then

$$\begin{aligned}&\mathbb{P}_{S\sim D^m}\left\{\exists w \in \mathcal{W} : (r_S(w) = 0) \wedge (r_D(w) > \epsilon)\right\}\\ &\qquad\leq \beta^{-1}\,\mathbb{P}_{S\oplus P\sim D^{m+u}}\left\{\exists w \in \mathcal{W} : (r_S(w) = 0) \wedge (r_P(w) > \alpha\epsilon)\right\} \quad .\end{aligned}$$

The proof of this result is similar in nature to that of Theorem 5.2, although some changes are necessary to accommodate the conditional reasoning.

Proof. Let w_S be a decision rule in $\mathcal{W}$ depending on S such that $r_S(w_S) = 0$ and $r_D(w_S) > \epsilon$ when $S \in \mathcal{Z}^m$ satisfies

$$\exists w \in \mathcal{W} : (r_S(w) = 0 \wedge r_D(w) > \epsilon) \quad .$$

[99] We later found the result to be implicit in the proof of Shawe-Taylor et al. (1993, Theorem 3.3).

Denote this set of samples by R_ϵ.

Now, since r_S and r_P are independent, w_S is independent of r_P. Thus, by the assumptions of the lemma,

$$\mathbb{P}_{P|S}\left\{r_P(w_S) > \alpha r_D(w_S)|S\right\} > \beta$$

for any S in R_ϵ.

Now

$$\begin{aligned}
& \beta\,\mathbb{P}_{S\sim D^m}\left\{S \in R_\epsilon\right\} \\
= \; & \int_{S\in R_\epsilon} \beta\, d(D^m) \\
< \; & \int_{S\in R_\epsilon} \mathbb{P}_{P|S}\left\{r_P(w_S) > \alpha r_D(w_S)|S\right\}\, d(D^m) \\
= \; & \int_{S\in \mathcal{Z}^m} \mathbb{P}_{P|S}\left\{r_P(w_S) > \alpha r_D(w_S)|S\right\} I(S \in R_\epsilon)\, d(D^m) \\
= \; & \mathbb{P}_{S\oplus P\sim D^{m+u}}\left\{(r_P(w_S) > \alpha r_D(w_S)) \wedge (S \in R_\epsilon)\right\} \\
= \; & \mathbb{P}_{S\oplus P\sim D^{m+u}}\left\{(r_P(w_S) > \alpha r_D(w_S)) \wedge (r_S(w_S) = 0) \wedge (r_D(w) > \epsilon)\right\} \\
\leq \; & \mathbb{P}_{S\oplus P\sim D^{m+u}}\left\{(r_S(w_S) = 0) \wedge (r_P(w_S) > \alpha\epsilon)\right\} \\
\leq \; & \mathbb{P}_{S\oplus P\sim D^{m+u}}\left\{\exists w \in \mathcal{W} : (r_S(w) = 0) \wedge (r_P(w) > \alpha\epsilon)\right\} \quad .
\end{aligned}$$

Dividing throughout by β yields the result. □

Applying this lemma to (5.23), we obtain:

$$\begin{aligned}
& \mathbb{P}_{S\sim D^m}\left\{\exists w \in \mathcal{W} : (e_S(w) = 0) \wedge (e_D(w) > \epsilon)\right\} \\
& \qquad \leq \beta^{-1}\, \mathbb{E}_{Q\sim D^{m+u}}\, \mathcal{N}_{1,Q}(\gamma, \mathcal{W}) \left(\frac{u}{m+u}\right)^{\left\lceil u\left(\alpha\epsilon - \frac{m+u}{u}\gamma\right)\right\rceil} \quad . \qquad (5.24)
\end{aligned}$$

In order to apply this lemma, we need to find valid choices of α and β to apply the symmetrization lemma. As with the general symmetrization lemma, we can obtain α in terms of β and u (and in this case ϵ) by employing concentration inequalities.

Simple manipulation shows that

$$\mathbb{P}_{P\sim D^u}\left\{r_P(w) > \alpha r_D(w)\right\} > \beta$$

is implied by the probability inequality

$$\mathbb{P}_{P\sim D^u}\left\{r_D(w) - r_P(w) > (1-\alpha)r_D(w)\right\} \leq (1-\beta) \ .$$

From this form, one can obtain a value for β by employing the Chebyshev inequality[100]: since

$$\mathbb{E}_{P\sim D^u}\, r_P(w) = r_D(w)$$

and

$$\mathbb{V}_{P\sim D^u}\, r_P(w) \leq \frac{r_D(w)[1-r_D(w)]}{u} \ ,$$

we have that

$$\begin{aligned} & \mathbb{P}_{P\sim D^u}\left\{r_D(w) - r_P(w) > (1-\alpha)r_D(w)\right\} \\ \leq\ & \frac{r_D(w)[1-r_D(w)]}{u(1-\alpha)^2[r_D(w)]^2} \\ =\ & \frac{1-r_D(w)}{u(1-\alpha)^2 r_D(w)} \ , \end{aligned}$$

which for $r_D(w) > \epsilon$ exceeds

$$\frac{\frac{1}{\epsilon}-1}{u(1-\alpha)^2} \ .$$

Setting this equal to $1-\beta$, we obtain that the conditions of the lemma hold for a given α with

$$\beta = 1 - \frac{\frac{1}{\epsilon}-1}{u(1-\alpha)^2} \ .$$

Solving this for α yields

$$\alpha_{C-R}(u,\beta,\epsilon) = 1 - \sqrt{\frac{\frac{1}{\epsilon}-1}{u(1-\beta)}}$$

where $C-R$ stands for "Chebyshev-Realizable".

Using this choice, and considering the bound as $\gamma \to 0^+$, one obtains, for any $\epsilon \in [0,1)$, $\beta \in (0,1]$:

$$\begin{aligned} & \mathbb{P}_{S\sim D^m}\left\{\exists w \in \mathcal{W} : (e_S(w) = 0) \wedge (e_D(w) > \epsilon)\right\} \\ & \qquad \leq \beta^{-1}\, \mathbb{E}_{Q\sim D^{m+u}}\, |Q_{\mathcal{W}}| \left(\frac{u}{m+u}\right)^{\lceil u\alpha_{C-R}(u,\beta,\epsilon)\epsilon\rceil} \ . \qquad (5.25) \end{aligned}$$

[100] Other options can be considered: for example, an improvement for the case $\beta = \frac{1}{2}$ is mentioned in Blumer et al. (1989, Lemma A.2.1).

By employing a number of relaxations, choosing $\beta = \frac{1}{2}$ and calculating a suitable u for a given ϵ such that (a relaxation of) $\alpha_{C-R}(u, \beta, \epsilon)$ is convenient, one can obtain an analog of the bound of Shawe-Taylor et al. (1993, Theorem 3.3). The details are omitted, since they require concepts not yet introduced. However, this result is a clear example of the benefits to be reaped by considering more general formulation of symmetrization lemmas and dual sample bounds.

We conclude this section by noting that while the realizable symmetrization lemma of Theorem 5.15 is new, the bound in Shawe-Taylor et al. (1993) is based on a very similar result, which we present in a general form below.

Theorem 5.16 (Alternative realizable symmetrization lemma for risk). *Consider $\epsilon \in [0, 1)$. Suppose there exist constants $\alpha, \beta > 0$ such that*

$$\mathbb{P}_{P \sim D^u} \left\{ r_D(w) - r_P(w) < \alpha \right\} > \beta$$

for every $w \in \mathcal{W}$ satisfying $r_D(w) > \epsilon$. Then

$$\begin{aligned} &\mathbb{P}_{S \sim D^m} \left\{ \exists w \in \mathcal{W} : (r_S(w) = 0) \wedge (r_D(w) > \epsilon) \right\} \\ &\qquad \leq \beta^{-1}\, \mathbb{P}_{S \oplus P \sim D^{m+u}} \left\{ \exists w \in \mathcal{W} : (r_S(w) = 0) \wedge (r_P(w) > \epsilon - \alpha) \right\} \quad . \end{aligned}$$

The earliest form of this result employed $m = u$, $\beta = \frac{1}{2}$ and $\alpha = \frac{\epsilon}{2}$, with a sample size restriction of $m > \frac{8}{\epsilon}$ obtained by employing Chebyshev's inequality — see Haussler (1986, Lemma 3.5). For further development of this result, see Blumer et al. (1986, 1989), Shawe-Taylor et al. (1993). To our knowledge, the result has not yet been presented in this generality.

Regular deviation of risk

Next we turn to employing the double sample bound for risk in Theorem 5.9. First we note that since the bound holds for $w \in \mathcal{W}$, we can not employ an external cover. Applying the Occam's razor method to the bound over a minimal γ-cover of $\mathcal{W}$ w.r.t. $d_{p,Q}$, $\mathcal{W}^\star$, yields

$$\mathbb{P}_{\tau \sim \mathrm{Unif}\, S_{2m}^\star} \left\{ \sup_{w \in \mathcal{W}^\star} \mathscr{E}_w(\tau(Q), \epsilon) | Q \right\} \\ < 2m \mathcal{N}_{p,Q}(\gamma, \mathcal{W}) \exp\left(-\frac{\epsilon^2 m^2 - 1}{m+1} \right)$$

where $\mathscr{E}_w$ is defined in Theorem 5.9. From (5.15), it follows that

$$r_{P_{\tau(Q)}}(w) - r_{S_{\tau(Q)}}(w) < r_{P_{\tau(Q)}}(w^\star) - r_{S_{\tau(Q)}}(w^\star) + 2\gamma$$

so that

$$\mathbb{P}_{\tau \sim \mathrm{Unif}\, S_{2m}^\star} \left\{ \sup_{w \in \mathcal{W}} \mathscr{E}_w(\tau(Q), \epsilon + 2\gamma) | Q \right\} \\ < 2m \mathcal{N}_{p,Q}(\gamma, \mathcal{W}) \exp\left(-\frac{\epsilon^2 m^2 - 1}{m+1} \right) \quad .$$

Taking expectations with respect to Q of both sides yields

$$\mathbb{P}_{S \oplus P \sim D^{2m}} \left\{ \sup_{w \in \mathcal{W}} [r_P(w) - r_S(w)] > \epsilon + 2\gamma \right\} \\ < 2m\, \mathbb{E}_{Q \sim D^{2m}} [\mathcal{N}_{p,Q}(\gamma, \mathcal{W})] \exp\left(-\frac{\epsilon^2 m^2 - 1}{m+1} \right) \quad ,$$

so that applying the general symmetrization lemma for regular deviation in (5.7) yields

$$\mathbb{P}_{S \sim D^m} \left\{ \sup_{w \in \mathcal{W}} [r_D(w) - r_S(w)] > \epsilon \right\} \\ < \frac{2m}{\beta} \mathbb{E}_{Q \sim D^{2m}} [\mathcal{N}_{p,Q}(\gamma, \mathcal{W})] \exp\left(-\frac{[\epsilon - \alpha(m,\beta) - 2\gamma]^2 m^2 - 1}{m+1} \right) \quad , \tag{5.26}$$

for any $0 < \beta \leq 1$; setting the right hand side to δ yields

$$\mathbb{P}_{S \sim D^m} \left\{ \begin{array}{c} \sup_{w \in \mathcal{W}} [r_D(w) - r_S(w)] > 2\gamma + \alpha(m,\beta) \\ + \quad \frac{\sqrt{(m+1)\left[\ln 2m\, \mathbb{E}_{Q \sim D^{2m}} \mathcal{N}_{1,Q}(\gamma, \mathcal{W}) - \ln \beta\delta\right] + 1}}{m} \end{array} \right\} < \delta \quad . \tag{5.27}$$

These results, although implicit in the derivations in the literature, are typically not stated in this generality. If we set $\beta = \frac{1}{2}$, requiring $m \geq 2\epsilon^{-2}$ implies $\alpha_{C'}(m,\beta) \leq \frac{\epsilon}{2}$. Comparing this result with the error bound in

(5.16), we see that the bound for general risk has an additional factor of $2m$ in the coefficient, and that the bound decreases more slowly with respect to ϵ due to an extra factor $\frac{1}{2}$ before the ϵ in the bound. The factor $\frac{1}{2}$ and a factor 2 in the coefficient are due to the use of the general symmetrization lemma rather than the Vapnik symmetrization lemma for regular deviation. The other factor m is introduced by employing the double sample bound for regular deviation of risk, rather than the Vapnik double sample bound.

If one further selects $\gamma = \frac{\epsilon}{3}$, one obtains essentially the result in Alon et al. (1993, Lemma 3.3)[101].

In the case $u = m$, it is not difficult to show that the bound of Theorem 5.10 is tighter than that of Theorem 5.9 roughly when[102]

$$\epsilon < \sqrt{\frac{2\ln(2m)}{m}} \ .$$

Due to the quick decay of the right hand side, the dual sample bound is generally not preferable for large sample sizes *if applied with $u = m$.* This is simply because the fraction $\frac{u}{m+u}$ is too small in this case. If we increase u, this fraction increases, making the dual sample bound more attractive as the distribution P becomes a better approximation to D. However, as u increases, the symmetrization lemma becomes weaker, so a trade-off will be necessary for the choice of u if the dual sample bound is to be employed to obtain tight bounds.

Obtaining a bound for regular deviation using the dual sample bound of Theorem 5.10 follows exactly the same lines as the one just derived based on Theorem 5.9, except that the argument is based on the symmetric group instead of the swapping subgroup, we employ the symmetrization lemma with unequal samples, and the extension from a cover to the whole of $\mathcal{W}$ is not so straightforward.

In order to extend the bound to $\mathcal{W}$ from a cover $\mathcal{W}^\star$, consider any $\tau \in S_{2m}$ and $w \in \mathcal{W}$. Suppose $w^\star \in \mathcal{W}^\star$ γ-approximates w w.r.t. $d_{p,Q}$. Let the first

[101] This result is one-sided, but a two-sided version can be seen to hold with a coefficient of 8. The coefficient of 12 in the referenced paper should actually be 8. See footnote 94.

[102] We approximate the bound in the double sample bound by $2m\exp(-\epsilon^2 m)$.

m points of $\tau(Q)$ be denoted by $S_{\tau(Q)}$. As before, we can write

$$\begin{aligned} & r_{P_{\tau(Q)}}(w) - r_{S_{\tau(Q)}}(w) \\ = & \left[r_{P_{\tau(Q)}}(w) - r_{P_{\tau(Q)}}(w^\star)\right] + \left[r_{P_{\tau(Q)}}(w^\star) - r_{S_{\tau(Q)}}(w^\star)\right] + \left[r_{S_{\tau(Q)}}(w^\star) - r_{S_{\tau(Q)}}(w)\right] \ , \end{aligned}$$

and we wish to use the fact that $w^\star$ γ-approximates w to bound the sum of the first and last terms on the right hand side. Using

$$r_{P_{\tau(Q)}}(w) - r_{S_{\tau(Q)}}(w) = \frac{m+u}{u}[r_Q(w) - r_{S_{\tau(Q)}}(w)]$$

for any decision rule, we can write the sum of these terms as

$$\begin{aligned} & \frac{m+u}{u}\left[(r_Q(w) - r_Q(w^\star)) - (r_{S_{\tau(Q)}}(w) - r_{S_{\tau(Q)}}(w^\star))\right] \\ = & \frac{m+u}{u}\left[\left(\frac{1}{m+u}\sum_{i=1}^{m+u}[w(q_i) - w^\star(q_i)]\right) + \left(\frac{1}{m}\sum_{i=1}^{m}[w^\star(s_i) - w(s_i)]\right)\right] \\ \leq & \frac{m+u}{u}\left[\left(\frac{1}{m+u}\sum_{i=1}^{m+u}|w(q_i) - w^\star(q_i)|\right) + \left(\frac{1}{m}\sum_{i=1}^{m+u}|w^\star(q_i) - w(q_i)|\right)\right] \\ = & \frac{m+u}{u}\left[d_{1,Q}(w,w^\star) + \frac{m+u}{m}d_{1,Q}(w,w^\star)\right] \\ < & \frac{(2m+u)(m+u)}{um}\gamma \ , \end{aligned}$$

where the s_i are the components of $S_{\tau(Q)}$ and the q_i are the components of Q. It follows that

$$r_{P_{\tau(Q)}}(w) - r_{S_{\tau(Q)}}(w) \leq r_{P_{\tau(Q)}}(w^\star) - r_{S_{\tau(Q)}}(w^\star) + \frac{(2m+u)(m+u)}{um}\gamma \ ,$$

with the same result holding for lower deviation.

Applying this, we obtain the following result.

Theorem 5.17 (Dual sample bound on regular deviation of risk). *Let $\alpha(u,\beta)$ be chosen such that $(\alpha(u,\beta),\beta)$ satisfies the requirements of the symmetrization lemma of (5.7). Let $0 < \beta \leq 1$.*

If $1 > \epsilon > \alpha(u,\beta)$ and $u, \gamma > 0$ are chosen such that

$$\epsilon - \alpha(u,\beta) - \frac{(2m+u)(m+u)}{um}\gamma > 0 \ ,$$

then

$$\begin{aligned}\mathbb{P}_{S\sim D^m}&\left\{\sup_{w\in\mathcal{W}}\left[r_D(w)-r_S(w)\right]>\epsilon\right\}\\ &< \begin{array}{l}\frac{1}{\beta}\,\mathbb{E}_{Q\sim D^{m+u}}\,\mathcal{N}_{p,Q}(\gamma,\mathcal{W})\\ \qquad\exp\left(-2m\left[\epsilon-\alpha(u,\beta)-\frac{(2m+u)(m+u)}{um}\gamma\right]^2\left(\frac{u}{m+u}\right)^2\right)\end{array}\quad .\end{aligned}$$

Unfortunately, the factor $\frac{(2m+u)(m+u)}{um}$ can be prohibitively large, particularly if we wish to select a large value of u. However, in cases where the loss function has a finite range, such as zero-one loss functions, we can make γ arbitrarily small, so that this factor does not influence the results below a certain resolution. Furthermore, some well-behaved function classes which have infinite ranges may exhibit similar behavior.

An alternative option is to use covers with respect to $d_{\infty,Q}$, if one can obtain a suitable estimate of $\mathbb{E}_{Q\sim D^{m+u}}\,\mathcal{N}_{\infty,Q}(\gamma,\mathcal{W})$. It is not difficult to show that if we replace p by ∞ in the covering numbers of Theorem 5.17, the result still holds if we replace $\frac{(2m+u)(m+u)}{um}\gamma$ by 2γ, yielding a bound of

$$\frac{1}{\beta}\,\mathbb{E}_{Q\sim D^{m+u}}\,\mathcal{N}_{\infty,Q}(\gamma,\mathcal{W})\exp\left(-2m\left[\epsilon-\alpha(u,\beta)-2\gamma\right]^2\left(\frac{u}{m+u}\right)^2\right)\quad .$$

Note that the reduction in the factor before γ is offset by the increased covering number obtained by using a more restrictive metric.

Theorem 5.17 generalizes somewhat the bound presented for error in (1.6) of Devroye (1982). First, that result selects $u=m^2-m$. Then it obtains β as the solution to $\alpha_{C'}(u,\beta)=\frac{1}{m}$, i.e.

$$\beta = 1-\frac{m^2}{4u}\quad .$$

Applying these settings, and taking the limit as $\gamma \to 0^+$, yields

$$\begin{aligned}
&\mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}}\left[e_D(w) - e_S(w)\right] > \epsilon\right\} \\
<\ & \frac{1}{1-\frac{m}{4(m-1)}}\,\mathbb{E}_{Q\sim D^{m+u}}\left|Q_{\mathcal{W}}\right| \exp\left(-2m\left[\epsilon - \frac{1}{m}\right]^2\left(\frac{m-1}{m}\right)^2\right) \\
=\ & \frac{4m-4}{3m-4}\,\mathbb{E}_{Q\sim D^{m+u}}\left|Q_{\mathcal{W}}\right| \exp\left(-2m\left[\epsilon - \frac{1}{m}\right]^2\left[1-\frac{1}{m}\right]^2\right) \\
<\ & \frac{4}{3}\,\mathbb{E}_{Q\sim D^{m+u}}\left|Q_{\mathcal{W}}\right| \exp\left(-2m\left[\epsilon^2 - \frac{2\epsilon}{m}\right]\left[1-\frac{2}{m}\right]\right) \\
<\ & \frac{4}{3}\,\mathbb{E}_{Q\sim D^{m+u}}\left|Q_{\mathcal{W}}\right| \exp\left(-2m\left[\epsilon^2 - \frac{2\epsilon^2}{m} - \frac{2\epsilon}{m}\right]\right) \\
=\ & \frac{4}{3}\exp\left(4\epsilon(\epsilon+1)\right)\mathbb{E}_{Q\sim D^{m+u}}\left|Q_{\mathcal{W}}\right|\exp(-2m\epsilon^2)
\end{aligned}$$

when $1 > \epsilon > \frac{1}{n}$. Doubling the coefficient to obtain a two-sided result yields (2.2) of Devroye (1982) except for the slightly improved constant of $\frac{8}{3}$ instead of 4 (which results in their paper from an unnecessary relaxation). In addition, the constraint $1 > \epsilon > \frac{1}{m}$, while not stated directly in their theorem, is necessary for their proof to hold.

It is instructive to compare this result to the bound on regular deviation of error in (5.16) for small enough γ. The bound in this result has a (potentially much) larger coefficient, and the covering numbers are with respect to a sample of size m^2, rather than $2m$. On the other hand, the negative exponential term in this result decays twice as quickly. Thus, if the difference in growth of the covering numbers is outweighed by the rate of decay of the negative exponential term, the second bound is better asymptotically. We shall see later that if these bounds are to be non-trivial, the gain in the negative exponential term will bear more weight. However, a large sample is necessary in practice to obtain improvements using this result.

We further note that this is the first straightforward covering number for regular deviation of risk exhibiting this asymptotic rate of decay. In particular, all the bounds which are asymptotically competitive with this bound make use of either chaining (see Section 5.7) or applications of more advanced concentration inequalities.

Bartlett-Lugosi deviation of risk

Next, we present a bound on the B-L deviation. This result employs similar techniques to those presented above, but combines the B-L deviation symmetrization lemma with a double sample bound for relative deviation. In order to do this, a result is needed relating probabilities w.r.t. the two deviations when $\nu \neq 0$. Let $\mathcal{W}^\star$ be a γ-cover of $\mathcal{W}$ w.r.t. $d_{p,Q}$. Applying the double sample bound for relative deviation of risk (Theorem 5.12) to the elements of this cover (note that this restricts us to $u = m$), and applying the uniform Occam's razor method, one obtains

$$\mathbb{P}_{\tau \sim \mathrm{Unif}\, S^\star_{2m}} \left\{ \exists w \in \mathcal{W}^\star : \mathscr{E}^0_w(\tau(Q), \epsilon) | Q \right\} < |\mathcal{W}^\star| \exp\left(\frac{-\epsilon^2 m}{4} \right) \quad , \tag{5.28}$$

where $\mathscr{E}^\nu_w(S \oplus P, \epsilon)$ denotes the event

$$\frac{r_P(w) - r_S(w) - \nu}{\sqrt{r_{S \oplus P}(w)}} > \epsilon \ .$$

Next, consider any decision rule $w \in \mathcal{W}$, any S, and any P. Let $w^\star$ denote an element of $\mathcal{W}^\star$ with $d_{p,Q}(w, w^\star) < \gamma$. (Note here that $Q = S \oplus P$). It will be convenient to handle the case $p = \infty$. Bounds for other p can be obtained by relationships between covering numbers. In that case, we have that $w(q) \leq w^\star(q) + \gamma$ for any $q \in Q$, so that $r_S(w^\star) - \gamma \leq r_S(w) \leq r_S(w^\star) + \gamma$ and similarly for P.

We can write

$$\begin{aligned}
& \frac{r_P(w) - r_S(w) - \nu}{\sqrt{r_{S\oplus P}(w)}} \\
= \;& \sqrt{2}\frac{\left(r_P(w) - \frac{\nu}{2}\right) - \left(\frac{\nu}{2} + r_S(w)\right)}{\sqrt{\left(r_P(w) - \frac{\nu}{2}\right) + \left(\frac{\nu}{2} + r_S(w)\right)}} \\
\leq \;& \sqrt{2}\frac{\left(r_P(w^\star) + \gamma - \frac{\nu}{2}\right) - \left(\frac{\nu}{2} + r_S(w)\right)}{\sqrt{\left(r_P(w^\star) + \gamma - \frac{\nu}{2}\right) + \left(\frac{\nu}{2} + r_S(w)\right)}} \qquad (5.29) \\
= \;& \sqrt{2}\frac{r_P(w^\star) + \gamma - \nu - r_S(w)}{\sqrt{r_P(w^\star) + \gamma + r_S(w)}} \\
= \;& \sqrt{2}\frac{\left(r_P(w^\star) + \gamma - \frac{\nu}{2}\right) - \left(\frac{\nu}{2} + r_S(w)\right)}{\sqrt{\left(r_P(w^\star) + \gamma - \frac{\nu}{2}\right) + \left(\frac{\nu}{2} + r_S(w)\right)}} \\
\leq \;& \sqrt{2}\frac{\left(r_P(w^\star) + \gamma - \frac{\nu}{2}\right) - \left(\frac{\nu}{2} + r_S(w^\star) - \gamma\right)}{\sqrt{\left(r_P(w^\star) + \gamma - \frac{\nu}{2}\right) + \left(\frac{\nu}{2} + r_S(w^\star) - \gamma\right)}} \qquad (5.30) \\
= \;& \frac{r_P(w^\star) - r_S(w^\star) - (\nu - 2\gamma)}{\sqrt{r_{S\oplus P}(w^\star)}}\ ,
\end{aligned}$$

where (5.29) follows from the monotonic increasing behaviour of $\frac{v-c}{\sqrt{v+c}}$ in $v > -c$ when $c \geq 0$, and (5.30) follows from the monotonic decreasing behaviour of $\frac{c-v}{\sqrt{c+v}}$ in $v > -c$ when $c \geq 0$.

It follows that for $\gamma \leq \frac{\nu}{2}$, $\mathscr{E}_w^\nu(S \oplus P, \epsilon)$ only occurs if $\mathscr{E}_{w^\star}^0(S \oplus P, \epsilon)$ occurs. Finally, choosing $\mathcal{W}^\star$ to be a minimal cover, and taking the expectation w.r.t. the double sample Q, one obtains

$$\begin{aligned}
& \mathbb{P}_{S\oplus P\sim D^{2m}}\left\{\sup_{w\in\mathcal{W}} \frac{r_P(w) - r_S(w) - \nu}{\sqrt{\frac{1}{2}(r_P(w) + r_S(w))}} > \epsilon\right\} \\
& \qquad < \mathbb{E}_{Q\sim D^{2m}}\,\mathcal{N}_{\infty,Q}\left(\frac{\nu}{2}, \mathcal{W}\right)\exp\left(\frac{-\epsilon^2 m}{4}\right)\ .
\end{aligned}$$

Combining this with the Bartlett-Lugosi symmetrization lemma yields the following result, where ψ_ν^U denotes the upper B-L ν-deviation measure:

$$\begin{aligned}
& \mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}} \psi_\nu^U(r_D(w), r_S(w)) > \epsilon\right\} \\
& \qquad \leq \beta^{-1}\,\mathbb{E}_{Q\sim D^{2m}}\,\mathcal{N}_{\infty,Q}\left(\frac{\nu - \alpha(m,\beta)}{2}, \mathcal{W}\right)\exp\left(\frac{-\epsilon^2 m}{4}\right)\ , \qquad (5.31)
\end{aligned}$$

where $\alpha(m, \beta)$ and β satisfy the condition of the symmetrization lemma.

An appropriate choice for α, used to derive the special case of this result presented in Bartlett and Lugosi (1999), is $\alpha_{CC'}$. A similar result to the one presented here can be obtained by a similar derivation combined with the form of the Bartlett-Lugosi symmetrization lemma appropriate for lower bounds.

Bartlett and Lugosi (1999) use these bounds to further obtain bounds on the P-H ν-deviation. However, they claim that the resulting bounds are not as tight as a direct argument, which we shall present next.

Pollard-Haussler deviation of risk

The bounds from the direct argument employ a pseudometric we have not yet encountered. To understand the metric, we present the following definition, based on Haussler (1992, Definition 12).

Definition 5.2 (Haussler extension of a metric). Let d be a metric defined on $\mathbb{R}^+$. The Haussler extension of d, denoted d^H, is defined on $(\mathbb{R}^+)^{2m}$ by

$$d^H(v_1, v_2) = \sup_{\tau \in S^\star_{2m}} \left[d\left(\phi_1(\tau, v_1), \phi_1(\tau, v_2)\right) + d\left(\phi_2(\tau, v_1), \phi_2(\tau, v_2)\right) \right] \quad ,$$

where

$$\phi_1(\tau, v) = \frac{1}{m} \sum_{i=1}^{m} v^{(\tau(i))}$$

and

$$\phi_2(\tau, v) = \frac{1}{m} \sum_{i=m+1}^{2m} v^{(\tau(i))} \quad .$$

Haussler (1992) verifies that d^H is a pseudometric. By employing a cover of $Q_{\mathcal{W}}$ w.r.t. this Haussler extension of the P-H metric, d^H_ν, one is able to control the two-sided P-H ν-deviation of all the decision rules in terms of those of the cover. The following derivation is a simple extension of a portion of the proof of Haussler (1992, Lemma 13).

Applying the double sample bound for two-sided P-H deviation of risk (Theorem 5.13) to a minimal γ-cover R of $Q_{\mathcal{W}}$ w.r.t. d_ν^H, one obtains that

$$\mathbb{P}_{\tau \sim \mathrm{Unif}\, S_{2m}^\star} \{\exists w \in \mathcal{W}^\star : \mathscr{E}_w(\tau(Q), \epsilon) | Q\} < 2\mathcal{N}(\gamma, \mathcal{W}, d_\nu^H) \exp(-2\epsilon^2 \nu m)$$

where $\mathscr{E}_w(S \oplus P, \epsilon)$ is the event

$$d_\nu(r_P(w), r_S(w)) > \epsilon \ ,$$

and $\mathcal{W}^\star = \{w \in \mathcal{W} : Q_w \in R\}$.

Based on the definition of d_ν^H, it can be shown that if

$$d_\nu(r_P(w), r_S(w)) > \epsilon$$

for some $w \in \mathcal{W}$, the corresponding $w^\star \in \mathcal{W}^\star$ satisfies

$$d_\nu(r_P(w^\star), r_S(w^\star)) > \epsilon - \gamma \ .$$

(This is a general property of the Haussler extension of a metric).

Thus

$$\begin{aligned} &\mathbb{P}_{\tau \sim \mathrm{Unif}\, S_{2m}^\star} \{\exists w \in \mathcal{W} : \mathscr{E}_w(\tau(Q), \epsilon) | Q\} \\ &\qquad\qquad < 2\mathcal{N}(\gamma, Q_{\mathcal{W}}, d_\nu^H) \exp(-2(\epsilon - \gamma)^2 \nu m) \ . \end{aligned} \tag{5.32}$$

Taking the expectation w.r.t. Q and applying the Haussler symmetrization lemma, one obtains

$$\begin{aligned} &\mathbb{P}_{S \sim D^m} \left\{ \sup_{w \in \mathcal{W}} d_\nu(r_D(w), r_S(w)) > \epsilon \right\} \\ &\qquad \le 2\beta^{-1}\, \mathbb{E}_{Q \sim D^{2m}} \mathcal{N}(\gamma, Q_{\mathcal{W}}, d_\nu^H) \exp\left(-2(\epsilon - \alpha(m, \beta) - \gamma)^2 \nu m\right) \ , \end{aligned}$$

where $\alpha(m, \beta)$ and β satisfy the requirements of the symmetrization lemma.

A problem with this bound is the covering numbers of $Q_{\mathcal{W}}$ w.r.t. d_ν^H. This problem is slightly alleviated by the following relationship between d_ν^H and the L^1 metric, from Haussler (1992, Lemma 14).

Theorem 5.18. *For $v_1, v_2 \in (\mathbb{R}^+)^{2m}$, and $\nu > 0$,*

$$d_\nu^H(v_1, v_2) \leq \frac{2}{\nu} d_{\ell^1_{2m}}(v_1, v_2) \ ,$$

where

$$d_{\ell^1_{2m}}(v_1, v_2) = \frac{1}{2m} \sum_{i=1}^{2m} |v_1^{(i)} - v_2^{(i)}| \ .$$

A consequence of this result is that the covering number $\mathcal{N}(\gamma, Q_{\mathcal{W}}, d_\nu^H)$ can be replaced by the covering number $\mathcal{N}_{1,Q}(\frac{\gamma\nu}{2}, \mathcal{W})$, which is similar to the other covering numbers we have employed.

As for obtaining an appropriate function α, the reader is referred to the discussion in Section 4.7. As an example, from the bound in (4.9), we obtain that α and β satisfying

$$1 - \beta = \exp\left(\frac{-18m\alpha^2\nu}{(3+\alpha)^2}\right)$$

are suitable choices. Inverting this numerically is an option, as described in Section 4.7. If a simpler alternative is desired, one can replace the right hand side by

$$\exp\left(-\frac{9}{8}m\alpha^2\nu\right) \ ,$$

which yields the function

$$\alpha_{BPH}(m, \beta, \nu) = \sqrt{\frac{-8\ln(1-\beta)}{9m\nu}} \ .$$

Putting this together yields the bound (for $\beta \in (0, 1]$),

$$\begin{aligned} &\mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}} d_\nu(r_D(w), r_S(w)) > \epsilon\right\} \\ &\leq 2\beta^{-1}\,\mathbb{E}_{Q\sim D^{2m}}\,\mathcal{N}_{1,Q}\left(\frac{\gamma\nu}{2}, \mathcal{W}\right)\exp\left(-2(\epsilon - \alpha_{BPH}(m,\beta,\nu) - \gamma)^2\nu m\right) \ . \end{aligned} \tag{5.33}$$

Setting $\beta = \frac{1}{2}$, using the Chebyshev inequality (instead of α_{BPH}) to obtain $\alpha = \frac{\epsilon}{2}$ for $m \geq \frac{2}{\epsilon^2\nu}$, and setting $\gamma = \frac{\epsilon}{4}$ essentially yields Haussler (1992, Theorem 3), barring the improvement we discuss next.

We have omitted one potential improvement to the results which we have presented here: the usage of replacing any trivial upper bound on probability by 1. This can be particularly useful when multiplying small probabilities by potentially large covering numbers. Examples of this approach are provided in Haussler (1992, Theorem 3), and some results in Pollard (1984). To give a flavour of these improvements, we shall simply state the improved bound obtained for the P-H deviation above.

For $\beta \in (0, 1]$, we have

$$\begin{aligned}&\mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}} d_\nu(r_D(w), r_S(w)) > \epsilon\right\}\\ &\leq \min\left\{\beta^{-1}\min\left\{2\,\mathbb{E}_{Q\sim D^{2m}}\mathcal{N}_{1,Q}\left(\frac{\gamma\nu}{2}, \mathcal{W}\right)\exp\left(-2(\epsilon - \alpha_{BPH}(m,\beta,\nu) - \gamma)^2\nu m\right), 1\right\}, 1\right\} \quad .\end{aligned}$$

We shall not pursue this issue further, however, because in practice there are very few opportunities to employ this strengthening.

5.5.6 The random subsample lemma and bound

The random subsample lemma, which we present next, is a well-established result which allows us to obtain bounds for general loss functions. The approach based on this lemma originated in the early 1980's, with Vladimir Koltchinskii (Koltchinskii, 1982) and David Pollard (Pollard, 1982) introducing it independently to prove central limit theorems for empirical measures (Pollard, 1984).

Theorem 5.19 (Random subsample lemma). *Let $\zeta = (\zeta_1, \cdots, \zeta_m)$ be a sequence of m independent Rademacher variables, i.e. $\zeta \sim \mathrm{Unif}(\{-1,1\}^m)$. Then, for $\epsilon > 0$,*

$$\begin{aligned}&\mathbb{P}_{S\oplus P\sim D^{2m}}\left\{\sup_{w\in\mathcal{W}}[r_P(w) - r_S(w)] > \epsilon\right\}\\ <\;& 2\,\mathbb{P}_{S\sim D^m,\zeta\sim\mathrm{Unif}(\{-1,1\}^m)}\left\{\sup_{w\in\mathcal{W}}\left[\frac{1}{m}\sum_{i=1}^m \zeta_i w(x_i)\right] > \frac{\epsilon}{2}\right\}\\ =\;& 2\,\mathbb{E}_{S\sim D^m}\,\mathbb{P}_{\zeta\sim\mathrm{Unif}(\{-1,1\}^m)}\left\{\sup_{w\in\mathcal{W}}\left[\frac{1}{m}\sum_{i=1}^m \zeta_i w(x_i)\right] > \frac{\epsilon}{2}|S\right\} \quad , \qquad (5.34)\end{aligned}$$

where x_i denotes the i-th component of S.

Proof. The proof again makes use of the swapping subgroup $S^\star_{2m}$ of the symmetric group S_{2m}. The left hand probability is

$$\mathbb{P}_{S \oplus P \sim D^{2m}} \left\{ \sup_{w \in \mathcal{W}} [r_P(w) - r_S(w)] > \epsilon \right\} \ .$$

We can rewrite this as

$$\mathbb{E}_{Q \sim D^{2m}} \, \mathbb{P}_{\tau \sim \mathrm{Unif}\, S^\star_{2m}} \left\{ \sup_{w \in \mathcal{W}} [r_{P_{\tau(Q)}}(w) - r_{S_{\tau(Q)}}(w)] > \epsilon | Q \right\} \ ,$$

where $S_{\tau(Q)}$ and $P_{\tau(Q)}$ denote the first and second half of $\tau(Q)$ respectively. Writing s_i for the i-th component of $S_{\tau(Q)}$ and analogously for p_i and $P_{\tau(Q)}$, we obtain

$$\mathbb{E}_{Q \sim D^{2m}} \, \mathbb{P}_{\tau \sim \mathrm{Unif}\, S^\star_{2m}} \left\{ \sup_{w \in \mathcal{W}} \left[\frac{1}{m} \sum_{i=1}^{m} [w(p_i) - w(s_i)] \right] > \epsilon | Q \right\} \ .$$

As in the proof of Theorem 5.12, we can replace τ by ζ. This yields

$$\begin{aligned} & \mathbb{E}_{Q \sim D^{2m}} \, \mathbb{P}_{\zeta \sim \mathrm{Unif}(\{-1,1\}^m)} \left\{ \sup_{w \in \mathcal{W}} \left[\frac{1}{m} \sum_{i=1}^{m} \zeta_i [w(q_{m+i}) - w(q_i)] \right] > \epsilon | Q \right\} \\ \leq \ & \mathbb{E}_{Q \sim D^{2m}} \, \mathbb{P}_{\zeta \sim \mathrm{Unif}(\{-1,1\}^m)} \left\{ \sup_{w \in \mathcal{W}} \left[\frac{1}{m} \sum_{i=1}^{m} \zeta_i w(q_{m+i}) \right] > \frac{\epsilon}{2} | Q \right\} \\ & + \mathbb{E}_{Q \sim D^{2m}} \, \mathbb{P}_{\zeta \sim \mathrm{Unif}(\{-1,1\}^m)} \left\{ \sup_{w \in \mathcal{W}} \left[\frac{1}{m} \sum_{i=1}^{m} (-\zeta_i w(q_i)) \right] < \frac{-\epsilon}{2} | Q \right\} \ . \end{aligned}$$

Since the sample Q is i.i.d., and the ζ_i are symmetric, this equals

$$\begin{aligned} & 2 \, \mathbb{E}_{S \sim D^m} \, \mathbb{P}_{\zeta \sim \mathrm{Unif}(\{-1,1\}^m)} \left\{ \sup_{w \in \mathcal{W}} \left[\frac{1}{m} \sum_{i=1}^{m} \zeta_i w(s_i) \right] > \frac{\epsilon}{2} | S \right\} \\ = \ & 2 \, \mathbb{P}_{S \sim D^m, \zeta \sim \mathrm{Unif}(\{-1,1\}^m)} \left\{ \sup_{w \in \mathcal{W}} \left[\frac{1}{m} \sum_{i=1}^{m} \zeta_i w(x_i) \right] > \frac{\epsilon}{2} \right\} \ . \end{aligned}$$

□

Clearly, the same result holds when $r_P(w) - r_S(w)$ is replaced by $r_S(w) - r_P(w)$. Furthermore, a scrutiny of the proof shows that it also holds if the suprema are taken over absolute values. i.e.

$$\begin{aligned} & \mathbb{P}_{S \oplus P \sim D^{2m}} \left\{ \sup_{w \in \mathcal{W}} |r_P(w) - r_S(w)| > \epsilon \right\} \\ & \qquad < 2 \, \mathbb{E}_{S \sim D^m} \, \mathbb{P}_{\zeta \sim \mathrm{Unif}(\{-1,1\}^m)} \left\{ \sup_{w \in \mathcal{W}} \left| \frac{1}{m} \sum_{i=1}^{m} \zeta_i w(s_i) \right| > \frac{\epsilon}{2} | S \right\} \ . \end{aligned}$$

Combining Theorem 5.19 with the symmetrization lemma of (5.7) (for $m = u$) we obtain, for any $0 < \beta \leq 1$, and $\epsilon > \alpha(u, \beta)$,

$$\begin{aligned}&\mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}}[r_D(w) - r_S(w)] > \epsilon\right\}\\ &\leq 2\beta^{-1}\,\mathbb{P}_{S\sim D^m,\zeta\sim\mathrm{Unif}(\{-1,1\}^m)}\left\{\sup_{w\in\mathcal{W}}\left[\frac{1}{m}\sum_{i=1}^{m}\zeta_i w(x_i)\right] > \frac{\epsilon - \alpha(u,\beta)}{2}\right\} \quad ,\end{aligned}$$

where α is a function of β so that the requirements of the symmetrization lemma are required. The same result holds with absolute values inside the suprema — see, for example, Giné and Zinn (1984, Lemma 2.7).

As in Section 5.5.4 our approach will be to work conditionally on a given sample. In this case, we investigate, for a given S and any decision rule w (not necessarily in $\mathcal{W}$), the probability on the right of (5.34). We can bound this using Hoeffding's tail inequality in a similar way to the proof of Theorem 5.12: since S is an i.i.d. sample, we have that $V_i(w) = \zeta_i w(x_i)$ for $i \in [1 : m]$ are independent r.v.'s, each with mean zero.

Applying Hoeffding's tail inequality, we obtain

$$\mathbb{P}_{\zeta\sim\mathrm{Unif}(\{-1,1\}^m)}\left\{\frac{1}{m}\sum_{i=1}^{m}V_i(w) > \epsilon|S\right\} \leq \exp\left(\frac{-2m^2\epsilon^2}{\sum_{i=1}^{m}[2w(x_i)]^2}\right) \quad .$$

Finally we note that $w(x_i) \leq 1$. We call the resulting inequality,

$$\mathbb{P}_{\zeta\sim\mathrm{Unif}(\{-1,1\}^m)}\left\{\frac{1}{m}\sum_{i=1}^{m}\zeta_i w(x_i) > \epsilon|S\right\} \leq \exp\left(\frac{-m\epsilon^2}{2}\right) \quad , \tag{5.35}$$

the *random subsample bound.*

The expression $\sup_{w\in\mathcal{W}}[\frac{1}{m}\sum_{i=1}^{m}\zeta_i w(x_i)]$ is known as the *Rademacher penalty* of $\mathcal{W}$ for the sample S, and we shall denote it by $\mathcal{R}_S(\mathcal{W})$.

In what follows, we show how to employ an ϵ-cover to bound the Rademacher penalty. Section 5.7 presents a more sophisticated method for bounding the Rademacher penalty, which uses a sequence of covers of various scales. In Section 7.2, we discuss methods for bounding the Rademacher penalty more directly by employing concentration inequalities.

5.5.7 Applying the cover with the random subsample bound

This section is analogous to Section 5.5.5. Here we apply the uniform Occam's razor method to the random subsample bound over a minimal γ-cover of $\mathcal{W}$ w.r.t. $d_{p,S}$, $\mathcal{W}^\star(S)$. Unlike in Section 5.5.5 we are not restricted to proper covers of $\mathcal{W}$, since the random subsample lemma applies to any decision rule w.

It follows that

$$\mathbb{P}_{\zeta\sim\mathrm{Unif}(\{-1,1\}^m)}\left\{\sup_{w\in\mathcal{W}^\star(S)}\left[\frac{1}{m}\sum_{i=1}^m \zeta_i w(x_i)\right] > \epsilon|S\right\}$$
$$\leq \bar{\mathcal{N}}_{p,S}(\gamma,\mathcal{W})\exp\left(\frac{-m\epsilon^2}{2}\right) \ .$$

In order to extend this result to $\mathcal{W}$, consider any $\zeta \in \{-1,1\}^{2m}$ and $w \in \mathcal{W}$. Suppose $w^\star \in \mathcal{W}^\star(S)$ γ-approximates w w.r.t. $d_{p,Q}$. Then

$$\frac{1}{m}\sum_{i=1}^m \zeta_i w(x_i) = \frac{1}{m}\sum_{i=1}^m \zeta_i w^\star(x_i) + \frac{1}{m}\sum_{i=1}^m \zeta_i[w(x_i) - w^\star(x_i)] \ .$$

The first term is subject to the bound above. We use the fact that $w^\star$ γ-approximates w to bound the other term. We have

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^m \zeta_i[w(x_i) - w^\star(x_i)] &\leq \frac{1}{m}\sum_{i=1}^m |w(x_i) - w^\star(x_i)| \\ &= d_{1,S}(w,w^\star) \\ &\leq d_{p,S}(w,w^\star) \\ &< \gamma \ .\end{aligned}$$

It follows that

$$\mathbb{P}_{\zeta\in\mathrm{Unif}(\{-1,1\}^{2m})}\left\{\sup_{w\in\mathcal{W}}\left[\frac{1}{m}\sum_{i=1}^m \zeta_i w(x_i)\right] > \epsilon+\gamma|S\right\}$$
$$< \bar{\mathcal{N}}_{p,Q}(\gamma,\mathcal{W})\exp\left(\frac{-m\epsilon^2}{2}\right) \ ,$$

with the right hand side a minimum for $p = 1$.

The next step is to integrate over the possible samples S, which yields

$$\mathbb{P}_{S\sim D^m,\zeta\sim\mathrm{Unif}(\{-1,1\})^m}\left\{\sup_{w\in\mathcal{W}}\left[\frac{1}{m}\sum_{i=1}^{m}\zeta_i w(x_i)\right] > \epsilon\right\} < \mathbb{E}_{Q\sim D^m}\,\bar{\mathcal{N}}_{p,Q}(\gamma,\mathcal{W})\exp\left(\frac{-m(\epsilon-\gamma)^2}{2}\right) .$$

Applying the random subsample lemma (Theorem 5.19) and the general symmetrization lemma for regular deviation in (5.7), we obtain for any $0 < \beta \leq 1$,

$$\begin{aligned}
&\mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}}[r_D(w)-r_S(w)] > \epsilon\right\}\\
\leq\ &\beta^{-1}\,\mathbb{P}_{S\oplus P\sim D^{2m}}\left\{\sup_{w\in\mathcal{W}}[r_P(w)-r_S(w)] > \epsilon-\alpha(m,\beta)\right\}\\
\leq\ &\frac{2}{\beta}\,\mathbb{P}_{S\sim D^m,\zeta\sim\mathrm{Unif}(\{-1,1\})^m}\left\{\sup_{w\in\mathcal{W}}\left[\frac{1}{m}\sum_{i=1}^{m}\zeta_i w(x_i)\right] > \frac{\epsilon-\alpha(m,\beta)}{2}\right\} &(5.36)\\
\leq\ &\frac{2}{\beta}\,\mathbb{E}_{Q\sim D^m}\,\bar{\mathcal{N}}_{p,Q}(\gamma,\mathcal{W})\exp\left(\frac{-m\left(\left[\frac{\epsilon-\alpha(m,\beta)}{2}\right]-\gamma\right)^2}{2}\right) . &(5.37)
\end{aligned}$$

Applying Theorem 5.3, we can replace β by $\frac{1}{2}$ and $\alpha(m,\beta)$ by $\frac{\epsilon}{2}$, so that[103]

$$\mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}}[r_D(w)-r_S(w)] > \epsilon\right\} \leq 4\,\mathbb{E}_{Q\sim D^m}\,\bar{\mathcal{N}}_{p,Q}(\gamma,\mathcal{W})\exp\left(\frac{-m\left(\frac{\epsilon}{4}-\gamma\right)^2}{2}\right) .$$

It is interesting to compare this result with the bound in (5.26). The bound in (5.26) has an extra factor of m in the coefficient. On the other hand, the current bound has an additional factor of approximately $\frac{1}{8}$ in the exponent. As a result, even if this bound begins narrower than the bound in (5.26), that bound is asymptotically tighter.

Choosing $\gamma = \frac{\epsilon}{8}$ yields a result very similar to Devroye et al. (1996, Theorem 29.1), which is in turn implicit in Pollard (1984, Theorem 24).

[103] Usually, we have the sample size restriction $m > 2\epsilon^{-2}$, but the bound on the right is trivial for $m \leq 2\epsilon^{-2}$ in this case, so it is not necessary. This is often the explanation for the disappearance of sample size restrictions when results are derived in the literature.

Setting the right hand side to δ, we obtain

$$\mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}}[r_D(w)-r_S(w)]>4\left[\gamma+\sqrt{\frac{2[\ln 4\,\mathbb{E}_{Q\sim D^m}\mathcal{N}_{p,Q}(\gamma,\mathcal{W})-\ln\delta]}{m}}\right]\right\}<\delta\ . \tag{5.38}$$

5.5.8 Thresholded class covering number bounds

Vapnik and Chervonenkis's original work in the 1960's only derived bounds for zero-one loss functions (Vapnik, 1998). Their approach to extending their results to general loss functions employed a so-called *thresholded decision class* $\mathcal{W}_t$ related to the the decision class $\mathcal{W}$. This associated class consisted of decision rules with range in $\{0,1\}$ so that their results for error bounds could be applied to $\mathcal{W}_t$.

The following lemma shows that any bound on the regular deviation of errors applies to the regular deviation of risk on the original decision class:

Theorem 5.20 (Thresholded class lemma for regular deviation).

$$\mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}}[r_D(w)-r_S(w)]>\epsilon\right\}\le\mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}_t}[e_D(w)-e_S(w)]>\epsilon\right\}\ ,$$

where the thresholded decision class $\mathcal{W}_t$ is defined by

$$\mathcal{W}_t=\{w_s : w\in\mathcal{W}, s\in[0,1]\}\ ,$$

where $w_s=I(w(\cdot)>s)$.

In other words, for a given $s\in[0,1]$ and a decision rule w, we construct a new decision rule which is 1 if $w(x)>s$, and 0 otherwise. The collection of all such binary decision rules forms the thresholded decision class $\mathcal{W}_t$. Any bound on error deviation over $\mathcal{W}_t$ thus yields a bound on risk deviation over $\mathcal{W}$.

The proof of this lemma is based on expanding the integrals in the definition of the risk as a limit, and investigating the limit directly. For proofs, the reader is referred to Vapnik (1998, Section 5.2.2) or Devroye et al. (1996, Lemma 29.1). As an example of the application of this result, we have from (5.17) that

$$\mathbb{P}_{S\sim D^m}\left\{\begin{array}{r} \sup_{w\in\mathcal{W}}\left[r_D(w)-r_S(w)\right] > 2\gamma \\ +\frac{1+\sqrt{(m+1)[\ln 2\,\mathbb{E}_{Q\sim D^{m+u}}\,\mathcal{N}_{1,Q}(\gamma,\mathcal{W}_t)-\ln\delta]+1}}{m} \end{array}\right\} < \delta \ . \qquad (5.39)$$

Let us now consider the case with relative deviation. The denominator presents an obstacle to obtaining as direct a result as the thresholded class lemma for regular deviation.

To proceed, we note that for a given w_c, we have

$$e_D(w_s) = \mathbb{P}_{Z\sim D}\{w(Z) > s\} \ .$$

The core of Vapnik (1998, Theorem 5.2)[104] essentially then consists of proving the following result:

Theorem 5.21 (Thresholded class lemma for relative deviation). *For $p \geq 1$,*

$$\mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}}\frac{r_D(w)-r_S(w)}{D_p(w)} > \epsilon\right\} \leq \mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}_t}\frac{e_D(w)-e_S(w)}{\sqrt[p]{e_D(w)}} > \epsilon\right\} \ ,$$

where

$$D_p(w) = \int_0^1 \sqrt[p]{\mathbb{P}_{Z\sim D}\{w(Z) > s\}}\, ds \ .$$

Applying this result with (5.19), we have

$$\mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}}\frac{r_D(w)-r_S(w)}{D_2(w)} > \epsilon\right\} < 4\,\mathbb{E}_{Q\sim D^{2m}}\left|Q_{\mathcal{W}_t}\right|\exp\left(\frac{-m\epsilon^2}{4}\right) \ . \qquad (5.40)$$

Finally, we present a result relating $D_p(w)$ and $\sqrt[p]{r_D(w)}$ which allows us to obtain relative deviation bounds.

Hölder's inequality states that if $\frac{1}{v_1}+\frac{1}{v_2}=1$ for $v_1, v_2 > 0$, then

$$\int_{c_1}^{c_2}|\phi_1(s)\phi_2(s)|ds \leq \left(\int_{c_1}^{c_2}|\phi(s)|^{v_1}\,dc\right)^{\frac{1}{v_1}}\left(\int_{c_1}^{c_2}|\phi_2(s)|^{v_2}\,dc\right)^{\frac{1}{v_2}} \ ,$$

[104] Note that this result is misstated in the reference. $H_{\text{ann}}^{\Lambda,\mathcal{B}}(l)$ should be replaced by $H_{\text{ann}}^{\Lambda,\mathcal{B}}(2l)$.

i.e. that $\|\phi_1\phi_2\|_1 \leq \|\phi_1\|_{v_1}\|\phi_2\|_{v_2}$, where $\|\cdot\|_p$ denotes the norm in $L^p[c_1, c_2]$. Applying this inequality with $c_1 = 0$, $c_2 = 1$,

$$\begin{aligned}\phi_1(s) &= \sqrt[p]{\mathbb{P}_{Z\sim D}\{w(Z) > s\}} \\ &= e_D(w_s)\end{aligned} \quad ,$$

$\phi_2(s) = 1$, and $v_1 = p$ yields, for $p > 1$,

$$D_p(h) \leq \left(\int_0^1 \mathbb{P}_{Z\sim D}\{w(Z) > s\}\, ds\right)^{\frac{1}{p}} 1^{1-\frac{1}{p}} \ .$$

Since

$$\begin{aligned}\int_0^1 \mathbb{P}_{Z\sim D}\{w(Z) > s\}\, ds &= \mathbb{E}_{Z\sim D}\, w(Z) \\ &= r_D(w) \ ,\end{aligned}$$

we have

$$D_p(h) \leq \sqrt[p]{r_D(w)} \ .$$

Combining this with (5.40) yields a bound on relative deviation of risk:

$$\mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}} \frac{r_D(w) - r_S(w)}{\sqrt{r_D(w)}} > \epsilon\right\} < 4\,\mathbb{E}_{Q\sim D^{2m}}\, |Q_{\mathcal{W}_t}| \exp\left(\frac{-m\epsilon^2}{4}\right) \ . \tag{5.41}$$

Note that this is the only result we have presented so far directly bounding relative deviation of risk.

5.6 Bounds from dominating loss functions

This section will focus on upper deviations. In many cases, similar results for lower deviations can be obtained.

Until now, we have obtained probabilistic bounds on a measure of deviation $\psi(r_D(w), r_S(w))$. However, we can generally study $\psi(r_D(w), v)$ for any v. If we can find a probabilistic bound on this expression, we can invert the probability statement to obtain an interval estimator of $r_D(w)$. Typically v will depend on the decision rule w, the sample S (as our source of information about D), and the loss function L. Thus $r_S(w)$ seems like a good choice. However, in some cases, we may be able to obtain better bounds using alternative choices for v.

A common choice of v is $r_S(w, L')$, where L' is a loss function which dominates L — a so-called proxy loss. In addition, convexity of the proxy loss is often desirable.

In the rest of this section, we will be considering such an approach to obtain so-called *margin bounds*. We consider the problem slightly more generally than the outline above, by considering loss functions using the output of the hypothesis class, prior to application of the strategy. Since the modified setting used so far in this chapter effectively discards this information, we will not be able to employ it here. For this section, then, we shall work in the regular setting.

Consider the unmodified problem specified by $\{\mathcal{X}, \mathcal{Y}, \mathcal{S}, \mathcal{A}, \mathcal{H}, L\}$, and a specific strategy g. We are interested in improving our risk estimators of $r_D(w)$ by employing information about the values of the hypothesis on S.

Suppose the hypotheses in $\mathcal{H}$ map into R. Then we may consider a related problem, specified by $\{\mathcal{X}, \mathcal{Y}, \mathcal{S}, R, \mathcal{H}, L'\}$, with the strategy restricted to id_R, where $L' : R \times \mathcal{Y} \to \mathrm{I\!R}^+$ is a loss function. We say that L' g-dominates L if the following holds: for any $(x, y) \in \mathcal{Z}$, and any $h \in \mathcal{H}$

$$L'(h(x), y) \geq L(g_h(x), y) \ .$$

If L' g-dominates L, it follows that

$$r_Q(g_h, L) \leq r_Q(h, L')$$

for any distribution Q on $\mathcal{Z}$. Thus, one can upper bound the true risk of a decision rule w by upper bounding the true risk of the underlying hypothesis with respect to a g-dominating loss of L on R.

Since the strategy in the related problem is id_R, the hypothesis class $\mathcal{H}$ effectively comprises the decision class for the related problem. Note that the bounds obtained in the earlier sections of this chapter were obtained in terms of the covering numbers of the loss class $\mathcal{F}_{\mathcal{W}}$.

Suppose for a given L', we can find a second loss function $L'' \leq L'$ on $R \times \mathcal{Y}$ which also g-dominates L, which further satisfies the following: for any distribution Q on $\mathcal{Z}$, if $d_{\infty,Q}(h_1, h_2) < \gamma$, we have that

$$L'(h_1(x), y) \geq L''(h_2(x), y)$$

and

$$L(g_{h_1}(x), y) \leq L''(h_2(x), y)$$

almost surely for $(x, y) \sim Q$. We shall call such an L'' an (L, L')-*intermediary with constant* γ.

Now, consider a measure of deviation $\psi(x, y)$, and consider a typical symmetrization lemma upper bounding

$$\mathbb{P}_{S\sim D^m}\left\{\sup_{h\in\mathcal{H}} \psi\left(r_D(g_h, L), r_S(g_h, L)\right) > \epsilon\right\}$$

by $\phi_1(v)$, with

$$v \geq \mathbb{P}_{S\oplus P\sim D^{m+u}}\left\{\sup_{h\in\mathcal{H}} \psi^\star(r_P(g_h, L), r_S(g_h, L)) > \phi_2(\epsilon)\right\} \quad , \tag{5.42}$$

for some functions ϕ_1 and ϕ_2, and a (potentially) modified measure of deviation $\psi^\star$.

In many cases, a similar argument can be used to show that we can upper bound

$$\mathbb{P}_{S\sim D^m}\left\{\sup_{h\in\mathcal{H}} \psi(r_D(g_h, L), r_S(h, L')) > \epsilon\right\}$$

by some $\phi_1'(v')$, with

$$v' \geq \mathbb{P}_{S\oplus P\sim D^{m+u}}\left\{\sup_{h\in\mathcal{H}} \psi'(r_P(g_h, L), r_S(h, L')) > \phi_2'(\epsilon)\right\} \quad , \tag{5.43}$$

for suitable ϕ_1',ϕ_2' and $\psi_\prime$.

It will be necessary to assume that $\psi'(v_1, v_2)$ is increasing in v_1 and decreasing in v_2. This is the case for the upper regular deviation and the upper relative deviation, and seems a fairly natural condition for an upper measure of deviation.

The key to improving bounds with this approach is to use L'' to obtain good bounds on v'. Let L'' be an (L, L')-intermediary with constant γ, and let $\mathcal{H}^\star$ be a γ-cover of $\mathcal{H}$ w.r.t. $d_{\infty,S\oplus P}$. For any $h \in \mathcal{H}$, let $h^\star \in \mathcal{H}^\star$ satisfy $d_{\infty,S\oplus P}(h, h^\star) < \gamma$. It follows that

$$r_P(h, L') \geq r_P(h^\star, L'')$$

and

$$r_S(g_h, L) \leq r_S(h^\star, L'') \ \ .$$

By our assumptions on ψ', it follows that

$$\psi'\left(r_P(g_h, L), r_S(h, L')\right) \leq \psi'\left(r_P(h^\star, L''), r_S(h^\star, L'')\right) \ \ .$$

Putting this together, we have

$$\begin{aligned}
&\mathbb{P}_{S\sim D^m}\left\{\sup_{h\in\mathcal{H}} \psi(r_D(g_h, L), r_S(h, L')) > \epsilon\right\} \\
\leq\ & \phi'_1\left(\mathbb{P}_{S\oplus P\sim D^{m+u}}\left\{\sup_{h\in\mathcal{H}} \psi'(r_P(g_h, L), r_S(h, L')) > \phi'_2(\epsilon)\right\}\right) \\
\leq\ & \phi'_1\left(\mathbb{P}_{S\oplus P\sim D^{m+u}}\left\{\sup_{h\in\mathcal{H}^\star} \psi'(r_P(h, L''), r_S(h, L'')) > \phi'_2(\epsilon)\right\}\right) \ \ . \qquad (5.44)
\end{aligned}$$

All that remains is bounding the supremum of the deviation over the cover, a technique which we have performed for a number of deviations already. In order to do this, we need a dual sample bound for ψ'. Conditioning on $S\oplus P$, employing minimal covers, applying the dual sample bound, adjusting the result to allow for the approximation of the whole class by a cover, and taking the expectation over the dual sample typically shows that (5.44) does not exceed

$$\phi'_1\left(\mathbb{E}_{Q\sim D^{m+u}} \mathcal{N}_{\infty,Q}(\gamma, \mathcal{H}) \phi_3(m, u, \phi'_2(\epsilon), \gamma)\right) \ \ ,$$

where ϕ_3 is a function based on the dual sample bound for risk for the measure of deviation ψ', and the corrections necessary for approximating by a cover.

Note that the dual sample bound employed may restrict the choice of u. In addition, if all the loss functions concerned are zero-one loss functions, one may employ a dual sample for error.

It seems that lower bounds will follow similarly, but we have not pursued this direction further.

5.6.1 Margin bounds for thresholding classifiers

We now show that many of the margin bounds developed for thresholding classifiers in the 1990's fall into the framework sketched above. The pioneering work in this direction was predominantly due to Peter Bartlett, John Shawe-Taylor and their collaborators — for the major developments in the literature, see Anthony and Bartlett (1994), Shawe-Taylor et al. (1996), Schapire et al. (1997), Shawe-Taylor et al. (1998), Bartlett (1998), Freund and Schapire (1999), Shawe-Taylor and Cristianini (1998a), and Shawe-Taylor and Cristianini (1998b).

We consider the thresholding classifiers introduced in Example 2.6. Thus we assume $\mathcal{A} = \mathcal{Y} = \{0, 1\}$, and that the hypotheses output real values. Furthermore, we are interested in error estimation for the strategy $g(v) = I(v \geq s)$, so we have

$$w(x) = g_h(x) = I(h(x) \geq s) \ .$$

Margin bounds are based on the intuition that when the real value calculated by the underlying hypothesis (i.e. $h(x)$) differs substantially from the threshold s , classification can be performed more confidently than if $h(x)$ is close to s. This idea is closely related to the Glick smoothed estimate described in Section 3.1.6.

The question of the validity of this intuition was, to our knowledge, first raised by Vapnik and Lerner in 1963. Vapnik later gave results supporting the intuition, and used his ideas with his development of canonical hyperplane classifiers for transductive learning (Vapnik, 1982). The ideas there later led to the development of support vector machines in Boser et al. (1992), Cortes and Vapnik (1995), Guyon et al. (1993).

Consider the zero-one loss function

$$L(w(x), y) = I(w(x) \neq y) \ .$$

This is the natural loss function for classification. We shall now propose the zero-one *margin loss* functions, defined on the related problem. Thus these loss functions map $\mathbb{R} \times \{0,1\} \to \{0,1\}$. Let

$$L_\gamma(h(x), y) = I([h(x) - s] \operatorname{sgn}(y) < \gamma)$$

for $\gamma \geq 0$. It is straightforward to verify that

$$L(g_h(\cdot), \cdot) = L_0(h(\cdot), \cdot) \ .$$

Moreover, it is clear that for $\gamma_1 \geq \gamma_2$ we have $L_{\gamma_1} \geq L_{\gamma_2}$. Finally, for any $\gamma \geq 0$, these observations mean that L_γ g-dominates L. The quantity $[h(x) - s] \operatorname{sgn}(y)$ is called the *margin* of h attained by (x, y)[105], so that the margin loss with parameter γ penalizes points not attaining a margin of at least γ. Also note that having a positive margin on a point implies correct prediction, while an incorrect prediction results from a negative margin.

The critical fact allowing us to obtain margin bounds is the following statement: $L_{\frac{\gamma}{2}}$ is an (L, L_u)-intermediary with constant $\frac{\gamma}{2}$.

Indeed, consider any distribution Q on $\mathcal{Z}$, and suppose $d_{\infty,Q}(h_1, h_2) < \frac{\gamma}{2}$, and let $(x, y) \in \operatorname{supp} Q$. Since $|h_1(x) - h_2(x)| \leq \frac{\gamma}{2}$, it follows that if

$$[h_1(x) - s] \operatorname{sgn}(y) \geq \gamma$$

then

$$[h_2(x) - s] \operatorname{sgn}(y) \geq \frac{\gamma}{2}$$

so that $L_{\frac{\gamma}{2}}(h_2(x), y)$ is always zero when $L_\gamma(h_1(x), y)$ is zero. As a result we have

$$L_\gamma(h_1(x), y) \geq L_{\frac{\gamma}{2}}(h_2(x), y) \ .$$

[105] Note that the margin has an implicit dependence on the strategy.

A similar argument shows that

$$L(g_{h_1}(x), y) \leq L_{\frac{\gamma}{2}}(h_2(x), y) \ \ .$$

Next we turn to the issue of obtaining an appropriate modified symmetrization lemma. In fact, a study of the original proof of the Vapnik symmetrization lemma for upper relative deviations shows that it still holds verbatim if we make the following changes to Theorem 5.7:

- replace $e_D(w)$ by $e_D(g_h, L)$;
- replace $e_P(w)$ by $e_P(g_h, L)$; and
- replace $e_S(w)$ by $e_S(h, L')$.

We will apply this modified result with $L' = L_\gamma$.

Note that the first two modifications are not really changes to the theorem, just modifying the notation to the new setting. The third replacement holds because of an optimization argument employed in the original proof. As a bonus, the derivative computations for the optimization show that

$$\psi'(x, y) = \frac{x - y}{\sqrt[p]{\frac{1}{2}\left[x + y + \frac{1}{\gamma}\right]}}$$

is increasing in x and decreasing in y. The same holds if we discard the $\frac{1}{\gamma}$ term.

The last component we need is a double sample bound on upper relative deviation of error. For this, we can apply Theorem 5.12 with respect to the loss function $L_{\frac{\gamma}{2}}$. Note that we are now restricted to $m = u$.

Thus we have

$$\phi_1'(v) = 4v \ \ ,$$

$$\phi_2'(\epsilon) = \epsilon \ \ ,$$

and

$$\phi_3(m, m, \epsilon, \gamma) = \exp\left(\frac{-\epsilon^2 m}{4}\right) \ \ .$$

Combining these ingredients we obtain the following result corresponding to Bartlett (1998, Theorem 6):[106]

$$\mathbb{P}_{S\sim D^m}\left\{\sup_{h\in\mathcal{H}}\frac{e_D(g_h,L)-e_S(h,L_\gamma)}{\sqrt{e_D(g_h,L)}}>\epsilon\right\}$$
$$<4\,\mathbb{E}_{Q\sim D^{2m}}\mathcal{N}_{\infty,Q}\left(\frac{\gamma}{2},\mathcal{H}\right)\exp\left(\frac{-\epsilon^2 m}{4}\right)\quad . \qquad (5.45)$$

Note that this result can also be used to obtain bounds for the realizable and realistic case employing empirical margin losses.

A similar approach can be followed for regular deviation, combining a modification of the regular symmetrization lemma in (5.7)[107] with the Vapnik double sample bound of Theorem 5.8. This yields the following result:

$$\mathbb{P}_{S\sim D^m}\left\{\sup_{h\in\mathcal{H}}\left[e_D(g_h,L)-e_S(h,L_\gamma)\right]>\epsilon\right\}$$
$$\leq\beta^{-1}\,\mathbb{E}_{Q\sim D^{2m}}\mathcal{N}_{\infty,Q}\left(\frac{\gamma}{2},\mathcal{H}\right)\exp\left(-\frac{(\epsilon-\alpha(m,\beta))^2m^2-1}{m+1}\right)\quad , \qquad (5.46)$$

where $\alpha(m,\beta),\beta$ satisfy the requirements of the symmetrization lemma, and $\epsilon>\alpha(m,\beta)$. We note that setting $\beta=\frac{1}{2}$ and $\alpha=\frac{\epsilon}{2}$ for large enough m yields a weaker result than Bartlett (1998, Lemma 4). Bartlett's proof employs a direct argument similar to the random subsample lemma, but avoids the reduction in resolution usually necessary. Our result could also be applied to unequal sample sizes by employing the dual sample bound for regular deviation of risk (Theorem 5.10) instead. This yields a bound with worse growth of covering numbers, but which would outperform Bartlett's result asymptotically. For reference, we state Bartlett's result in a relevant form below.

Theorem 5.22 (Lemma 4 of Bartlett, 1998). *Select* $\gamma>0$*, and* $0<$

[106] Although this result implicitly has the sample size restriction $m>\epsilon^{-2}$ derived from employing (5.19), the restriction has been dropped since the bound is trivial for smaller values of m.

[107] The proof of the modified symmetrization lemma follows directly by applying Theorem 5.2 to appropriate processes, so we have $\phi_1'=\phi_1$, and $\phi_2'=\phi_2$.

$\delta < \frac{1}{2}$. *Then,*

$$\mathbb{P}_{S \sim D^m} \left\{ \sup_{h \in \mathcal{H}} [e_D(g_h, L) - e_S(h, L_\gamma)] > \epsilon \right\} \\ \leq 2\, \mathbb{E}_{Q \sim D^{2m}}\, \mathcal{N}_{\infty,Q}\left(\frac{\gamma}{2}, \mathcal{H}\right) \exp\left(-\frac{\epsilon^2 m}{2}\right) \ .$$

The last bound we discuss is a bound for the realizable case. We could obtain a realizable-case bound from the relative deviation result, but we can do better by directly employing the realizable symmetrization lemma and the realizable double sample bound in Theorems 5.15 and 5.14.

Once again, the symmetrization lemma can be shown to hold in the same form with the replacement of the training risk by the training margin risk. This leads to the following realizable margin bound, which can be compared to the bounds in (5.24) and (5.25) for the same choices of α and β:

$$\mathbb{P}_{S \sim D^m} \{\exists h \in \mathcal{H} : (e_S(h, L_\gamma) = 0) \wedge (e_D(g_h, L) > \epsilon)\} \\ \leq \beta^{-1}\, \mathbb{E}_{Q \sim D^{m+u}}\, \mathcal{N}_{\infty,Q}\left(\frac{\gamma}{2}, \mathcal{H}\right) \left(\frac{u}{m+u}\right)^{\lceil u \alpha_{C-R}(u,\beta,\epsilon)\epsilon \rceil} \ . \quad (5.47)$$

5.6.2 ε-insensitive loss

Consider a problem employing the ε-insensitive indicator loss function of Example 2.13,

$$L_\varepsilon(y_1, y_2) = I(|y_1 - y_2| > \varepsilon) \ .$$

By considering a related problem associated with this one, we can obtain bounds from the margin bounds for thresholded classifiers.[108] Suppose the original problem is specified by $\{\mathcal{X}, \mathbb{R}, \mathcal{S}, \mathbb{R}, \mathcal{H}, L\}$. We specify the new problem as $\{\mathcal{X} \times \mathbb{R}, \{0,1\}, \mathcal{S}', \{0,1\}, \mathcal{H}', L^\star\}$, where the first four components are identical to those of the modified learning problem, and $L^\star(y_1, y_2) = I(y_1 \neq y_2)$. The hypotheses in the new problem are related to the original hypotheses by

$$h'(x') = h'(x, y) = |h(x) - y| \ .$$

[108] This approach is very similar to that performed in Shawe-Taylor and Cristianini (1998b).

In this new problem setting, we specify the strategy g' by

$$g'(v) = I(v > \varepsilon) \ .$$

Furthermore, we restrict $\mathcal{S}'$ to the products of distributions in $\mathcal{S}$ with the distribution concentrated entirely on $\{0\}$. It follows that for any distribution $D' \in \mathcal{S}'$, for $(x', y') \sim D'$, we have $y' = 0$, so that

$$L^\star(g'(h'(x')) \neq y') = g'(h'(x')) \ .$$

Suppose (x, y) erred on h in the original problem. Then $|h(x) - y| > \varepsilon$, so that

$$g'(h'(x')) = I(|h(x) - y| > \varepsilon) \ .$$

A similar argument when (x, y) does not err on h shows that the losses for both problems are identical, and thus the risk of corresponding decision rules are the same.

With this setup, we can apply the margin bounds described above. Note that the loss function $L^\star_\gamma$ in this scenario can be expanded as follows:

$$\begin{aligned} L^\star_\gamma(h'(x'), 0) &= I(-[h'(x') - \varepsilon] < \gamma) \\ &= I(h'(x') < \varepsilon - \gamma) \\ &= I(|h(x) - y| > \varepsilon) \\ &= L_\gamma(h(x), y) \ . \end{aligned}$$

Thus, intuitively, a point (x, y) is associated with a γ-margin loss in this scenario if it falls outside a narrower insensitive tube, of width $2(\varepsilon - \gamma)$.

Applying the margin bound for upper relative deviation to this problem, we have that

$$\mathbb{P}_{S' \sim D'^m} \left\{ \sup_{h' \in \mathcal{H}'} \frac{r_{D'}({g'}_{h'}, L^\star) - r_{S'}(h', L^\star_\gamma)}{\sqrt{r_{D'}({g'}_{h'}, L^\star)}} > \epsilon \right\} \\ < 4\, \mathbb{E}_{Q \sim D'^{2m}}\, \mathcal{N}_{\infty, Q}\left(\frac{\gamma}{2}, \mathcal{H}'\right) \exp\left(\frac{-\epsilon^2 m}{4}\right) \ .$$

Noting that

$$r_{D'}({g'}_{h'}, L^\star) = r_D(g_h, L)$$

and that

$$r_{S'}(h', L^\star_\gamma) = r_S(h, L_\gamma) \ ,$$

we can reformulate this as

$$\mathbb{P}_{S\sim D^m}\left\{\sup_{h\in\mathcal{H}} \frac{r_D(g_h, L) - r_S(h, L_\gamma)}{\sqrt{r_D(g_h, L)}} > \epsilon\right\}$$
$$< 4\,\mathbb{E}_{Q\sim D^{2m}}\,\mathcal{N}_{\infty,Q}\left(\frac{\gamma}{2}, \mathcal{H}'\right)\exp\left(\frac{-\epsilon^2 m}{4}\right) \ .$$

It is important to note that the covering number on the right is for covers of $\mathcal{H}'$, not $\mathcal{H}$. However, in practice, this is not a major obstacle.

If we instead apply the bound in (5.47), choosing $\beta = \frac{1}{2}$ and $m = u$, we essentially recover the covering number result underlying one of the earliest (if not the earliest) margin bounds, Anthony and Bartlett (1994, Theorem 5) (except that the covering number is once again of $\mathcal{H}'$). It was perhaps not realized at the time, but the interpretation of the result as a margin bound for classification was presented in Shawe-Taylor et al. (1998).

Finally, note that the analysis in this section could easily be generalized by replacing the absolute value in the original loss function by a norm in some other space instead of $\mathbb{R}$, thus generalizing the concept of an ϵ-insensitive tube to other spaces (particularly, higher-dimensional Euclidean spaces may be useful).

5.6.3 Margin bounds for other classifiers

Earlier, we defined the margin of a point w.r.t. an hypothesis for thresholded classifiers. In this section, we continue our consideration of classifiers. Thus, we assume our strategy maps into $\{0, 1\}$. Furthermore, we assume the hypotheses map into a metric space $(\mathcal{E}, d)$.

We begin by generalizing the margin concept.

Definition 5.3 (Generalized margin). The margin of an hypothesis $h \in \mathcal{H}$ on a point $z = (x, y) \in \mathcal{X} \times \{0, 1\}$ is

$$\rho(z, h) = \operatorname{sgn} y \operatorname{sgn}\left(I(g_h(x) = 1)\right) d\left(h(x), \{\eta \in \mathcal{E} : g(\eta) \neq g_h(x)\}\right) \ .$$

To demystify this definition, the first two factors determine the sign of the margin: it is positive when the x is correctly classified by g_h, and negative otherwise. The final term measures the distance from $h(x)$ to the closest point in $\mathcal{E}$ that the strategy would classify differently to $h(x)$. In the case of thresholded classifiers, this is the distance from $h(x)$ to s, where the classification changes.

Effectively, since the strategy is a map from $\mathcal{E}$ to $\{0,1\}$, g partitions $\mathcal{E}$ into a portion it maps to zero, and a portion it maps to one. The size of the margin can then informally be seen as the distance of $h(x)$ from the boundary between these two portions, while the sign is determined by whether x was correctly classified by g_h or not. In the thresholding classifier, the portion of $\mathbb{R}$ mapping to zero, were the points less than s, while those mapping to one were the points at least s. Thus, the boundary between the portions consisted of s, showing that this margin is a generalization of that for thresholded classifiers.

We can also generalize the margin loss easily, by defining

$$L_\gamma(h(x),y) = I(\rho((x,y),h) < \gamma) \ .$$

Again we have

$$L(g_h(\cdot),\cdot) = L_0(h(\cdot),\cdot) \ ,$$

$L_{\gamma_1} \geq L_{\gamma_2}$ for $\gamma_1 \geq \gamma_2$, and that L_γ g-dominates L for any $\gamma \geq 0$. Furthermore, the same approach shows that $L_{\frac{\gamma}{2}}$ is an (L, L_γ)-intermediary with constant $\frac{\gamma}{2}$.

It follows that *margin bounds obtained for thresholded classifiers apply for arbitrary zero-one strategies*, when the margin loss is defined with respect to the partition of $\mathcal{E}$ induced by g.

Example 5.9. Consider the strategy on $\mathcal{E}$ defined by

$$g(\eta) = I(||\eta - \eta_0|| > \varepsilon)$$

for some $\eta_0 \in \mathcal{E}$, and $\varepsilon \in \mathbb{R}$, so that the portion classified as zero lies in a ball of radius ε in $\mathcal{E}$.

In this case, we can see that the margin of h on z is defined by

$$\rho(z,h) = (||h(x) - \eta_0|| - \varepsilon)\,\mathrm{sgn}\,y \ \ ,$$

and the bounds of thresholded classifiers are directly applicable here.

It is interesting to note that the form for the margin here is extremely similar to that observed when considering thresholding classifiers and the ε-insensitive loss.

Note furthermore, that a number of strange shaped sets can be obtained by using an appropriate metric on $\mathcal{E}$. For example, if we consider $\mathbb{R}^n$ with the Manhattan metric, we can apply margin bounds to classification based on whether $h(x)$ lies in a box. □

Example 5.10. Let us consider a somewhat more complex strategy. Consider a voting machine composed of N thresholded classifiers, each with threshold s_i. Let $s = (s_1, \cdots, s_N)$. We can obtain a margin bound based on the output of each classifier directly.

Consider $\eta = (\eta_1, \cdots, \eta_N) \in \mathcal{E} = \mathbb{R}^N$. Let $g(\eta)$ be one exactly when at least j coordinates of η exceed the corresponding thresholds in s. Assume without loss of generality that $s = 0$. Then the strategy partitions $\mathbb{R}^N$ into two sets of orthants, those with j or more nonnegative coordinates, and those with less than j.

Our results show that if we can calculate the margin of any point $\eta \in \mathbb{R}^n$, we can apply the margin bounds for thresholding classifiers using the corresponding margin loss.

Consider the Manhattan metric d_1, and assume a point $\eta \in \mathbb{R}^N$ has $i \geq j$ nonnegative coordinates. For η' to be classified differently to η, it needs less than j nonnegative coordinates. Such a point must have a distance from η of at least the sum of the $i - j + 1$ smallest nonnegative coordinates of η. Similarly comparing a point η with $i < j$ nonnegative coordinates to an η' with at least j nonnegative coordinates, shows that the distance between them is at least the negated sum of the $i - j$ smallest negative coordinates of η. Furthermore, these distances can be approximated arbitrarily closely by appropriate selection of η', so that these expressions define the margin.

With this definition, it is not difficult to calculate the margin, and thus apply the margin bound, yielding a bound which uses the real outputs of each classifier of a voting classifier. □

5.6.4 γ and Occam's razor

Note that pre-specifying γ with the bounds above may lead one to overly pessimistic bounds: if a large margin is attained over the whole sample, but γ is selected too small, the bound is likely to be overly pessimistic — picking a larger γ would be more useful. Thus selecting γ on the basis of the selected hypothesis and its behaviour on the data is desirable.

The most obvious approach to achieve this while avoiding dependence problems is to make the bound uniform for all γ by employing the Occam's razor method in the same way we used it to obtain uniform bounds over a countable decision class. Unfortunately, however, γ is a real value, so that it seems we need to employ a cover over a grid of potential values of γ. This is very similar to the approach we outline next.

For an hypothesis class $\mathcal{H}$, define the sets $R_1, R_2, \cdots$ by

$$R_i = \left\{ \gamma : \left\lfloor \log_2 \left(\mathbb{E}_{Q \sim D^{m+u}} \mathcal{N}_{\infty,Q} \left(\frac{\gamma}{2}, \mathcal{H} \right) \right) \right\rfloor = i - 1 \right\} \ .$$

That is, R_i consists of those γ for which the expected covering number lies in $[2^{i-1}, 2^i)$. Clearly, the R_i form intervals which partition $\mathbb{R}^+$.

Suppose we have a bound on

$$\mathbb{P}_{S \sim D^m} \left\{ \sup_{h \in \mathcal{H}} \psi(r_D(g_h, L), r_S(h, L_\gamma)) > \epsilon \right\}$$

in terms of the expected covering number above for any pre-specified γ. Let $\gamma_i = \inf R_i$.

Now consider any $\gamma \geq \gamma_i$, so that

$$L_\gamma \geq L_{\gamma_i} \ .$$

Assuming that ψ is decreasing in its second argument, this implies that

$$\psi(r_D(g_h, L), r_S(h, L_\gamma)) \leq \psi(r_D(g_h, L), r_S(h, L_{\gamma_i})) \ .$$

Thus

$$\mathbb{P}_{S\sim D^m}\left\{\forall\gamma\geq\gamma_i : \sup_{h\in\mathcal{H}}\psi(r_D(g_h,L),r_S(h,L_\gamma))>\epsilon\right\}$$
$$=\mathbb{P}_{S\sim D^m}\left\{\sup_{h\in\mathcal{H}}\psi(r_D(g_h,L),r_S(h,L_{\gamma_i}))>\epsilon\right\} \ .$$

It follows that by employing the union bound on the probability of ψ-deviation at each γ_i will yield a bound for all γ. Let $\alpha^\star(i)$ be a "prior" for $i\in\mathbb{N}$ for this purpose.

We now illustrate the results of this approach using the realizable margin bound of (5.47). Let us write

$$\delta(\epsilon,i,m,u)=\beta^{-1}\,\mathbb{E}_{Q\sim D^{m+u}}\mathcal{N}_{\infty,Q}\left(\frac{\gamma_i}{2},\mathcal{H}\right)\left(\frac{u}{m+u}\right)^{\lceil u\alpha(u,\beta,\epsilon)\epsilon\rceil} \ .$$

From the union bound on γ_i, $i\in\mathbb{N}$, we have

$$\mathbb{P}_{S\sim D^m}\left\{\forall i\in\mathbb{N}:\exists h\in\mathcal{H}:(e_S(h,L_{\gamma_i})=0)\wedge(e_D(g_h,L)>\epsilon_i)\right\}$$
$$\leq\sum_{i=1}^{\infty}\delta(\epsilon_i,i,m,u) \ . \quad (5.48)$$

Since the bound for any γ_i applies simultaneously to all $\gamma\geq\gamma_i$, we have

$$\mathbb{P}_{S\sim D^m}\left\{\forall\gamma\in\mathbb{R}:\exists h\in\mathcal{H}:(e_S(h,L_\gamma)=0)\wedge(e_D(g_h,L)>\epsilon_i)\right\}$$
$$\leq\sum_{i=1}^{\infty}\delta(\epsilon_i,i,m,u) \ . \quad (5.49)$$

Some important points to note:

- Note that the right hand side is fixed, regardless of γ, for any selection of the ϵ_i. Any selection of $\epsilon_i>0$ is valid, but typically, they can be specified by employing the "prior" $\alpha^\star(i)$. This is done by setting

$$\delta(\epsilon_i,i,m,u)=\delta\alpha^\star(i)$$

for some desired overall confidence level $1-\delta$. This ensures that

$$\sum_{i=1}^{\infty}\delta(\epsilon_i,i,m,u)=\delta \ .$$

- Even if this inversion can not be done precisely, we still obtain a uniform bound if we choose the ϵ_i to satisfy

  $$\delta(\epsilon_i, i, m, u) \leq \delta\alpha^\star(i) \ .$$

  Generally, the idea is to select smaller ϵ_i for small i. In practice, we would generally like the bounds to be tightest for the smallest choice of i satisfying $\gamma_0 \geq \gamma_i$, where γ_0 is the largest choice of γ for which $e_S(h, L_\gamma) = 0$.

- Note that the same analysis could be performed with respect to any upper bound on $\mathbb{E}_{Q\sim D^{m+u}}\mathcal{N}_{\infty,Q}(\frac{\gamma}{2}, \mathcal{H})$. In practice, this is extremely important, because we can not generally know the exact values of γ_i if defined with respect to the expected covering numbers. However, if we use some computable alternative, we can find an appropriate choice of i for implementing this bound.

- One guiding principle in applying a "prior" to such a bound is to consider the form of the bound. There is no point allocating confidence to choices of i and ϵ_i for which the bound will be trivial.

- Note that $\lim_{i\to\infty}\gamma_i = 0$, so the bound can always be applied.

In order to apply a "prior" to the result above will require significant numerical calculations due to the symmetrization lemma employed. We will present another method for making this bound uniform soon.

To illustrate the approach, however, we shall use the relative deviation bound of (5.45), and the uniform "prior" $\alpha^\star(i) = \frac{1}{2m}$ on $[1 : 2m]$[109]. We thus set

$$\frac{\delta}{2m} = 4\,\mathbb{E}_{Q\sim D^{2m}}\mathcal{N}_{\infty,Q}\left(\frac{\gamma_i}{2}, \mathcal{H}\right)\exp\left(\frac{-\epsilon_i^2 m}{4}\right)$$

and solve for ϵ_i in terms of γ_i. This yields

$$\epsilon_i = \epsilon(\gamma_i) = \sqrt{\frac{4\ln\left(\frac{8m\,\mathbb{E}_{Q\sim D^{2m}}\mathcal{N}_{\infty,Q}\left(\frac{\gamma_i}{2},\mathcal{H}\right)}{\delta}\right)}{m}} \ ,$$

[109] Thus $\alpha^\star(i) = 0$ for $i > 2m$, so the bound will not apply to any γ such that the expected covering number at scale $\frac{\gamma}{2}$ exceeds 2^{2m}.

for $i \in [1 : 2m]$.

Thus, we have

$$\mathbb{P}_{S \sim D^m}\left\{\forall i \in [1 : 2m] : \left(\forall \gamma \geq \gamma_i : \sup_{h \in \mathcal{H}} \frac{r_D(g_h, L) - r_S(h, L_\gamma)}{\sqrt{r_D(g_h, L)}} > \epsilon_i\right)\right\} < \delta \ .$$

What makes this inversion more attractive is that in this form we can make use of a relaxation which obviates the need to know the relationship of γ to any of the γ_i: since covering numbers are decreasing functions of the resolution, $\gamma \leq \gamma_i$ implies that $\epsilon(\gamma) \geq \epsilon(\gamma_i)$. It follows that

$$\mathbb{P}_{S \sim D^m}\left\{\forall \gamma \geq \gamma_{2m} : \sup_{h \in \mathcal{H}} \frac{r_D(g_h, L) - r_S(h, L_\gamma)}{\sqrt{r_D(g_h, L)}} > \epsilon(\gamma)\right\} < \delta \ .$$

Converting this to a realizable case bound provides a weaker analog of the bound in Shawe-Taylor and Cristianini (1999, Theorem 4.7).[110]

An example of the benefits reaped from not allocating confidence in the "prior" when the resulting bound is trivial is the improvement from Shawe-Taylor and Cristianini (1999, Theorem 4.7), which employed a uniform bound on $[1 : 2m]$, to Shawe-Taylor and Cristianini (2000, Theorem VII.7), which noted that allocating confidence to $i > \frac{m}{2}$ was pointless given the form of the bound, so instead applied a uniform "prior" on $[1 : \frac{m}{2}]$ instead.

An alternative, but highly similar approach is based on the following result from Bartlett (1998), which we shall call the *margin unification lemma.*

Lemma 5.1 (Proposition 8 of Bartlett, 1998). *Let $(\mathcal{E}, \Sigma, \tau)$ be a probability space, and let*

$$\{\mathscr{E}(v_1, v_2, v_3) : 0 < v_1, v_2, v_3 \leq 1\}$$

be a set of events satisfying the following conditions:

1. *for all $0 < c \leq 1$ and $0 < v_3 \leq 1$, $\tau(\mathscr{E}(c, c, v_3)) \leq v_3$;*

[110] That bound is obtained by combining the realizable dual sample bound of Theorem 5.14 with a variant of the traditional symmetrization lemma, modified for the realizable case.

2. *for all* $0 < c < 1$ *and* $0 < v_3 \leq 1$,

$$\bigcup_{v_2 \in (0,1]} \mathscr{E}(cv_2, v_2, (1-c)v_2v_3)$$

is measurable; and

3. *for all* $0 < v_1 \leq c \leq v_2 \leq 1$ *and* $0 < v' \leq v_3 \leq 1$,

$$\mathscr{E}(v_1, v_2, v') \subseteq \mathscr{E}(c, c, v') \ .$$

Then, for $0 < c, v_3 < 1$,

$$\tau\left(\mathscr{E}(cv_2, v_2, (1-c)v_2v_3)\right) \leq v_3 \ .$$

The reader is referred to the article for the proof. In our case, the lemma assumes that we have some event (such as large deviation of true loss from margin loss) which holds with probability at most δ for any individual $\gamma \leq 1$ (condition 1). If the modified event obtained by

- decreasing some occurences of γ to any smaller γ_1;
- increasing the other occurences of γ to any larger γ_2; and
- reducing any occurence of δ to a smaller δ'

implies the original event (condition 3), then (ignoring measurability issues) we can obtain a uniform bound for all $\gamma \in (0, 1]$[111].

We now illustrate the results of this approach using the realizable margin bound of (5.47). Suppose $\epsilon(\delta, \gamma)$ is the solution (usually obtained numerically) for ϵ of

$$\delta = \beta^{-1}\, \mathbb{E}_{Q \sim D^{m+u}}\, \mathcal{N}_{\infty,Q}\left(\frac{\gamma}{2}, \mathcal{H}\right) \left(\frac{u}{m+u}\right)^{u\alpha(u,\beta,\epsilon)\epsilon}$$

for an appropriate α and β.

[111] Extending the maximum margin permissible here from 1 to some other value requires a straightforward modification to the margin unification lemma, akin to scaling the margin. For such a modified result, see Kroon (2003, Theorem 68).

Let

$$\mathscr{E}(v_1, v_2, v_3) = \{S \in \mathcal{Z}^m : (\exists h \in \mathcal{H} : (e_S(h, L_{v_2}) = 0) \wedge (e_D(g_h, L) > \epsilon(v_3, v_1)))\} \ .$$

Clearly, condition 1 of the lemma is met with $c = \gamma$ and $v_3 = \delta$. We shall now consider condition 3. Suppose that a sample $S \in \mathscr{E}(\gamma_1, \gamma_2, \delta')$, with $0 < \gamma_1 \leq \gamma \leq \gamma_2$ and $0 < \delta' \leq \delta \leq 1$. Consider any $h \in \mathcal{H}$ such that $e_S(h, L_{\gamma_2}) = 0$ and

$$e_D(g_h, L) > \epsilon(\delta', \gamma_1) \ .$$

Since $\gamma \leq \gamma_2$, $e_S(h, L_\gamma) = 0$.

When the α in (5.47) is determined by α_{C-R}, $\epsilon(\delta, \gamma)$ solves for ϵ in

$$\frac{\ln \frac{\delta\beta}{\mathbb{E}_{Q \sim D^{m+u}} \mathcal{N}_{\infty,Q}\left(\frac{\gamma}{2}, \mathcal{H}\right)}}{u \ln \frac{u}{m+u}} = \epsilon \left(1 - \sqrt{\frac{\frac{1}{p} - 1}{u(1-\beta)}}\right) \ .$$

The left hand side here is decreasing in δ, and, since covering numbers decrease with an increase in the scale, in γ. (Note the negative denominator). Furthermore, the right hand side is increasing in ϵ. It follows that $\epsilon(v_3, v_1)$ is decreasing in v_3 and v_1. Therefore,

$$\epsilon(\delta, \gamma) \leq \epsilon(\delta', \gamma_1) \ ,$$

so that

$$e_D(g_h, L) > \epsilon(\delta, \gamma) \ ,$$

implying that $S \in \mathscr{E}(\gamma, \gamma, \delta)$. Thus condition 3 holds, and we can apply the margin unification lemma. We obtain the following result.

Theorem 5.23 (Realizable margin bound for error). *Select some* $0 < v, \delta < 1$. *Then*

$$\mathbb{P}_{S \sim D^m} \{\forall \gamma \in (0, 1] : (\exists h \in \mathcal{H} : (e_S(h, L_\gamma) = 0) \wedge (e_D(g_h, L) > \epsilon(v\gamma, \gamma\delta(1-v))))\} \leq \delta \ .$$

In short, taking as an example $v = \frac{1}{2}$, we can select the margin from $(0, 1]$ after selecting the hypothesis at the cost of doubling the resolution (i.e. dividing the scale parameter by 2) of the cover of $\mathcal{H}$, and replacing δ by

$\frac{\delta\gamma}{2}$. Since we are interested in large values of γ, the lost confidence from this application of the margin unification lemma should hopefully not be too punitive.

This approach is very convenient, since any reasonable margin bound will satisfy the conditions of the margin unification lemma. A study of the proof of the lemma shows that it employs a geometrically decreasing "prior" over geometrically decreasing portions of $(0, 1]$. This result then effectively seems to allocate confidence uniformly over the entire region, in contrast to the first method discussed, where an arbitrary "prior" can be employed.

5.6.5 Margin distribution bounds

Shawe-Taylor and Cristianini (1998b) provides an incredibly interesting method for transforming a problem to ensure that a pre-specified hard margin is attained by all the points in the sample on the modified problem. Standard realizable margin bounds can then be applied to the transformed problem, yielding bounds which apply to the original problem. Furthermore, the exact margin which will be attained is known a priori. This approach is, however, limited to thresholding classifiers.

The process involves a trade-off: the transformation of the problem increases the covering numbers, but at the same time it allows application of the more rapidly decaying realizable bounds. We shall also see that this approach removes the need to make the bound uniform over the margin at the expense of introducing a new value over which the bound must be made uniform.

The approach employs a generalization of an ingenious technique seemingly first used for a different, but related, purpose in Klasner and Simon (1995), and further developed in this direction by Freund and Schapire (1999). The resulting *margin embedding* technique results in so-called *margin distribution*, or *soft margin* bounds: it was shown (Shawe-Taylor and Cristianini, 1998b) that the approach has strong parallels with the soft-margin SV machine classifier proposed in Cortes and Vapnik (1995).

The approach relies on transforming the hypotheses for a given strategy to ensure that all the points in the training sample will attain some pre-specified minimum margin γ. Realizable margin bounds can then be applied to the transformed hypotheses. The transformation is further chosen so that the new hypotheses behave identically to the original hypotheses everywhere in $\mathcal{X}$ *except on the points in* S. A result of this is that bounds on the error of transformed hypotheses also appply to the original hypotheses. We detail the transformation in what follows.

Select some desired margin γ_0, and consider any hypothesis $h \in \mathcal{H}$. Then, we can consider the *margin shortfall* $\max(0, \gamma_0 - \rho(z, h))$ of any point in $\mathcal{Z}$. This is the amount by which the margin achieved by h on z, $\rho(z, h)$, falls short of γ_0. We employ these margin shortfalls on the points in the training sample S, along with the Kronecker delta.

We define $\mathcal{L}(\mathcal{X})$ as the set of functionals on $\mathcal{X}$ which have finite support[112]. We endow $\mathcal{L}(\mathcal{X})$ with an inner product defined by

$$\langle v_1, v_2 \rangle = \sum_{x \in \mathrm{supp}(v_1)} v_1(x) v_2(x) \ \ ,$$

where $\mathrm{supp}(v_1)$ denotes the support of v_1.

For any $x_0 \in \mathcal{X}$, the Kronecker delta function, $\delta_{x_0} = I(x = x_0)$, belongs to $\mathcal{L}(\mathcal{X})$. We now define the mapping $\xi : \mathcal{X} \to \mathcal{X} \times \mathcal{L}(\mathcal{X})$ by $\xi(x) = (x, \delta_x)$. We will transform h so that it operates on this transformed input space, and attains a margin of γ_0 on the transformed training sample. We define $h_\xi : \mathcal{X} \times \mathcal{L}(\mathcal{X}) \to \mathbb{R}$ by

$$h_\xi((x, v)) = h(x) + \left\langle \sum_{(x', y') \in S} \max\left(0, \gamma_0 - \gamma((x', y'))\right) y' \delta_{x'}, v \right\rangle \ \ .$$

Note that the first function in the inner product above is a sum of weighted Kronecker delta functions over the sample, with each weight equal to the margin shortfall for the corresponding element of the sample. Thus, if v is a Kronecker delta function corresponding to a sample point, the inner product term will add just enough to the original predicted output to attain a

[112] i.e. each function is non-zero at finitely many points of $\mathcal{X}$.

margin of γ_0, if necessary. However, if v is the Kronecker delta function corresponding to a point not in the training sample, the inner product will be zero, so the modified hypothesis behaves identically to the original hypothesis. Assuming that no misclassified points are assigned non-zero measure by D, it follows that h and h_ξ are equal D-almost everywhere, and that h_ξ achieves a margin of γ_0 on the transformed sample, allowing the application of the optimistic margin bound. The complication is that the new hypothesis comes from a larger class than $\mathcal{H}$, and it is the covering number of the larger class that needs to be used: if we define $h_1((x,v)) = h(x)$ and

$$h_2((x,v)) = \left\langle \sum_{(x',y')\in S} \max\left(0, \gamma_0 - \gamma((x',y'))\right) y'\delta_{x'}, v \right\rangle \ ,$$

we can write $h_\xi = h_1 + h_2 \in \mathcal{H} + \mathcal{L}(\mathcal{X})$, where

$$\mathcal{H} + \mathcal{L}(\mathcal{X}) = \{h + v : h \in \mathcal{H}, v \in \mathcal{L}(\mathcal{X})\} \ .$$

It seems tempting to simply apply the optimistic margin bound in this situation. However, the class $\mathcal{L}(\mathcal{X})$ is too rich, so the covering numbers involved will be too large. The solution is to restrict the choice of functions in $\mathcal{L}(\mathcal{X})$. One approach is to upper bound the allowable norm of h_2: note that for any training sample S, we have that $\|h_2\|$ equals

$$\begin{aligned} & \left\| \sum_{(x',y')\in S} \max\left(0, \gamma_0 - \gamma((x',y'))\right) y'\delta_{x'} \right\| \\ = & \sqrt{\left\langle \sum_{(x',y')\in S} \max(0, \gamma_0 - \gamma((x',y')))y'\delta_{x'}, \sum_{(x',y')\in S} \max(0, \gamma_0 - \gamma((x',y')))y'\delta_{x'} \right\rangle} \\ = & \sqrt{\sum_{(x',y')\in S} \max(0, \gamma_0 - \gamma((x',y')))^2} \ . \end{aligned}$$

In addition, it is useful to note that this approach only ever considers $h((x,v))$ for

$$v \in \delta_{\mathcal{X}} = \{\delta_x : x \in \mathcal{X}\} \ .$$

This restriction allows us to employ so-called radius-margin bounds[113] for the calculation of covering numbers, since for any x, $\|\delta_x\| = 1$.

[113] Radius margin bounds are discussed in Section 5.9.4.

Now, if $J \geq \|h_2\|$ we can view h_ξ as an element of $\mathcal{H} + \mathcal{L}_J(\mathcal{X})|_{\delta_\mathcal{X}}$, where $\mathcal{L}_J(\mathcal{X})$ consists of the functions in $\mathcal{L}(\mathcal{X})$ with norm not exceeding J.

By considering balls in $\mathcal{L}(\mathcal{X})$ with successively increasing radii J_i, similarly to the approach in Shawe-Taylor and Cristianini (1998b), and applying the union bound, a bound which holds regardless of $\|h_2\|$ can be obtained.

For each i, we wish to apply the realizable bound of (5.47). For any $h \in \mathcal{H}$, denote the transformed hypothesis by h_ξ as above. Let $\mathcal{H}_\xi = \{h_\xi : h \in \mathcal{H}\}$. For any $h' \in \mathcal{H}_\xi$ we write h'_2 for the component of h' in $\mathcal{L}(\mathcal{X})$. Then, for the appropriate choices of α and β for that bound, we have

$$\begin{aligned}
& \mathbb{P}_{S\sim D^m}\left\{\exists h \in \mathcal{H} : ((h_\xi)_2 \in \mathcal{L}_{J_i}(\mathcal{X})|_{\delta_\mathcal{X}}) \wedge (e_D(g_h, L) > \epsilon)\right\} \\
= \; & \mathbb{P}_{S\sim D^m}\{\exists h' \in \mathcal{H}_\xi : (h'_2 \in \mathcal{L}_{J_i}(\mathcal{X})|_{\delta_\mathcal{X}}) \wedge (e_S(h', L_{\gamma_0}) = 0) \wedge (e_D(g_{h'}, L) > \epsilon)\} \\
= \; & \mathbb{P}_{S\sim D^m}\{\exists h' \in \mathcal{H}_\xi \cup (\mathcal{H} + \mathcal{L}_{J_i}(\mathcal{X})|_{\delta_\mathcal{X}}) : (e_S(h', L_{\gamma_0}) = 0) \wedge (e_D(g_{h'}, L) > \epsilon)\} \\
\leq \; & \beta^{-1}\, \mathbb{E}_{Q\sim D^{m+u}}\, \mathcal{N}_{\infty,Q}\left(\frac{\gamma}{2}, \mathcal{H} + \mathcal{L}_{J_i}(\mathcal{X})|_{\delta_\mathcal{X}}\right)\left(\frac{u}{m+u}\right)^{\lceil u\alpha(u,\beta,\epsilon)\epsilon\rceil} . \qquad (5.50)
\end{aligned}$$

Note that the transformed class $\mathcal{H}_\xi$ depends on the training sample S. This is why the covering number in the bound of the last line is not over the intersection

$$\mathcal{H}_\xi \cap (\mathcal{H} + \mathcal{L}_{J_i}(\mathcal{X})|_{\delta_\mathcal{X}}) \;,$$

but over the larger class $\mathcal{H} + \mathcal{L}_{J_i}(\mathcal{X})|_{\delta_\mathcal{X}}$. This is because this class contains the intersection for every sample, allowing one to obtain the bound despite the data-dependence of the class.

Finally, we apply the weighted union bound over the potential choices of J_i to obtain[114]

$$\begin{aligned}
& \mathbb{P}_{S\sim D^m}\left\{\exists i \in \mathbb{N} : (\exists h \in \mathcal{H} : ((h_\xi)_2 \in \mathcal{L}_{J_i}(\mathcal{X})|_{\delta_\mathcal{X}}) \wedge (e_D(g_h, L) > \epsilon_i))\right\} \\
& \leq \sum_{i=1}^{\infty} \beta^{-1}\, \mathbb{E}_{Q\sim D^{m+u}}\, \mathcal{N}_{\infty,Q}\left(\frac{\gamma}{2}, \mathcal{H} + \mathcal{L}_{J_i}(\mathcal{X})|_{\delta_\mathcal{X}}\right)\left(\frac{u}{m+u}\right)^{\lceil u\alpha(u,\beta,\epsilon_i)\epsilon_i\rceil} . \qquad (5.51)
\end{aligned}$$

[114] One may be able to apply the margin unification lemma instead.

Notes similar to those after (5.49) apply here regarding the choice of the ρ_i. This result is based on a modified version of Theorem 5.15. The original bounds of this form were based on a modification of Vapnik's symmetrization lemma for the realizable case. Early forms of this modified symmetrization lemma employed $m = u$, so those results appear in that form. However, Herbrich and Williamson (2002, 2004) provide an extension of this modification to the case $m \neq u$, so that the original margin distribution results can be extended to that case — for the modified symmetrization lemma, see Theorem 6.1.

Regarding selection of the J_i, we suggest using an arithmetic sequence if the underlying functions are bounded in a small interval (allowing one to obtain a maximal J_i), while a geometric sequence is recommended if the underlying functions are unbounded.

Finally, we mention that in practice, the ϵ_i will be determined by specifying a desired confidence level δ, together with a "prior" over the J_i. These values permit a numerical inversion of each individual bound to obtain the appropriate choice of ϵ_i. In this case, the right hand side of this bound sums to δ.

5.6.6 Discussion

The development presented here does not match the development of the results in theory. Bounds were first developed for the so-called hard-margin case, which corresponds to our realizable bounds, first in the case of ε-insensitive function learning, and then for thresholding classifiers. Note that such results are (strongly) optimistic: the bound only applies to functions correctly classifying the whole training sample with the unthresholded function achieving a certain (pre-specified) margin on all the points.

Bartlett (1998) provided the extension to the general case, providing the capability to still obtain bounds when the minimal margin on the sample did not exceed the selected γ. These bounds are then called *margin percentile* bounds, since the choice of γ corresponds to a quantile of the empirical distribution of the margin on the sample.

These general case bounds are particularly useful when there are no hypotheses in $\mathcal{H}$ such that g_h is consistent with the data. In this case, no choice of γ can yield a non-trivial bound based on the realizable bounds.

The general case bounds of Bartlett also make it clear that generalization of thresholded classifiers not only depends on the minimum margin on the sample, but on the distribution over the whole sample. Consider a situation in which an hypothesis attains a margin near 1 on all but one of the points in a sample, while achieving a margin very close to 0 on the remaining point. The realizable bound could only obtain a bound by selecting γ less than the smallest margin, which is not representative of the whole sample's margins. With the general case bounds, one can probably substantially improve the bound by using a larger γ, near 1: the margin error $e_S(h, L_\gamma)$ increases from 0 to $\frac{1}{m}$, and the bound decays slower, but the reduction in the covering number due to the reduced resolution required may well far outweigh this downside.

5.7 Chaining

Chaining is a powerful technique in the field of stochastic processes — an excellent discussion of the technique by one of the pioneering authors is Talagrand (2005), while Alexander (1984) attributes the first application of the technique to empirical processes to Dudley (1978). The contents of this section are closely related to the derivation of Dudley's entropy bound on the expected supremum of a stochastic process, as detailed in the above reference.

The core concepts employed in chaining were pioneered by Kolmogorov, and their application to bounding stochastic processes was promoted in the work of Richard Dudley in the 1960's and 1970's.

Chaining allows one to obtain bounds on the Rademacher penalty by expressing each function as a telescoping series. The terms in the telescoping series are differences between carefully selected successive approximations to the given function.

Consider $\mathcal{W}$, a sample $S = [\![x_1, \cdots, x_m]\!]$, and the metric $d = d_{2,S}$[115]. Then

$$\operatorname{diam}(\mathcal{W}) = \sup\{d(w_1, w_2) : w_1, w_2 \in \mathcal{W}\} \leq 1 \ .$$

We consider a sequence of subclasses

$$\mathcal{W}_0 \subseteq \mathcal{W}_1 \subseteq \cdots \subseteq \mathcal{W} \ ,$$

where each $\mathcal{W}_j$ is a maximal (2^{-j})-separated subset of $\mathcal{W}$.

Consider any $j \geq 0$. Now, by the definition of a maximal ϵ-separated subset, it follows that for any $w \in \mathcal{W}$, there is at least one element of $\mathcal{W}_j$ within distance 2^{-j} of w. Pick an arbitrary such element w_j, and define $\mathscr{P}_j(w) = w_j$. This $\mathscr{P}_j$ is then a kind of "projection function", which maps any w onto an approximation $\mathscr{P}_j(w)$ such that $d(w, \mathscr{P}_j(w)) \leq 2^{-j}$.[116] Clearly, $\mathcal{W}_0$ has exactly one element, since no two elements of $\mathcal{F}$ are 2^n-seperated, and we can choose that element arbitrarily. We select it to be the constant decision rule w_0 which always outputs zero.[117] We thus have $\mathscr{P}_0(w) = w_0$ for all $w \in \mathcal{W}$.

Since $\mathscr{P}_j(w)$ and $\mathscr{P}_{j-1}(w)$ are both close to w, it follows that they are close to each other. Employing the triangle inequality, going via w, we obtain that $d(\mathscr{P}_j(w), \mathscr{P}_{j-1}(w)) \leq 3 \cdot 2^{-j}$.

In order for chaining to yield useful results, we need some relationship to hold between values assumed by the process at points close to each other in $\mathcal{W}$. We shall be interested in applying chaining to the stochastic process $\{\frac{1}{m}\sum_{i=1}^{m} \zeta_i w(x_i) : w \in \mathcal{W}\}$, where $\zeta = (\zeta_1, \cdots, \zeta_m)$ is a sequence of independent Rademacher variables. We shall call this process the *Rademacher process*. Note that the supremum of this process over the index set is the

[115] The reason for the choice of $p = 2$ will be made clear later.

[116] Such a projection function is not generally unique, but any one will suffice for our purposes.

[117] Technically, this decision rule is not necessarily in $\mathcal{W}$. However, if we add it to $\mathcal{W}$ the results we obtain also hold for the original class, so we shall not worry about this further.

Rademacher penalty $\mathcal{R}_S(\mathcal{W})$. So, for two decision rules w_1 and w_2, we consider

$$(r_D(w_1) - r_S(w_1)) - (r_D(w_2) - r_S(w_2)) = r_S(w_2) - r_S(w_1) \ .$$

The kind of result we desire is that if two decision rules in $\mathcal{W}$ are close (w.r.t. d), then it is highly likely that the process assumes similar values at those two points.

Such a result is easily obtained from Hoeffding's inequality. Consider any $w_1, w_2 \in \mathcal{W}$. Then

$$\begin{aligned} & \mathbb{P}_{\zeta \sim (\mathrm{Unif}\{-1,1\}^m)} \left\{ \frac{1}{m} \sum_{i=1}^{m} \zeta_i (w_1(x_i) - w_2(x_i)) > \epsilon | S \right\} \\ \leq \ & \exp\left(\frac{-2m^2\epsilon^2}{\sum_{i=1}^{m} (2[w_1(x_i) - w_2(x_i)])^2} \right) \\ = \ & \exp\left(\frac{-m\epsilon^2}{2(d(w_1, w_2))^2} \right) \ . \end{aligned}$$

Note that the last inequality does not generally hold for $d_{p,S}$ where $p \neq 2$. Clearly, the closer w_1 and w_2 are, the less likely it is that the value assumed by the Rademacher process at w_1 will exceed its value at w_2 by any prespecified amount.

Now, we can decompose any $w \in \mathcal{W}$ on the basis of its projections:

$$w = w_0 + \sum_{j=1}^{\infty} [\mathscr{P}_j(w) - \mathscr{P}_{j-1}(w)] \ .$$

It follows that

$$\sum_{i=1}^{m} \zeta_i w(x_i) = \sum_{i=1}^{m} \zeta_i w_0(x_i) + \sum_{j=1}^{\infty} \zeta_i \sum_{i=1}^{m} (\mathscr{P}_j(w)(x_i) - \mathscr{P}_{j-1}(w)(x_i)) \ .$$

For any given j and w, we know that $d(\mathscr{P}_j(w), \mathscr{P}_{j-1}(w)) \leq 3 \cdot 2^{-j}$. It follows that

$$\begin{aligned} \mathbb{P}_{\zeta \sim (\mathrm{Unif}\{-1,1\}^m)} & \left\{ \frac{1}{m} \sum_{i=1}^{m} \zeta_i ([\mathscr{P}_j(w)](x_i) - [\mathscr{P}_{j-1}(w)](x_i)) > \epsilon | S \right\} \\ & \leq \exp\left(\frac{-m\epsilon^2}{2(3 \cdot 2^{-j})^2} \right) \ . \end{aligned}$$

For a specific w, we can thus obtain an upper bound on

$$\mathbb{P}_{\zeta\sim(\mathrm{Unif}\{-1,1\}^m)}\left\{\frac{1}{m}\sum_{i=1}^{m}\zeta_i w(x_i) > \epsilon|S\right\}$$

by employing a countable union bound over each link (since the first term is zero). We would like to extend this result to hold for all $w \in \mathcal{W}$. Unfortunately, we can not simply apply the Occam's razor method over $\mathcal{W}$, since $\mathcal{W}$ is generally uncountable. One approach to circumventing this problem is to apply the Occam's razor method to $\mathcal{W}_j$, for a large enough j, and then correct for any inaccuracies. This is, however, effectively the same as using the normal covering number approach to obtain bounds. Chaining obtains a great improvement by getting a result which applies to all of $\mathcal{W}_j$ by applying the Occam's razor method over the individual links, and employing the union bound within each set of links! This results in a great reduction in "lost confidence" in the Occam's razor method.

It is clear that there are at most $|\mathcal{W}_j|\cdot|\mathcal{W}_{j-1}|$ possible pairs $(\mathscr{P}_j(w), \mathscr{P}_{j-1}(w))$. However, we can do better by directly employing the size of the set

$$\mathcal{B}_j = \{(\mathscr{P}_j(w), \mathscr{P}_{j-1}(w)) : w \in \mathcal{W}\} \quad .$$

Applying the uniform Occam's razor method over the links in $\mathcal{B}_j$ yields

$$\mathbb{P}_{\zeta\sim(\mathrm{Unif}\{-1,1\}^m)}\left\{\sup_{(w_1,w_2)\in\mathcal{B}_j}\left[\frac{1}{m}\sum_{i=1}^{m}\zeta_i(w_1(x_i)-w_2(x_i))\right] > \epsilon|S\right\} \\ \leq |\mathcal{B}_j|\exp\left(\frac{-m\epsilon^2}{2(3\cdot 2^{-j})^2}\right) \quad .$$

Equating the probability to $\delta(j)$ and solving for ϵ we obtain

$$\mathbb{P}_{\zeta\sim(\mathrm{Unif}\{-1,1\}^m)}\left\{\begin{array}{c}\sup_{(w_1,w_2)\in\mathcal{B}_j}\left[\frac{1}{m}\sum_{i=1}^{m}\zeta_i(w_1(x_i)-w_2(x_i))\right] \\ > 3\cdot 2^{-j}\sqrt{\frac{2(\ln|\mathcal{B}_j|-\ln\delta(j))}{m}}\end{array}|S\right\} \leq \delta(j) \quad .$$

We now have a probabilistic bound for the j-th link in the chain stretching from w_0 to any $w \in \mathcal{W}$, for any j. Next, we can apply the Occam's razor method non-uniformly to all j to obtain

$$\mathbb{P}_{\zeta\sim(\mathrm{Unif}\{-1,1\}^m)}\left\{\exists j\in\mathbb{N}:\begin{array}{c}\sup_{(w_1,w_2)\in\mathcal{B}_j}\left[\frac{1}{m}\sum_{i=1}^m\zeta_i(w_1(x_i)-w_2(x_i))\right]\\ >3\cdot 2^{-j}\sqrt{\frac{2(\ln|\mathcal{B}_j|-\ln\delta(j))}{m}}\end{array}|S\right\}$$
$$\leq\sum_{j=1}^{\infty}\delta(j)\ .$$

Suppose the $\delta(j)$ are selected in such a way that the sum on the right is finite and equal to $\delta\in(0,1]$. Then, with probability at least $1-\delta$, for every decision rule $w\in\mathcal{W}$, we have that every term in the telescoping series expansion of $\frac{1}{m}\sum_{i=1}^m\zeta_i w(x_i)$ is bound by

$$3\cdot 2^{-j}\sqrt{\frac{2(\ln|\mathcal{B}_j|-\ln\delta(j))}{m}}\ ,$$

so that the infinite sum does not exceed

$$3\sqrt{\frac{2}{m}}\sum_{j=1}^{\infty}\left(2^{-j}\sqrt{\ln|\mathcal{B}_j|-\ln\delta(j)}\right)\ .$$

Together this yields a chaining bound in full generality, which provides an alternative to the random subsample bound of (5.35):

$$\mathbb{P}_{\zeta\sim\mathrm{Unif}(\{-1,1\}^m)}\left\{\frac{1}{m}\sum_{i=1}^m\zeta_i w(x_i)>3\sqrt{\frac{2}{m}}\sum_{j=1}^{\infty}\left(2^{-j}\sqrt{\ln|\mathcal{B}_j|-\ln\delta(j)}\right)|S\right\}\leq\delta\ , \tag{5.52}$$

The next section will discuss a futher generalization of this approach, known as generic chaining. However, before doing that, we turn our attention to the non-trivial problem of applying this rather abstract bound.

To apply this bound, we need to select a "prior" $\alpha(j)$ such that $\delta(j)=\delta\alpha(j)$, and we need to know more about the sizes of the link sets $\mathcal{B}_j$. We also need to be able to evaluate or upper bound the infinite sum in the expression above.

We now demonstrate a useful technique for obtaining an upper bound on sums of this form: essentially, we replace the infinite sum by an appropriate integral.

Suppose a function ϕ is decreasing on some interval $(0, c) \subset \mathbb{R}$. Let $K = \lfloor -\log_2 c \rfloor$, so that $2^{-K} \leq c \leq 2^{-K+1}$. Now we consider a step function approximating ϕ from below. Specifically define $\phi' : (0, 2^{-K}) \to \mathbb{R}$ by $\phi'(v) = \phi(2^{-j})$ when $v \in [2^{-j-1}, 2^{-j})$, for $j \geq K$.

Thus, ϕ' approximates ϕ on $[2^{-j-1}, 2^{-j})$ by the value of ϕ at the right-most side of the interval. Since ϕ is decreasing, ϕ' never exceeds ϕ. Thus

$$\int_0^{v'} \phi'(v)\, dv \leq \int_0^{v'} \phi(v)\, dv$$

for all $v' \leq 2^{-K}$. However, since ϕ' is a step function, the left-hand integral can be replaced by an infinite sum. Assuming $v' = 2^{-i}$ for some $i \in \mathbb{Z}$, we have

$$\begin{aligned} \int_0^{v'} \phi'(v)\, dv &= \sum_{j=i}^{\infty} (2^{-j} - 2^{-j-1}) \phi(2^{-j}) \\ &= \sum_{j=i}^{\infty} \phi(2^{-j}) 2^{-j-1} \ . \end{aligned}$$

We apply this result by finding a ϕ such that the infinite series in (5.52) corresponds to integrating/summing the corresponding ϕ'. Then we can upper bound the sum by the integral of ϕ. If we write

$$\Upsilon(j) = 3\sqrt{\frac{2}{m}}\sqrt{\ln|\mathcal{B}_j| - \ln\delta(j)} \ ,$$

we would like it if we could have $\phi(2^{-j}) = \Upsilon(j)$, so that the upper bound we would like to evaluate is simply

$$\sum_{j=1}^{\infty} 2^{-j-1} \phi(2^{-j}) \ ,$$

yielding an upper bound of

$$\int_0^{\frac{1}{2}} \phi(v)\, dv \ .$$

In order to do this, we need ϕ to be a decreasing function passing through $(2^{-j}, \Upsilon(j))$ for all $j \in [2 : \infty)$. Such a ϕ can be found if we know that $\Upsilon(j)$ is an increasing function of j. Thus, for a specific construction of the $\mathcal{W}_j$, one must choose $\delta(j)$ such that $\frac{|\mathcal{B}_j|}{\delta(j)}$ is increasing in j. For most reasonable constructions, $|\mathcal{B}_j|$ is increasing in j, and one will typically work with a decreasing sequence for $\delta(j)$, so that this condition will hold in almost any reasonable situation.

It is clear that for any $\phi^\star \geq \phi$, $\int_0^{\frac{1}{2}} \phi^\star(v)\, dv$ is also an upper bound on the sum under consideration. For any construction of the $\mathcal{W}_j$, we know that

$$|\mathcal{W}_j| \leq \mathcal{M}_{2,S}(2^{-j}, \mathcal{W}) \ ,$$

so that

$$|\mathcal{B}_j| \leq (\mathcal{M}_{2,S}(2^{-j}, \mathcal{W}))^2 \ .$$

This upper bound on $|\mathcal{B}_j|$ is convenient since it provides a clear extension to values of s which are not powers of 2. If one constructs each $\mathcal{W}_j$ as a refinement of $\mathcal{W}_{j-1}$ (i.e. by adding points to $\mathcal{W}_{j-1}$), it can be shown that $|\mathcal{B}_j| = |\mathcal{W}_j|$ (since each element of $\mathcal{W}_j$ has a unique element of $\mathcal{W}_{j-1}$ close enough to form a link). In this case, we have

$$|\mathcal{B}_j| \leq \mathcal{M}_{2,S}(2^{-j}, \mathcal{W}) \ ,$$

a considerably tighter bound.

In conjunction with this, one typically selects a prior which can easily be extended from powers of 2. Since we are dealing with a countably infinite sequence, the prior must be arbitrarily small for infinitely many j. Furthermore, we would like $\delta(j)$ to be decreasing. However, using a geometric sequence for the "prior" α will typically yield an undesirably rapid decrease of the $\delta(j)$ to zero. Popular sequences for α in practice are generally based on variations of $\alpha(j) = \frac{1}{j^2}$. The corresponding series sums to $\frac{\pi^2}{6}$. Some references reduce this to less than 1 by instead using $\alpha(j) = \frac{1}{(j+1)^2}$, but this wastes confidence unnecessarily. Another option is scaling α to ensure summation to one, i.e. $\alpha(j) = \frac{6}{\pi^2 j^2}$. We recommend an option similar to scaling, but without the extraneous constants: if one employs $\alpha(j) = \frac{1}{j(j+1)}$, we have that the series sums to one directly. We consider this choice of $\alpha(j)$ in what follows.

Let us now consider the function

$$\alpha'(v) = \frac{1}{(-\log_2 v)(-\log_2 v + 1)} = \frac{(\ln 2)^2}{(\ln v)\left(\ln \frac{v}{2}\right)} \quad .$$

Then $\alpha'(2^{-j}) = \alpha(j)$, and α' is increasing for $v > 0$.

With this choice of $\delta(j)$, and the $\mathcal{W}_j$ constructed by refining $\mathcal{W}_{j-1}$, we have that

$$3\sqrt{\frac{2}{m}}\sqrt{\ln \frac{\mathcal{M}_{2,S}(v, \mathcal{W})}{\delta\alpha'(v)}}$$

is a potential choice of $\phi^\star$, an upper bound on ϕ.

With this choice of $\phi^\star$, we have

$$\mathbb{P}_{\zeta\sim(\mathrm{Unif}\{-1,1\}^m)}\left\{\sup_{w\in\mathcal{W}}\left[\frac{1}{m}\sum_{i=1}^{m}\zeta_i w(x_i)\right] > \int_0^{\frac{1}{2}} \phi^\star(v)\, dv|S\right\} \le \delta \quad .$$

The bound obtained here is for a fixed m-sample S. In order to apply the random subsample lemma, we need to take an expectation w.r.t. the possible samples S. However, the form above is not convenient for that. Consider any $\epsilon > 0$. We note that

$$\int_0^{\frac{1}{2}} \phi^\star(v)\, dv$$

is a function of δ and S, say $\mathit{\Pi}(\delta, S)$. Let $\delta(\epsilon, S)$ be such that

$$\mathit{\Pi}(\delta(\epsilon, S), S) = \epsilon \quad .$$

Then

$$\mathbb{P}_{\zeta\sim(\mathrm{Unif}\{-1,1\}^m)}\left\{\sup_{w\in\mathcal{W}}\left[\frac{1}{m}\sum_{i=1}^{m}\zeta_i w(x_i)\right] > \epsilon|S\right\} \le \delta(\epsilon, S) \quad .$$

Taking expectations w.r.t. S on both sides yields

$$\mathbb{P}_{S\sim D^m,\zeta\sim\mathrm{Unif}(\{-1,1\}^m)}\left\{\sup_{w\in\mathcal{W}}\left[\frac{1}{m}\sum_{i=1}^{m}\zeta_i w(x_i)\right] > \epsilon\right\} \le \mathbb{E}_{S\sim D^m}\, \delta(\epsilon, S) \quad .$$

It is common to replace the packing numbers by covering numbers in the expression above, using

$$\mathcal{M}_{2,S}(v, \mathcal{W}) \le \mathcal{N}_{2,S}\left(\frac{v}{2}, \mathcal{W}\right) \quad .$$

In summary, in this section we defined a sequence of approximating subclasses of $\mathcal{W}$, and expanded each decision rule w into a telescoping series based on its projections on consecutive subclasses. By construction, the successive projections for each decision rule w are close together (because they are both close to w). Because the process behaviour is in someway related to the structure of the index set, we could obtain a probabilistic bound on the difference between the process values at these projections. We employed the Occam's razor method over each link set, and for each successive approximation (non-uniformly). In this manner, we obtained a bound on the Rademacher penalty. Finally, we took the expectation over all m-samples S to obtain a bound which can be combined with the random subsample lemma and a symmetrization lemma. This approach provides a bound on the regular deviation between empirical and true risk for all of the decision rules in $\mathcal{W}$.

5.7.1 Generic chaining

Generic chaining is a generalization of chaining originally presented using the concept of *majorizing measures*, which first arose in Gaussian process theory (Talagrand, 1996a). Essentially, generic chaining replaces the uniform application of the Occam's razor method in the regular chaining method over the various link sets $\mathcal{B}_j$ by a non-uniform application of the Occam's razor method.

To motivate this, we consider the elements of a link set $\mathcal{B}_j$. For each pair $(w_1, w_2) \in \mathcal{B}_j$, define the set

$$A_j((w_1, w_2)) = \{w \in \mathcal{W} : (\mathscr{P}_j(w), \mathscr{P}_{j-1}(w)) = (w_1, w_2)\} \quad .$$

So w is an element of $A_j((w_1, w_2))$ if the element of the j-th link set corresponding to w is (w_1, w_2). It follows that the collection of sets

$$\mathcal{D}_j = \{A_j((w_1, w_2)) : (w_1, w_2) \in \mathcal{B}_j\}$$

forms a partition of $\mathcal{W}$. Effectively, the standard chaining bound assigns a uniform "prior" ($\frac{1}{|\mathcal{B}_j|}$) to all elements of the link set $\mathcal{B}_j$. Generic chaining attempts to improve this "prior" by relating the "prior" probability of a link in $\mathcal{B}_j$ to the "relative size" of the partition element associated with that link.

We first need to formalize the concept of the "relative size" of an element A of a partition of some set R. First, suppose R is a bounded subset of $\mathbb{R}^n$. A natural concept of relative size in this case is the ratio of the volume of A to that of R, or equivalently the ratio of the Lebesgue measure of these sets.[118] If we wish to work in more abstract spaces than $\mathbb{R}^n$, and possibly unbounded sets, the concept of the ratio of the measure of the sets is more useful. However, the choice of measure is typically arbitrary, and can also be seen as similar to that of the "prior" of the Occam's razor method — in fact, the measure we select here will translate directly into a "prior" for each partition $\mathcal{D}_j$. We shall assume the measure we use is a probability measure. This is a minor restriction, since we are dealing with ratios of measures, so that any bounded measure can be replaced by a probability measure with identical results.

Consider a probability measure τ on $\mathcal{W}$. Then the relative size of A (with respect to $\mathcal{W}$) is $\frac{\tau(A)}{\tau(\mathcal{W})} = \tau(A)$. Since $\mathcal{D}_j$ forms a partition of $\mathcal{W}$, the sum of the relative sizes of each partition element will sum to one. We use these relative sizes to assign a "prior" to the elements of $\mathcal{B}_j$: specifically, $\alpha_j((w_1, w_2)) = \tau(A_j((w_1, w_2)))$. Once again, note that τ is merely a technical prior, which may, but need not, reflect prior belief and information on the appropriateness of the elements in $\mathcal{W}$. Furthermore, if we desire, for each choice of j the corresponding α_j can be defined with reference to a different measure τ_j.

We now consider the application of this idea to modify the chaining bound. A scrutiny of the reasoning employed for the chaining bound shows that one can obtain a result for this modified "prior" over $\mathcal{B}_j$ simply by replacing the old "prior" value $\frac{1}{|\mathcal{B}_j|}$ by the new "prior" value

$$\alpha_j((w_1, w_2)) = \tau_j(A_j((w_1, w_2))) \ ,$$

and noting that the appropriate link (w_1, w_2) for a given link number j and

[118] Assuming we have equipped $\mathbb{R}^n$ with the Lebesgue measure.

decision rule w is $(\mathscr{P}_j(w), \mathscr{P}_{j-1}(w))$. This yields the generic chaining bound

$$\mathbb{P}_{\zeta\sim(\text{Unif}\{-1,1\})^m}\left\{\begin{array}{l} \sup_{w\in\mathcal{W}}\left[\frac{1}{m}\sum_{i=1}^m \zeta_i w(x_i)\right] \\ \quad > 3\sqrt{\frac{2}{m}}\sum_{j=1}^\infty 2^{-j}\sqrt{\ln\frac{1}{\delta(j)\tau_j(A_j((\mathscr{P}_j(f),\mathscr{P}_{j-1}(f))))}} \end{array} |S\right\} \leq \delta \ .$$

5.8 Dimension measures of complexity

In practice, expectations involving covering numbers are not obtained exactly. Even for a known distribution D, this is a very difficult problem. In general, we do not know D, and the expected covering number may be as large as the worst covering number obtained w.r.t. any specific sample. Thus the classical approach to this problem is to bound the expectation involving the covering number from above by replacing the covering number by the supremum of the covering number over all possible samples of the relevant length. This supremum is then an upper bound for the expected covering number for all distributions D. We shall write $\mathcal{N}_{p,m}(\gamma, \mathcal{W})$ for this supremum, i.e.

$$\mathcal{N}_{p,m}(\gamma, \mathcal{W}) = \sup_{Q\in\mathcal{Z}^m} \mathcal{N}_{p,Q}(\gamma, \mathcal{W}) \ .$$

We extend the notation for external covering and packing numbers similarly. In addition, we shall write $\mathcal{N}_{\mathcal{W}}(m)$ for

$$\sup_{Q\in\mathcal{Z}^m} |Q_{\mathcal{W}}| \ .$$

$\mathcal{N}_{\mathcal{W}}(m)$ is known as the *shatter coefficient* of $\mathcal{W}$ for sample size m.

Even the calculation of these shatter coefficients is very difficult or impossible in practice, so bounds on them are typically employed. Many of these bounds rely on quantities called dimensions. Before focusing on methods for bounding covering numbers and their suprema, we shall spend some time introducing the three most important[119] dimension quantities.

[119] A wide variety of other related dimension quantities have been studied, but these will be sufficient for presenting and discussing our results.

5.8.1 VC dimension

The first bounds on covering numbers were based on the concept of the Vapnik-Chervonenkis (VC) dimension. To explain VC dimension, we begin by introducing the auxiliary concept of *shattering*. Note that for a small m, a zero-one loss function, and a reasonably-sized decision class $\mathcal{W}$, it is highly likely that there exists a sample Q containing n *distinct* points such that $Q_{\mathcal{W}} = \{0,1\}^n$, i.e. that for any sequence of n decisions in $\{0,1\}$ there is some decision rule in $\mathcal{W}$ yielding that sequence of decisions on Q. When this is the case, we say that $\mathcal{W}$ shatters Q. Clearly, when $\mathcal{W}$ shatters an n-sample Q, we have $|Q_{\mathcal{W}}| = \mathcal{N}_{\mathcal{W}}(m) = 2^n$.

Consider the 1-concept class corresponding to $\mathcal{W}$, $\mathcal{C}_1(\mathcal{W})$. Consideration of the definitions show that $\mathcal{W}$ shatters an m-sample Q of distinct points if for each subsample Q_0 of Q, there is a $c \in \mathcal{C}_1(\mathcal{W})$ such that $Q_0 \subseteq c$ and $Q \setminus Q_0 \cap c = \emptyset$. The same result holds for $\mathcal{C}_0(\mathcal{W})$. In general, we can extend our definition of shattering to arbitrary classes of sets: we say a class of sets $\mathscr{C}$ in a space $\mathcal{E}$ shatters a set $R \subseteq \mathcal{E}$ if, for each subset R_0 of R, there is a $c \in \mathscr{C}$ such that $R_0 \subseteq c$ and $R \setminus R_0 \cap c = \emptyset$. An alternative formulation: $\mathscr{C}$ shatters a finite set R if

$$\{c \cap R : c \in \mathscr{C}\} = 2^R \ .$$

If we write

$$\mathcal{N}_{\mathscr{C}}(R) = |\,\{c \cap R : c \in \mathscr{C}\}\,| \ ,$$

we can restate this condition as

$$\mathcal{N}_{\mathscr{C}}(R) = 2^{|R|} \ .$$

Furthermore, we can see that

$$\mathcal{N}_{\mathcal{C}_1(\mathcal{W})}(Q) = |Q_{\mathcal{W}}| \ .$$

We can also extend the concept of shatter coefficient to classes of sets using this approach, motivating our choice of the symbol $\mathcal{N}$: the shatter coefficient of a class of sets is defined by

$$\mathcal{N}_{\mathscr{C}}(n) = \sup\{\mathcal{N}_{\mathscr{C}}(R) : |R| = n, R \subseteq \mathcal{E}\} \ .$$

Example 5.11. Consider the collection $\mathscr{C}$ of intervals in $\mathbb{R}$ of the form $c_s = (-\infty, s]$. Consider any point $s_0 \in \mathbb{R}$. If $s \geq s_0$, $s_0 \in c_s$ and $c_s \cap \{s_0\} = \{s_0\}$. If $s < s_0$, we have $c_s \cap \{s_0\} = \emptyset$. Since the power set of $\{s_0\}$ has two elements, we see that $\mathscr{C}$ shatters $\{s_0\}$.

Furthermore, it should be easy to see that given any two points $s_0, s_1 \in \mathbb{R}$ with $s_1 > s_0$, no interval in $\mathscr{C}$ intersects with s_1 but not with s_0. Thus, no two points can be shattered by $\mathscr{C}$.

Given n points, $s_0 < s_1 < \cdots < s_{n-1}$, the upper endpoint of such an interval must lie below s_0, between s_{i-1} and s_i for $i = 1, \cdots, n-1$, or above s_{n-1}, yielding $n+1$ classifications. We thus have

$$\mathcal{N}_{\mathscr{C}}(n) = n + 1 = \binom{n}{0} + \binom{n}{1} \ .$$

□

Example 5.12. Let us expand $\mathscr{C}$ to allow all intervals in $\mathbb{R}$ with at least one infinite endpoint. These intervals are also known as *halfspaces* in $\mathbb{R}$. Consider $s_0, s_1 \in \mathbb{R}$ with $s_1 > s_0$.

- For $s < s_0$, $(-\infty, s] \cap \{s_0, s_1\} = \emptyset$.
- For $s_0 < s < s_1$, $(-\infty, s] \cap \{s_0, s_1\} = \{s_0\}$.
- For $s_0 < s < s_1$, $[s, \infty) \cap \{s_0, s_1\} = \{s_1\}$.
- For $s > s_1$, $(-\infty, s] \cap \{s_0, s_1\} = \{s_0, s_1\}$.

Thus $\mathscr{C}$ shatters $\{s_0, s_1\}$.

However, for any three points $\{s_0, s_1, s_2\}$, it is clear we can not have

$$c \cap \{s_0, s_1, s_2\} = \{s_0, s_2\}$$

for any $c \in \mathscr{C}$, since any interval containing s_0 and s_2 has to contain t_1.

In this case, it can be shown (Devroye et al., 1996, Theorem 13.8) that

$$\mathcal{N}_{\mathscr{C}}(n) = \frac{n(n+1)}{2} + 1 = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} \ .$$

□

VC dimension is useful due to a lemma known as the Vapnik-Chervonenkis-Sauer-Shelah (VCSS) lemma. This effectively states that once n becomes large enough that no n-sample can be shattered by $\mathcal{W}$, $\mathcal{N}_{\mathcal{W}}(n)$ stops growing exponentially in n, and starts growing no faster than a polynomial in n. The n at which this change occurs, known as the VC dimension of $\mathcal{W}$, is also the degree of this polynomial. In some cases, there is no n large enough — $\mathcal{W}$ can shatter some sequence of arbitrary length. In this (not so uncommon) case, we say $\mathcal{W}$ has infinite VC dimension.

Definition 5.4 (VC dimension). The VC dimension of a zero-one decision class $\mathcal{W}$, VC($\mathcal{W}$), is the size of the largest sample of distinct points which are shattered by $\mathcal{W}$, and infinity if no such largest set exists.

Equivalently, the VC dimension of $\mathcal{W}$ is the VC-dimension of $\mathcal{C}_1(\mathcal{W})$ (or $\mathcal{C}_0(\mathcal{W})$), where the VC-dimension of a class of sets $\mathscr{C}$ is the size of the largest set of points R shattered by $\mathscr{C}$.

If the VC dimension of a class is finite, we call the class a VC class.

Example 5.13. From Example 5.11 it is clear that the VC dimension of the set of intervals on $\mathbb{R}$ with lower endpoint $-\infty$ is one.

Similarly, from Example 5.12, the VC dimension of the set of intervals on $\mathbb{R}$ and of the set of halfspaces on $\mathbb{R}$ is two. □

Example 5.14. Consider the class $\mathscr{C}$ of all (Borel) closed sets in $\mathbb{R}^N$, and any finite set $R = \{\eta_1, \cdots, \eta_n\}$. For each η_i, we can construct a closed ball B_i in $\mathbb{R}^N$ with centre η_i not containing any other elements of R.

Consider any $R_0 \subseteq R$. It is clear that

$$R' = \bigcup_{\{i:\eta_i \in R_0\}} B_i$$

is (Borel) closed, and $R' \cap R = R_0$. Thus $\mathscr{C}$ shatters any finite set R, so that the VC dimension of $\mathscr{C}$ is infinite. □

The VCSS lemma was independently discovered (in slightly varying strengths) by Vapnik and Chervonenkis (Vapnik and Chervonenkis, 1971), by Sauer (Sauer, 1972), and by Shelah (Shelah, 1972), their initials resulting in the current name.[120]

[120] The intermediate inequalities presented here are from Blumer et al. (1989).

Theorem 5.24 (VCSS lemma). *Let $\varrho < n$. Then*

$$\sum_{i=0}^{\varrho} \binom{n}{i} \leq \frac{2n^{\varrho}}{\varrho!} \leq \sqrt{\frac{2}{\pi\varrho}} \left(\frac{en}{\varrho}\right)^{\varrho} \leq \left(\frac{en}{\varrho}\right)^{\varrho} \ .$$

For any decision class of indicator functions $\mathcal{W}$, we have that $\mathcal{N}_{\mathcal{W}}(n) = 2^n$ for $1 \leq n \leq \mathrm{VC}(\mathcal{W})$, and, for $n > \mathrm{VC}(\mathcal{W})$,

$$\mathcal{N}_{\mathcal{W}}(n) \leq \sum_{i=0}^{\mathrm{VC}(\mathcal{W})} \binom{n}{i} \ ,$$

which can in turn be bounded by the results above.

A formulation for classes of sets is obtained by replacing $\mathcal{W}$ by $\mathscr{C}$, and $\mathcal{N}_{\mathcal{W}}(m)$ by $\mathcal{N}_{\mathscr{C}}(m)$.

We note that the first inequality in the lemma is tight, since equality holds in Examples 5.11 and 5.12.

Proofs for this result may be found in, amongst others, Vidyasagar (2002, Section 4.2.1) and Vapnik (1998, Lemma 4.4) — the conversion from the combinatorial form to the polynomial expression is made by using Stirling's approximation (Blumer et al., 1989). A number of other related and useful results based on other bounds on the combinatorial form are presented in Devroye et al. (1996, Section 13.1) and Vapnik (1998, Section 4.10).

It should be clear that this lemma effectively provides an upper bound on the covering number-based bounds employing $Q_{\mathcal{W}}$ in previous sections. As one example, applying the VCSS lemma to the bound in (5.18), we obtain that (for $2m > \mathrm{VC}(\mathcal{W})$),

$$\mathbb{P}_{S\sim D^m} \left\{ \begin{array}{l} \sup_{w\in\mathcal{W}}[e_D(w) - e_S(w)] \\ \quad > \frac{1+\sqrt{(m+1)\left[\mathrm{VC}(\mathcal{W}) \ln \frac{4em}{\mathrm{VC}(\mathcal{W})} - \ln\delta\right]+1}}{m} \end{array} \right\} < \delta \tag{5.53}$$

since for any distribution D, we have

$$\begin{aligned} \mathbb{E}_{Q\sim D^{2m}} |Q_{\mathcal{W}}| &\leq \mathcal{N}_{\mathcal{W}}(2m) \\ &\leq \left(\frac{2em}{\mathrm{VC}(\mathcal{W})}\right)^{\mathrm{VC}(\mathcal{W})} \end{aligned}$$

for $2m$ exceeding $\mathrm{VC}(\mathcal{W})$.

The VCSS lemma represents the first generic approach to obtaining non-trivial distribution-independent bounds (although very large samples are needed in practice): without the VCSS lemma, bounding the supremum of the covering number by the naïve 2^{2m} yields trivial confidence intervals in all cases.

5.8.2 Pseudodimension

The approach above employing the VC dimension is clearly limited to the bounds stated in terms of expectations involving $Q_{\mathcal{W}}$. This works well for decision classes of indicator functions, so we can also obtain bounds for real functions by employing bounds based on the thresholded loss class, along with applying the VCSS lemma to bound $\mathcal{N}_{\mathcal{W}_t}(m)$, the supremum of $Q_{\mathcal{W}_t}$ for m-samples Q on the thresholded loss class. It turns out that a fascinating connection exists between $\mathcal{N}_{\mathcal{W}_t}$ and the packing numbers of $\mathcal{W}$ when the domain of the functions in $\mathcal{W}$ is $\mathbb{R}^N$ for some N. This connection is quantified by the following result relating the expected γ-packing number of $\mathcal{W}$ to $\mathcal{N}_{\mathcal{W}_t}$. The result is a refinement of a result by Pollard, based on work by Dudley:

Theorem 5.25 (Theorem 29.3 of Devroye et al., 1996). *Let the domain of the elements of $\mathcal{W}$ be $\mathbb{R}^N$ for some $N \in \mathbb{N}$. For every γ and every distribution Q,*

$$\mathcal{M}_{1,Q}(\gamma, \mathcal{W}) \leq \mathcal{N}_{\mathcal{W}_t}\left(\left\lceil \frac{1}{\gamma} \log_2 \frac{e\gamma \mathcal{M}_{1,Q}^2(\gamma, \mathcal{W})}{2} \right\rceil\right) .$$

It follows from this theorem together with the VCSS lemma that we can bound the packing numbers (and hence the covering numbers) from above by

$$2\left(\frac{2e}{\gamma} \ln \frac{2e}{\gamma}\right)^{\mathrm{VC}(\mathcal{W}_t)} , \tag{5.54}$$

which can be seen as a more widely applicable companion to the VCSS lemma. This is Theorem 6 in Haussler (1992).

We thus see that the VC dimension of the thresholded decision class can be used to obtain bounds on results employing general covering and packing numbers (at least those employing a L^1 norm). The VC dimension of the thresholded decision class $\mathcal{W}_t$ is called (Dudley's) *pseudodimension* of the decision class $\mathcal{W}$[121], and is denoted by

$$\operatorname{pdim}(\mathcal{W}) = \mathrm{VC}(\mathcal{W}_t) \ .$$

Considering these results, the relationship between covering numbers and packing numbers, and the bound of (5.37) with $p = 1$, we obtain

$$\begin{aligned} & \mathbb{P}_{S \sim D^m} \left\{ \sup_{w \in \mathcal{W}} [r_D(w) - r_S(w)] > \epsilon \right\} \\ \leq \ & \frac{2}{\beta} \mathbb{E}_{Q \sim D^m} \bar{\mathcal{N}}_{1,Q}(\gamma, \mathcal{W}) \exp \left(\frac{-m \left(\frac{\epsilon - \alpha(m,\beta)}{2} - \gamma \right)^2}{2} \right) \\ \leq \ & \frac{2}{\beta} \mathbb{E}_{Q \sim D^m} \mathcal{M}_{1,Q}(\gamma, \mathcal{W}) \exp \left(\frac{-m \left(\frac{\epsilon - \alpha(m,\beta)}{2} - \gamma \right)^2}{2} \right) \\ \leq \ & \frac{4}{\beta} \left(\frac{2e}{\gamma} \ln \frac{2e}{\gamma} \right)^{\operatorname{pdim}(\mathcal{W})} \exp \left(\frac{-m \left(\frac{\epsilon - \alpha(m,\beta)}{2} - \gamma \right)^2}{2} \right) \ . \end{aligned}$$

Haussler also provided a packing number bound w.r.t. $d_{1,Q}$ for zero-one loss functions which is tighter than the general-purpose one above. Specifically, in Haussler (1991, Theorem 1), he shows that for any Q,

$$\begin{aligned} \mathcal{M}_{1,Q}(\gamma, \mathcal{W}) \ & \leq \ e(\mathrm{VC}(\mathcal{W}) + 1) \left(\frac{2e(m+1)}{m\gamma + 2\,\mathrm{VC}(\mathcal{W}) + 2} \right)^{\mathrm{VC}(\mathcal{W})} \\ & \leq \ e(\mathrm{VC}(\mathcal{W}) + 1) \left(\frac{2e}{\gamma} \right)^{\mathrm{VC}(\mathcal{W})} \ , \end{aligned} \tag{5.55}$$

provided that $m\gamma \in [1 : m]$. Note that this result is in terms of the VC dimension, not the pseudodimension, of $\mathcal{W}$. This result is useful for bounds on zero-one loss functions where γ is large enough that $\bar{\mathcal{N}}_{1,Q}(\gamma, \mathcal{W}) < |Q_{\mathcal{W}}|$. An example is if we would like to employ the bound of (5.16) directly without relying on the bound of (5.18).

[121] Other names for this quantity include the combinatorial or Pollard dimension. An equivalent formulation can be found in Anthony (1994, Definition 12.1)

5.8.3 Fat-shattering dimension

Further refinements were obtained in Alon et al. (1993), Cesa-Bianchi and Haussler (1998), Haussler and Long (1995). The authors generalized the VCSS Lemma, employing alternative concepts of *strong shattering* and the *strong dimension* of an hypothesis class. The formulation of the final result, however, can be provided in terms of the more widely used *fat-shattering dimension*. Unlike the VC dimension and the pseudodimension, which are numbers, the fat-shattering dimension is a function, encoding the complexity of a function class at various resolutions. In fact, the fat-shattering dimension can be seen as a three-level generalization of the VC dimension.

We say that a decision class $\mathcal{W}$ of real-valued functions γ-shatters a set $R \subseteq \mathcal{X}$ if there is a real-valued function ϕ such that for any $R_0 \subseteq R$ there is a $w \in \mathcal{W}$ such that $w(x) \geq \phi(x) + \gamma$ for $x \in R_0$ and $w(x) \leq \phi(x) - \gamma$ for $x \in R \setminus R_0$. The function mapping $\gamma > 0$ to the size of the largest set which $\mathcal{W}$ γ-shatters (or infinity), is called the fat-shattering dimension (or P_γ dimension — see Alon et al., 1993) of $\mathcal{W}$, and is denoted by $\mathrm{fat}_{\mathcal{W}}$. We refer to $\mathrm{fat}_{\mathcal{W}}(\gamma)$ as the fat-shattering dimension[122] of $\mathcal{W}$ at scale γ. Clearly the fat-shattering dimension is decreasing on $(0, \infty)$. If the fat-shattering dimension of $\mathcal{W}$ is finite at all scales $\gamma > 0$, we call $\mathcal{W}$ a *uniform Glivenko-Cantelli (GC) class*.

If we restrict ϕ to be a constant function, we speak of *uniform* γ-shattering, and the uniform fat-shattering dimension (or V_γ-dimension — see Alon et al., 1993), $\mathrm{fatV}_{\mathcal{W}}$. It is interesting to note that all bounds formulated in terms of the fat-shattering dimension can be reformulated using the uniform fat-shattering dimension — see Alon et al. (1993, Lemma 2.2). The limit of the fat-shattering dimension of a function class as γ tends to 0 can be seen to

[122] Introduced by Kearns and Schapire (Kearns and Schapire, 1994), although they refer to γ as the *width* of shattering. Interestingly, their work used the concept to derive *lower bounds* on sample size for learning.

be the pseudodimension of the class[123]:

$$\mathrm{pdim}(\mathcal{W}) = \lim_{\gamma \to 0^+} \mathrm{fat}_{\mathcal{W}}(\gamma) \enspace .$$

Thus, any class with finite pseudodimension is a uniform GC class.

Since the fat-shattering dimension is a decreasing function, we also have that $\mathrm{pdim}(\mathcal{W}) \geq \mathrm{fat}_{\mathcal{W}}(\gamma)$ for all $\gamma > 0$, and that if $\mathrm{pdim}(\mathcal{W})$ is finite, there is some $\gamma_0 > 0$ such that for $0 < \gamma < \gamma_0$, $\mathrm{pdim}(\mathcal{W}) = \mathrm{fat}_{\mathcal{W}}(\gamma)$.

Earlier, we discussed the VC-dimension: it can be viewed as a measure of the ability of a class of functions to classify randomly labelled inputs. If one considers the definition of the (uniform) fat-shattering dimension in this light, we can interpret it as a measure of the ability of a class of functions to "classify" randomly labelled points with a kind of *safety buffer* (corresponding to γ) when thresholded. Informally, this safety buffer is the margin, and provides extra confidence in the classification. Thus, we shall see that traditional classification bounds will employ the VC dimension, while margin bounds will typically employ the fat-shattering dimension at some scale.

Improvements to the VCSS lemma were based on generalizing the lemma to loss functions with a finite range. The original work in Alon et al. (1993) employed a generalization of shattering known as strong shattering, leading to the concept of strong dimension of the function class. It turns out that the packing number of $\mathcal{W}$ w.r.t. $d_{\infty,Q}$ can be related to the packing number of a discretized version of $\mathcal{W}$. This discretized version has a finite range (related to the resolution desired for the packing number of $\mathcal{W}$), and thus a generalized VCSS lemma can be applied to bound its packing numbers.

A bound on the packing numbers of $\mathcal{W}$ is stated in Lemma 3.4 of Alon et al. (1993), with a slight refinement implicit in the previous proof presented in Lemma 4 of Anthony and Bartlett (1994).

Theorem 5.26 (Packing number bound from fat-shattering dimension).

[123] The corresponding limit of the uniform fat-shattering dimension is known as the Vapnik dimension (Guermeur, 2004).

Let $0 < \gamma < 1$. *Then*

$$\begin{aligned}\mathcal{M}_{\infty,n}(\gamma,\mathcal{W}) &\leq 2\left(\frac{4n}{\gamma^2}\right)^{\left\lceil \varrho\log_2\frac{4n}{\gamma}-\log_2\varrho!\right\rceil}\\ &\leq 2\left(\frac{4n}{\gamma^2}\right)^{\left\lceil \varrho\log_2\frac{2en}{\varrho\gamma}\right\rceil},\end{aligned}$$

where $\varrho = \mathrm{fat}_{\mathcal{W}}(\frac{\gamma}{4})$ *satisfies* $\varrho < m$.[124]

In addition, define

$$j = \sum_{i=1}^{\varrho}\binom{n}{i}\left\lceil\frac{2}{\gamma}\right\rceil^{i} \quad .$$

Then, for $n \geq \log_2 j + 1$,

$$\mathcal{M}_{\infty,n}(\gamma,\mathcal{W}) < 2\left(n\left\lceil\frac{2}{\gamma}\right\rceil^{2}\right)^{\log_2 j} \quad .$$

The first form is more often used, although the second form is slightly stronger, and is analogous to the combinatorial bound in the VCSS Lemma. These results allow us to use the fat-shattering dimension as an analog of the VC dimension for general loss functions, as well as a measure of complexity when employing margin bounds. As a result, the fat-shattering dimension of a class at a suitable scale is sometimes referred to as the *effective* VC dimension of the class, e.g. Cristianini and Shawe-Taylor (2000, Section 4.3).

Applying the bound above to the bound on regular deviation of risk in

[124] The original statement of this Theorem did not include this condition or the ceiling in the final exponent; neither did it include the central inequality. The condition is necessary because the proof effectively employs the VCSS lemma, which also allows the introduction of the slightly tighter bound. The ceiling seems to have been dropped for convenience in the original proof.

(5.37), we obtain that

$$\begin{aligned}
&\mathbb{P}_{S\sim D^m}\left\{\sup_{w\in\mathcal{W}}\left[r_D(w)-r_S(w)\right]>\epsilon\right\}\\
\leq\ & \frac{2}{\beta}\,\mathbb{E}_{Q\sim D^m}\,\bar{\mathcal{N}}_{p,Q}(\gamma,\mathcal{W})\exp\left(\frac{-m\left(\frac{\epsilon-\alpha(m,\beta)}{2}-\gamma\right)^2}{2}\right)\\
\leq\ & \frac{2}{\beta}\mathcal{M}_{\infty,m}(\gamma,\mathcal{W})\exp\left(\frac{-m\left(\frac{\epsilon-\alpha(m,\beta)}{2}-\gamma\right)^2}{2}\right)\\
\leq\ & \frac{4}{\beta}\left(\frac{4m}{\gamma^2}\right)^{\varrho\log_2\frac{2em}{\varrho\gamma}}\exp\left(\frac{-m\left(\frac{\epsilon-\alpha(m,\beta)}{2}-\gamma\right)^2}{2}\right)\quad,
\end{aligned}$$

where $\varrho = \mathrm{fat}_{\mathcal{W}}(\frac{\gamma}{4})$.

Haussler and Long (1995) provide an alternative result by directly generalizing the VCSS lemma to classes of functions with finite range for various generalizations of the concept of shattering. This led to the following result, which bounds $\mathcal{N}_{\infty,n}(\gamma,\mathcal{W})$:

Theorem 5.27 (Covering number bound from pseudodimension).

$$\mathcal{N}_{\infty,n}(\gamma,\mathcal{W})\leq\sum_{i=1}^{\varrho}\binom{n}{i}\left\lfloor\frac{1}{2\gamma}\right\rfloor^i\quad,$$

where $\varrho = \mathrm{pdim}(\mathcal{W})$.

Note that for $n\geq\varrho$ and $\gamma\leq\frac{1}{2}$ we can relax the right hand side of the above bound to

$$\left(\frac{en}{\varrho}\right)^{\varrho}(2\gamma)^{-\varrho}=\left(\frac{en}{2\gamma\varrho}\right)^{\varrho}\quad,$$

by employing the same bound used to relax the VCSS lemma. This result looks substantially tighter, but employs the pseudodimension, rather than the fat-shattering dimension. In some cases, the pseudodimension is not finite, even though the fat-shattering dimension is finite at every scale $\gamma>0$.

Employing the results above for bounding the covering numbers in terms of complexity dimensions are confounded by two complications. The first is that the results only help when the appropriate dimension of the decision class is known. Although these dimensions have been determined for a number of popular decision classes, finding these values for a decision class in general is a very difficult problem, if a way to do it can be found at all. Two points are important in this regard. First, adding functions to a decision class can only increase its dimension. Thus, one often obtains the dimension of a slightly expanded decision class for which finding the dimension is an easier task. Second is a related issue: often bounds on the dimension are used instead of the dimension itself, when finding the dimension directly is not feasible.

The second issue leads us to the realization that we need further alternatives besides these bounds: many decision classes have infinite VC, pseudo-, and/or fat-shattering dimension. This means that bounds employing them never become non-trivial, regardless of the sample size.

5.9 Bounding covering numbers

Having introduced some dimension quantities in the previous section, and illustrated why they are helpful, we now turn to obtaining bounds on covering numbers.

In general, finding a dimension quantitiy corresponding to a complex decision class directly is quite difficult. However, results have been found for a number of simple classes, and these results can often be leveraged to obtain bounds on more complex classes. Due to the theoretical importance of finite dimensions in statistical and computational learning theory, the VC dimension in particular has been studied extensively. Unfortunately, many of the results only establish finiteness of the VC dimension of a class, rather than providing a specific bound. Fortunately, a large number of results were derived for complex classes such as artificial neural networks (for example Anthony, 1997). These complex classes and their layered structure encouraged the development of methods for relating the VC dimension of complex classes to the VC dimension of simple "building block" classes.

5.9.1 Shatter coefficients and VC dimension

Our first focus will be on obtaining bounds on the VC dimension. First we present some useful properties of shatter coefficients. These results are taken from Devroye et al. (1996, Theorem 13.5) and Devroye and Lugosi (2001, Theorem 4.1).

Theorem 5.28 (Properties of shatter coefficients). *Let $\mathscr{C}_1, \mathscr{C}_2$ be classes of subsets of a set $\mathcal{E}$, and $n_1, n_2 \geq 1$ be integers.*

- $\mathcal{N}_{\mathscr{C}_1}(n_1 + n_2) \leq \mathcal{N}_{\mathscr{C}_1}(n_1)\mathcal{N}_{\mathscr{C}_1}(n_2)$.
- $\mathcal{N}_{\mathscr{C}_1 \cup \mathscr{C}_2}(n_1) \leq \mathcal{N}_{\mathscr{C}_1}(n_1) + \mathcal{N}_{\mathscr{C}_2}(n_1)$.
- *Let*
  $$\mathscr{C}_3 = \{\mathcal{E} \setminus c : c \in \mathscr{C}_1\} \ .$$
  Then $\mathcal{N}_{\mathscr{C}_3}(n_1) = \mathcal{N}_{\mathscr{C}_1}(n_1)$.
- *Let*
  $$\mathscr{C}_3 = \{c_1 \cap c_2 : c_i \in \mathscr{C}_i\} \ .$$
  Then $\mathcal{N}_{\mathscr{C}_3}(n_1) \leq \mathcal{N}_{\mathscr{C}_1}(n_1)\mathcal{N}_{\mathscr{C}_2}(n_1)$.
- *Let*
  $$\mathscr{C}_3 = \{c_1 \cup c_2 : c_i \in \mathscr{C}_i\} \ .$$
  Then $\mathcal{N}_{\mathscr{C}_3}(n_1) \leq \mathcal{N}_{\mathscr{C}_1}(n_1)\mathcal{N}_{\mathscr{C}_2}(n_1)$.
- *Let*
  $$\mathscr{C}_3 = \{c_1 \times c_2 : c_i \in \mathscr{C}_i\} \ .$$
  Then $\mathcal{N}_{\mathscr{C}_3}(n_1) \leq \mathcal{N}_{\mathscr{C}_1}(n_1)\mathcal{N}_{\mathscr{C}_2}(n_1)$.

It follows that if $\mathscr{C}_1$ and $\mathscr{C}_2$ have finite VC dimension, so do the four forms of $\mathscr{C}_3$ considered above, as well as $\mathscr{C}_1 \cup \mathscr{C}_2$. In fact, it can further be shown (by combining the above result with the VCSS lemma) that $\mathrm{VC}(\mathscr{C}_1 \cup \mathscr{C}_2) \leq \mathrm{VC}(\mathscr{C}_1) + \mathrm{VC}(\mathscr{C}_2) + 1$.

As pointed out in Examples 5.13 and 5.14:

- the class of (Borel) closed sets has infinite VC dimension;
- the class of intervals on $\mathbb{R}$ with lower endpoint $-\infty$ has VC dimension 1;

- the class of halfspaces on $\mathbb{R}$ has VC dimension 2;
- and the class of intervals on $\mathbb{R}$ has VC dimension 2.

We shall now generalize the last three of these results in various ways. We begin by extending the result for intervals on $\mathbb{R}$ with lower endpoint $-\infty$.

Example 5.15. Consider the class $\mathscr{C}$ of Cartesian products of intervals on $\mathbb{R}$ with lower endpoint $-\infty$: given $s = (s_1, s_2, \cdots, s_N) \in \mathbb{R}^N$, define

$$c_s = \prod_{i=1}^{N} (-\infty, s_i] = \{\eta \in \mathbb{R}^N : \eta \leq s\} \ .$$

The VC dimension of $\mathscr{C}$ is N (Devroye et al., 1996, Theorem 13.8). □

Example 5.16. A monotone layer is a generalization of these intervals. Specifically, a set c is a monotone layer if $\eta_1 \in c$ implies $\eta_2 \in c$ for all $\eta_2 \leq \eta_1$. This definition applies in an arbitrary setting, not just in $\mathbb{R}^N$.

It turns out that the class of monotone layers in $\mathbb{R}^N$ has infinite VC dimension (see Devroye et al., 1996, Problem 13.19). □

Definition 5.5 (Linearly ordered by inclusion). We say a class of sets $\mathscr{C}$ is *linearly ordered by inclusion* (Devroye et al., 1996, Problem 13.15) if, for any two elements $c_1, c_2 \in \mathscr{C}$, either $c_1 \subseteq c_2$ or $c_2 \subseteq c_1$.

It is clear that the class of intervals with lower endpoint $-\infty$ is linearly ordered by inclusion. By a proof analogous to Example 5.11 one obtains the following result:

Theorem 5.29. *Suppose a class of sets $\mathscr{C}$ satisfies $|\mathscr{C}| \geq 2$. If $\mathscr{C}$ is linearly ordered by inclusion,* $\mathrm{VC}(\mathscr{C}) = 1$.

Example 5.17. Let R be a non-empty subset of a vector space over $\mathbb{R}$, which is star-shaped around 0.[125] For any $v > 0$, define $vR = \{v\eta : \eta \in R\}$.

Then the convex cone generated by R, i.e. the class $\mathscr{C} = \{vR : v > 0\}$, is linearly ordered by inclusion, so that $\mathrm{VC}(\mathscr{C}) = 1$. □

[125] A subset R of a real vector space $\mathcal{E}$ is star-shaped around a point η if, for all $v \in [0, 1]$, we have $\eta + v(\eta' - \eta) \in R$ for all $\eta' \in R$. Thus any line segment connecting a point in R to η lies entirely in R.

One can generalize the fact that the class of intervals in $\mathbb{R}$ has VC dimension 2 to Euclidean N-space in two obvious ways. One analog of the interval is the class of N-dimensional axis-parallel rectangles[126], for which it can be shown that the VC dimension is $2N$ (Blumer et al., 1989). The other analog of the interval is the class of balls in N-dimensional space — the VC dimension of this class is $N+1$ (Devroye and Lugosi, 2001, Dudley, 1979).

Furthermore, one can generalize the result for intervals in a different way. Consider the class of unions of up to K intervals in $\mathbb{R}$. Then the class of intervals corresponds to $K = 1$. In general, the VC dimension of such a class is $2K$ (Blumer et al., 1989).

5.9.2 Thresholded classes — VC and pseudodimension

Next we turn to a generalization of halfspaces. Note that we can write any halfspace in $\mathbb{R}$ as

$$\{\eta \in \mathbb{R} : v\eta \geq s\}$$

for some $v, s \in \mathbb{R}$. This is easily generalized to $\mathbb{R}^N$: a halfspace in $\mathbb{R}^N$ is a set of the form

$$\{\eta \in \mathbb{R}^N : \langle v, \eta \rangle \geq s\}$$

for $v \in \mathbb{R}^N, s \in \mathbb{R}$. The boundary of this set,

$$\{\eta \in \mathbb{R}^N : \langle v, \eta \rangle = s\}$$

is an N-dimensional plane, which we call a *hyperplane*. It can be shown that the VC dimension of the class of all halfspaces in $\mathbb{R}^N$ is $N+1$ (Vidyasagar, 2002, Wenocur and Dudley, 1981).

Steele (1975) and Dudley (1978) provide a powerful bound on the VC dimension of a wide variety of classes:

Theorem 5.30 (VC dimension of a thresholded affine space). *Let $\mathcal{V}$ be an affine space*[127] *of real-valued functions on $\mathbb{R}^N$ with dimension K. For*

[126] Note that the axes in $\mathbb{R}^N$ can be rotated by a change of coordinates: axis-parallel here thus merely means that all the rectangles should be oriented similarly.

[127] An affine space is a generalization of a vector space. The generalization of this result from vector spaces to affine spaces is due to Hush and Scovel (2004).

each $\phi \in \mathcal{V}$ define

$$c_\phi = \{\eta \in \mathrm{I\!R}^n : \phi(\eta) \geq 0\} \ .$$

Then the class of sets $\mathscr{C}(\mathcal{V}) = \{c_\phi : \phi \in \mathcal{V}\}$ has VC dimension K (Anthony, 1994, Devroye et al., 1996).

Equivalently, given that $\phi_1(\eta), \cdots, \phi_K(\eta)$ are linearly independent real-valued functions on $\mathrm{I\!R}^N$, consider the sign of a linear combination of these functions, $\mathrm{sgn}(\sum_{i=1}^{K} v_i \phi_i(\eta))$*. The VC dimension of the class of such thresholded linear combinations equals K (Vapnik, 1998).*

Example 5.18. Let $\phi_i(\eta) = \eta^{(i)}$, the i-th component of $\eta \in \mathrm{I\!R}^N$, for $i \in [1 : N]$; and let $\phi_{N+1}(\eta) = 1$. These ϕ_i are linearly independent, and thresholding their span at zero yields the set of halfspaces in $\mathrm{I\!R}^N$. This confirms that the VC dimension of the set of halfspaces in $\mathrm{I\!R}^N$ is $N + 1$.

Another important result concerns the restricted class of halfspaces where we have $s = 0$. In this case, the corresponding hyperplanes pass through the origin in $\mathrm{I\!R}^n$. This restricted class is obtained by thresholding the span of $\phi_1, \cdots, \phi_N$ at zero. It follows that the VC dimension of this restricted class is N. □

Example 5.19. This example is based on Devroye et al. (1996, Corollary 13.2).

Consider the class of closed balls in $\mathrm{I\!R}^N$. A closed ball is a set of the form

$$\left\{\eta \in \mathrm{I\!R}^N : ||\eta - v||^2 \leq s\right\} \ ,$$

for $v \in \mathrm{I\!R}^N$, $s \in \mathrm{I\!R}^+$.

Setting $\phi_i(\eta) = \eta^{(i)}$, $\phi_{N+1}(\eta) = 1$, and $\phi_{N+2}(\eta) = ||\eta||^2$ and expanding the norms coordinate-wise, one sees that any closed ball can be obtained by thresholding a function in the span of the ϕ_i.

It follows that the VC dimension of the class of closed balls does not exceed $N + 2$. As we have seen, it actually equals $N + 1$. The reason the bound is not exact is because the thresholded span of the ϕ_i contains more sets than just the closed balls. Note that the class of closed balls are parametrized by only $N + 1$ real values, with the coefficient of $||\eta||^2$ fixed at 1. □

Example 5.20. In this example we consider a larger class than the closed balls, the class of closed axis-parallel ellipsoids in $\mathrm{I\!R}^N$. A closed axis-parallel ellipsoid is a set of the form

$$\{\eta \in \mathrm{I\!R}^N : (\eta - v)^T \mathscr{S} (\eta - v) \leq 1\} \ ,$$

where $v \in \mathbb{R}^N$ and $\mathscr{S}$ is a positive definite symmetric $N \times N$ matrix.

Expanding $(\eta - v)^T \mathscr{S}(\eta - v) - 1$ into its components leads one to consider the basis consisting of the N functions of the form $[\eta^{(i)}]^2$, the $\frac{N(N-1)}{2}$ functions of the form $\eta^{(i)}\eta^{(j)}$, and the constant function 1. Clearly thresholding the span of these functions yields a class containing all these ellipsoids.

It follows that the VC dimension of the class of closed axis-parallel ellipsoids does not exceed

$$N + \frac{N(N-1)}{2} + 1 = \frac{N(N+1)}{2} + 1 \ .$$

□

Example 5.21. This example shows that the pseudodimension of $\mathcal{V}$ in Theorem 5.30 is K or $K+1$.

The pseudodimension of $\mathcal{V}$ is the VC dimension of $\mathcal{V}_t$. Clearly $\mathscr{C}(\mathcal{V}) \subseteq \mathcal{V}_t$, so that $\mathrm{pdim}(\mathcal{V}) \geq K$.

Now, consider $\mathcal{V}'$, the smallest vector space containing $\mathcal{V}$ and the constant function $\{1\}$. Any element of $\mathcal{V}_t$ can be written as

$$c_{\phi,s}\{\eta \in \mathbb{R}^N : \phi(\eta) \geq s\} = \{\eta : \phi(\eta) - s \geq 0\}$$

for some $s \in \mathbb{R}$. Thus, it follows that $\mathcal{V}_t \subseteq \mathscr{C}(\mathcal{V}')$, where $\mathscr{C}(\mathcal{V}')$ is defined by analogy to $\mathscr{C}(\mathcal{V})$. Furthermore, the dimension of $\mathcal{V}'$ is either K or $K+1$, yielding the result we seek. □

Some modifications to an hypothesis class do not affect certain dimension measures of complexity.

Example 5.22. Let $\mathcal{V}$ be as in Theorem 5.30, and let v be an arbitrary real-valued function. Define the class

$$v + \mathcal{V} = \{v + \phi : \phi \in \mathcal{V}\} \ .$$

For each $\phi \in v + \mathcal{V}$ define

$$c_\phi = \{\eta \in \mathbb{R}^N : c_\phi \geq 0\} \ .$$

Then the class of sets $\mathscr{C}(v + \mathcal{V}) = \{c_\phi : \phi \in v + \mathcal{V}\}$ has VC dimension K (Wenocur and Dudley, 1981).

In other words, adding a fixed function to every function in $\mathcal{V}$ before thresholding does not change the VC dimension of the class.

It follows immediately that $\mathrm{pdim}(\mathscr{C}(v+\mathcal{V}))$ is K or $K+1$. Thanks to the first result in the following example, we actually have $\mathrm{pdim}(v+\mathcal{V}) = \mathrm{pdim}(\mathcal{V})$. □

Example 5.23. Consider an arbitrary real-valued function class $\mathcal{V}$, and define $v+\mathcal{V}$ as in Example 5.22. Then $\mathrm{pdim}(v+\mathcal{V}) = \mathrm{pdim}(\mathcal{V})$. This result is due to Wenocur and Dudley (1981).

Define the class

$$v \circ \mathcal{V} = \{v \circ \phi : \phi \in \mathcal{V}\}$$

for a function $v : [0,1] \to \mathbb{R}$. If v is nondecreasing, $\mathrm{pdim}(v \circ \mathcal{V}) \le \mathrm{pdim}(\mathcal{V})$ (Dudley, 1987, Nolan and Pollard, 1987). □

Thus, when working with a real-valued hypothesis class, one can add a specific function to each element of the class to form a new class with the same pseudodimension; in addition, composing each element of a real-valued hypothesis class with a fixed, non-decreasing function cannot result in a new class with a larger pseudodimension.

Blumer et al. (1989) provides a valuable result for bounding the VC dimension of classes constructed by finite unions or intersections from classes with known or bounded VC dimension, based on a slightly weaker result from Haussler (1986):

Theorem 5.31 (Lemma 3.2.3 of Blumer et al., 1989). *Let $\mathscr{C}^\star$ denote the class of unions of up to K elements of a class $\mathscr{C}$. Then*

$$\mathrm{VC}(\mathscr{C}^\star) \le 2\,\mathrm{VC}(\mathscr{C}) K \log_2(3K) \ .$$

The same result applies when $\mathscr{C}^\star$ is the class of intersections of up to K elements of $\mathscr{C}$.

Note that the result for intervals above shows that this bound can be quite loose: the union of up to K intervals has VC dimension exactly $2K$, and the original class of intervals has VC dimension 1.

Example 5.24. Consider the class $\mathscr{C}^\star$ of intersections of up to K halfspaces in $\mathbb{R}^N$. Since the VC dimension of the underlying class is $N+1$, we have that $\mathrm{VC}(\mathscr{C}^\star) \le 2(N+1)K\log_2(3K)$.

Note that if $K \geq N+1$, $\mathscr{C}^\star$ contains the convex polytopes in $\mathbb{R}^N$ with up to K facets[128]

The restriction to a finite number K of intersections or unions is necessary in this case (and many others). A direct derivation of the VC dimension of the class of convex polytopes with up to K facets by a direct shattering argument yields $\mathrm{VC}(\mathscr{C}^\star) = 2K-1$ (Blumer et al., 1989, Example 3.2.2). It follows that the class of all convex polytopes in $\mathbb{R}^N$ (with an arbitrary number of facets) has infinite VC dimension.

In addition, we note that if $K = N+1$, the class of convex polytopes with K facets is the class of N-simplices[129]. Thus the class of N-simplices has VC dimension at most $2N+1$. □

A rather irksome restriction in Theorem 5.31 is the restriction to employing either unions or intersections. In practice, we may like to employ combinations of the two, as well as set difference. The following result allows us to do so:

Theorem 5.32 (VC dimension of a set-theoretic formula). *Let $\mathscr{C}^\star$ denote the class of all sets obtainable by any set-theoretic formula (using unions, intersections, and set differences) involving K elements of a class $\mathscr{C}$. Then*

$$\mathrm{VC}(\mathscr{C}^\star) \leq 2\,\mathrm{VC}(\mathscr{C})K\log_2(2\,\mathrm{VC}(\mathscr{C})K) \ .$$

Example 5.25. It follows from this result that the class of all (convex and non-convex) polygons in $\mathbb{R}^2$ with at most 5 sides has VC dimension at most

$$2\cdot(2+1)\cdot 5\log_2(2\cdot(2+1)\cdot 5) = 30\log_2 30 < 148 \ .$$

□

Note that a number of classification bounds were provided in terms of covering numbers, after which the limit is often taken as the scale tends to zero. The following result from Alexander (1984) on covering numbers for zero-one functions seems to underpin the result in Theorem 5.32.

Theorem 5.33 (Covering numbers of a set-theoretic formula). *Let $\mathcal{V}$ denote the class of indicator functions corresponding to the concepts in $\mathscr{C}$ in*

[128] A polytope is a higher-dimensional generalization of a polygon or polyhedron, and a facet corresponds to the concept of the edge of a polygon or the face of a polyhedron.

[129] A simplex is a higher dimensional analog of a triangle: a convex polytope with the least number of facets.

Theorem 5.32, and let $\mathcal{V}^\star$ be the corresponding class of indicator functions of concepts in $\mathscr{C}^\star$.

Then, for any distribution P,

$$\mathcal{N}_{1,P}(\gamma, \mathscr{C}^\star) \leq \left[\mathcal{N}_{1,P}\left(\frac{\gamma}{K}, \mathscr{C}\right)\right]^K \quad .$$

5.9.3 Covering number bounds, pseudodimension and Euclidean classes

We now move our focus from the VC dimension and shatter coefficients to more general covering numbers and their associated dimension quantities. We begin with some bounds relating covering numbers of simple classes to those of classes built from them. To do this, we generally assume that the functions in any function classes are defined on $\mathrm{I\!R}^n$.

Definition 5.6 (Envelope). Suppose a function $\phi^\star$ satisfies $\phi^\star \geq \phi$ for all $\phi \in \mathcal{V}$. Then $\phi^\star$ is called *an envelope* of $\mathcal{V}$. The smallest envelope of $\mathcal{V}$ is called the *natural envelope* of $\mathcal{V}$, denoted by $\mathrm{env}_{\mathcal{V}}$.

Example 5.26. All the decision classes we consider have the constant function 1 as an envelope. If, for every point $x \in \mathcal{X}$, there is a $w \in \mathcal{W}$ such that $w(x) = 1$, 1 is a natural envelope for $\mathcal{W}$, □

Many of the results we have stated can be generalized to classes of bounded and even unbounded loss by incorporating a type of normalization based on the envelope of the relevant function classes. Since we are considering building functions into $[0, 1]$ from other components which do not necessarily map into $[0, 1]$, we need to introduce the envelope concept here.

The following theorem is mostly based on the section on packing numbers in Pollard (1990, Chapter 5).

Theorem 5.34 (Properties of covering numbers). *Let $\mathcal{V}_1, \mathcal{V}_2$ be real-valued function classes, d be a pseudometric, and let $\gamma_1, \gamma_2 > 0$.*

1. $\mathcal{N}(\gamma_1, \mathcal{V}_1 \cap \mathcal{V}_2, d) \leq \min(\mathcal{N}(\gamma_1, \mathcal{V}_1, d), \mathcal{N}(\gamma_1, \mathcal{V}_2, d))$.

2. *$\mathcal{N}(\gamma_1, \mathcal{V}_1 \cup \mathcal{V}_2, d) \leq \mathcal{N}(\gamma_1, \mathcal{V}_1, d) + \mathcal{N}(\gamma_1, \mathcal{V}_2, d)$.*
3. *Let $\mathcal{V}_3$ be any of the following classes:*
   - $\{\phi_1 + \phi_2 : \phi_1 \in \mathcal{V}_1, \phi_2 \in \mathcal{V}_2\}$*;*
   - $\{\max(\phi_1, \phi_2)(\cdot) : \phi_1 \in \mathcal{V}_1, \phi_2 \in \mathcal{V}_2\}$*; or*
   - $\{\min(\phi_1, \phi_2)(\cdot) : \phi_1 \in \mathcal{V}_1, \phi_2 \in \mathcal{V}_2\}$,

   where the max *and* min *of the functions are defined pointwise (i.e.* $[\max(\phi_1, \phi_2)](x) = \max(\phi_1(x), \phi_2(x))$*).*

   Then we have

   $$\mathcal{N}(\gamma_1 + \gamma_2, \mathcal{V}_3, d) \leq \mathcal{N}(\gamma_1, \mathcal{V}_1, d)\mathcal{N}(\gamma_2, \mathcal{V}_2, d) \ .$$
4. *Suppose for $i = 1, 2$, we have that the constant function K_i is an envelope for $\mathcal{V}_i$, and that d satisfies*

   $$d(K\phi_1, K\phi_2) \leq |K| d(\phi_1, \phi_2)$$

   for all $K \in \mathbb{R}$ and functions $\phi_1 \in \mathcal{V}_1, \phi_2 \in \mathcal{V}_2$. Let

   $$\mathcal{V}_3 = \{\phi_1\phi_2 : \phi_1 \in \mathcal{V}_1, \phi_2 \in \mathcal{V}_2\} \ .$$

   Then we have

   $$\mathcal{N}(K_2\gamma_1 + K_1\gamma_2, \mathcal{V}_3, d) \leq \mathcal{N}(\gamma_1, \mathcal{V}_1, d)\mathcal{N}(\gamma_2, \mathcal{V}_2, d) \ .$$

The first two results are trivial, and all three cases for the third result are easily obtained using the triangle inequality. The fourth result also uses the triangle inequality, but in a slightly more sophisticated way. It is a special case of the corresponding result in Pollard (1990), since it assumes the functions in the classes are bounded.

Clearly applying these results repeatedly yields bounds on covering numbers of sums, products, minima, maxima, intersections and unions of k functions.

Example 5.27. We apply the result to the sum of K functions with equal γ_i. Let $\mathcal{V} = \{\sum_{i=1}^{K} \phi_i : \phi_i \in \mathcal{V}_i\}$ for some function classes $\mathcal{V}_1, \cdots, \mathcal{V}_k$. Then, for every $\gamma > 0$ and every sample Q,

$$\mathcal{N}_{p,Q}(\gamma, \mathcal{W}) \leq \prod_{i=1}^{K} \mathcal{N}_{p,Q}\left(\frac{\gamma}{K}, \mathcal{V}_i\right) \ .$$

This result, with $p = 1$, is Devroye et al. (1996, Theorem 29.6). □

We recall that one can generally obtain a bound for an algorithm employing a given hypothesis class in terms of covers of a surrogate hypothesis class. Many such classes are obtained by composition with some kind of well-behaved function.

Next, we present a covering number result for the composition of functions where the outer functions are Lipschitz, which is a reasonably straightforward generalization of Anthony and Bartlett (1999, Lemma 14.3).

Theorem 5.35. *Let $\mathcal{V}_1$ be a class of functions from a set $\mathcal{E}_1$ to a metric space $(\mathcal{E}_2, d)$, and $\mathcal{V}_2$ be a class of K-Lipschitz (w.r.t. d) functions from $\mathcal{E}_2$ into* $\mathbb{R}$.

Then, for any distribution P_1 on $\mathcal{E}_1$,

$$\mathcal{N}_{\infty,P_1}(K\gamma_1 + \gamma_2, \mathcal{V}_2 \circ \mathcal{V}_1) \leq \mathcal{N}_{\infty,P_1}(\gamma_1, \mathcal{V}_1)\mathcal{N}_{\infty,P_2}(\gamma_2, \mathcal{V}_2) \ ,$$

where

$$\mathcal{V}_2 \circ \mathcal{V}_1 = \{\phi_2 \circ \phi_1 : \phi_1 \in \mathcal{V}_1, \phi_2 \in \mathcal{V}_2\} \ ,$$

and P_2 is any distribution on $\mathcal{E}_2$ with support equal to $\mathcal{E}_2$.

Example 5.28. A rather trivial, yet very important, example of this is when $\mathcal{V}_2$ consists of a single K-Lipschitz squashing function $\phi^\star$ being applied to an hypothesis class $\mathcal{H}$. In that case, the second covering number is clearly one, and we obtain

$$\mathcal{N}_{\infty,P}(\gamma, \phi^\star(\mathcal{H})) \leq \mathcal{N}_{\infty,P}\left(\frac{\gamma}{K}, \mathcal{H}\right) \ .$$

The first notable field of application of this result is to the zero-one loss function $L(y_1, y_2) = I(y_1 \neq y_2)$ when $y_1, y_2 \in \{0, 1\}$. This function is 1-Lipschitz. A second useful application of this result is for margin bounds: the trimming function $\pi_{(s-\gamma', s+\gamma')}$ is 1-Lipschitz. In both cases, it follows that composition with the relevant function can not increase the covering numbers employed in the bounds. Note that for margin bounds, γ' must be selected at least as large as the margin required for the empirical margin loss, since otherwise, no points achieve the margin for any function in the trimmed class. Thus, γ' is typically chosen equal to the margin under consideration.

See Bartlett (1998, Proposition 25) for related results. □

Furthermore, note that the loss class can be seen as the composition of the decision class with the loss function, and the decision class can often be seen as the composition of the hypothesis class with the strategy. Thus results like the one above are powerful tools for obtaining covering numbers on the loss class.[130]

We now present a closely related result which can provide a connection between the covering number of a decision class and a loss class for a wide variety of loss functions, or between those of a decision and hypothesis class for a variety of strategies. It is a basic generalization of Lemma 2 of Williamson et al. (1998b) to p-norms, and allowing P to be an arbitrary distribution.

Theorem 5.36. *Let $\mathcal{V}$ be a class of functions mapping $\mathcal{X}$ into $[\mathscr{L}, \mathscr{U}] \subseteq \mathrm{I\!R}$, and $v : [\mathscr{L}, \mathscr{U}]^2 \to \mathrm{I\!R}^+$. For $\phi \in \mathcal{V}$, define*

$$\phi'((x, y)) = v(\phi(x), y)$$

and

$$\mathcal{V}' = \{\phi' : \phi \in \mathcal{V}\} \ .$$

Suppose there is a function v' which is K-Lipschitz[131] *on $[\mathscr{L} - \mathscr{U}, \mathscr{U} - \mathscr{L}]$ such that $v(y_1, y_2) = v'(y_1 - y_2)$ for all $y_1, y_2 \in [\mathscr{L}, \mathscr{U}]$.*

Then, for all distributions P on $\mathcal{Z} = \mathcal{X} \times [\mathscr{L}, \mathscr{U}]$, all $p \geq 1$, and all $\gamma > 0$,

$$\mathcal{N}_{p,P}(\gamma, \mathcal{V}') \leq \mathcal{N}_{p,P}\left(\frac{\gamma}{K}, \mathcal{V}\right) \ .$$

When P is an empirical distribution, a bound on $\mathcal{N}_{\infty,P}(\gamma, \mathcal{V}')$ can be found from the bound on $\mathcal{N}_{1,P}(\gamma, \mathcal{V}')$.

Example 5.29. Suppose for a problem we have $\mathcal{A} = \mathcal{Y} = [\mathscr{L}, \mathscr{U}]$, with decision class $\mathcal{W}$.

Consider the polynomial loss function $L(r, y) = |r - y|^K$. Then L can be written as $L'(r - y)$, and L' can be shown to be Lipchitz with constant $K(\mathscr{U} - \mathscr{L})^{K-1}$ (Williamson et al., 1998b, Lemma 2). This allows one to easily obtain bounds on two very common loss functions: the absolute loss, and the squared error loss.

[130] Note that in the modified setting, we assume the loss function is the identity function. This is useful for theoretical derivations. However, in practice, bounds must be obtained for the loss class, while the loss function is not typically the identity function.

[131] Actually, a slightly weaker condition often suffices — see the original article for more details.

For example, the covering numbers of the loss class for squared error loss, when $\mathcal{Y} = [0, 1]$, is bounded by

$$\mathcal{N}_{p,P}\left(\gamma, \mathcal{F}_{\mathcal{W}}\right) \leq \mathcal{N}_{p,P}\left(\frac{\gamma}{2}, \mathcal{W}\right) \quad .$$

□

We shall see that many function classes $\mathcal{V}$ satisfy the following condition: for every distribution Q with $\mathbb{E}_{\eta\sim Q}\, \mathrm{env}_{\mathcal{V}}(\eta)$ finite, we have

$$\mathcal{N}_{1,Q}(\gamma, \mathcal{V}) \leq E_1 \left(\frac{\mathbb{E}_{\eta\sim Q}\, \mathrm{env}_{\mathcal{V}}(\eta)}{\gamma}\right)^{E_2} \quad ,$$

for some constants $E_1, E_2 \geq 0$. Following Nolan and Pollard (1987), we call such a class (E_1, E_2)-Euclidean[132]. A Euclidean class is any class which is Euclidean for some constants (E_1, E_2). If the result holds with $\mathrm{env}_{\mathcal{V}}$ replaced by another envelope $\phi^{\star}$, we say that the class is (E_1, E_2)-Euclidean for $\phi^{\star}$. This is a slightly stronger condition, since $\phi^{\star} \geq \mathrm{env}_{\mathcal{V}}$.

If an envelope $\phi^{\star}$ of $\mathcal{V}$ is bounded, the covering numbers are bounded for every distribution Q, including those where Q is an empirical distribution. Note that the covering number bound is then independent of the sample size underlying the empirical distribution.

We have the following result implicit in Pollard (1984, Lemma II.25).

Theorem 5.37. *If $\varrho = \mathrm{pdim}(\mathcal{V})$ is finite, then $\mathcal{V}$ is $(E_1(\varrho), E_2(\varrho))$-Euclidean, where $E_1(\varrho) = 2\varrho$, and $E_2(\varrho) = \max(v_0(\varrho), [v_1(\varrho)]^2)$, with $B(\varrho)$ and $v_0(\varrho)$ as follows:*

- *$v_0(\varrho)$ is the solution to*

$$(1 + 4 \ln v)^{\varrho} = \sqrt{v} \quad . \tag{5.56}$$

- *$v_1(\varrho)$ is such that*

$$\sum_{i=0}^{\varrho} \binom{j}{i} \leq v_1(\varrho) j^{\varrho}$$

for all $j \geq \varrho$, and

$$2^j \leq v_1(\varrho) j^{\varrho}$$

[132] Because their covering numbers grow at a rate similar to an E_2-dimensional Euclidean space.

for $1 \le j < \varrho$.[133]

It is natural to consider whether there are any other Euclidean classes besides the classes of finite pseudodimension. Since any (E_1, E_2)-Euclidean class is also (E_1', E_2')-Euclidean when $(E_1', E_2') \ge (E_1, E_2)$, we can show this is the case by showing that for any (E_1, E_2), there is a ϱ such that $(E_1(\varrho), E_2(\varrho)) \ge (E_1, E_2)$.

Clearly, for any $\varrho > \frac{E_1}{2}$, we have $E_1(d) > E_1$. Let us next consider $v_0(\varrho)$: rewriting (5.56), we have that $v_0(\varrho)$ is the solution to

$$\varrho = \frac{\ln v}{2\ln(1 + 4\ln v)} \quad .$$

The right hand side is $O(\frac{\ln v}{\ln \ln v})$ so that it becomes infinitely large as $v \to \infty$. Thus, we can obtain $v_0(\varrho)$, and hence $E_2(\varrho)$, arbitrarily large by selecting d large enough.

We thus have the following characterization of Euclidean classes.

Theorem 5.38. *A function class $\mathcal{V}$ is Euclidean if and only if it has finite pseudodimension.*

It follows from this result that function classes with finite fat-shattering dimension at all scales, but infinite pseudodimension, are uniform Glivenko-Cantelli classes, but not Euclidean classes.

It is not difficult to show that the set of Euclidean classes is closed under finite unions, intersections, maxima, minima, addition, and multiplication, based on Theorem 5.34. In fact, Theorem 5.34 may often allow us to obtain upper bounds for E_1 and E_2 for the modified classes given the corresponding values for the original classes. Combining such results with Theorem 5.37 would yield bounds for many complicated classes of finite pseudodimension, but for which it is not practical to find the pseudodimension directly. However, it is generally more desirable to employ Theorem 5.34 together with

[133] Clearly this is possible, since the left hand side of the first inequality (which corresponds to the expression in the VCSS lemma) are merely points on a polynomial of degree ϱ. This interpretation explains why VC classes are sometimes called polynomial classes.

a more direct bound on the covering numbers of a class of finite pseudodimension, such as Theorem 5.27.

Anthony and Bartlett (1999, Theorems 11.3 and 11.14) provides essentially the following results.

Theorem 5.39. *Let v be a monotonic function and $\mathcal{V}$ a class of real-valued functions. Let $v \circ \mathcal{V} = \{v \circ \phi : \phi \in \mathcal{V}\}$. Then*

$$\operatorname{pdim}(v \circ \mathcal{V}) \leq \operatorname{pdim}(\mathcal{V}) \enspace .$$

Theorem 5.40. *If a function class $\mathcal{V}$ is closed under scalar multiplication, then*

$$\operatorname{pdim}(\mathcal{V}) = \operatorname{fat}_{\mathcal{F}}(\gamma)$$

for all $\gamma > 0$.

The first result basically says that one can not increase the pseudodimension of a function class by smoothing the functions in the class. The second result applies in particular to any function classes which form a vector space, so that we may be able to get bounds on the pseudodimension from Theorem 5.30. In these cases, it is recommended to bound covering numbers by using results based on the pseudodimension. Bounds on covering numbers based on fat-shattering dimension rely on the fact that the fat-shattering dimension at resolution γ can be substantially less than the pseudodimension. This improvement in the dimension quantity often outweighs increases in other components of the bound. However, this is not the case when the fat-shattering dimension is constant (or decreases too slowly).

5.9.4 More covering number bounds, and fat-shattering dimension

For other basic function classes, bounds on the fat-shattering dimension have relied on knowledge of the functions in the class. We present two such results next, after providing an introductory definition.

Definition 5.7 (Total and bounded variation). A real-valued function ϕ defined over an interval $[\mathscr{L}, \mathscr{U}]$ is said to have bounded variation on

$[\mathscr{L}, \mathscr{U}]$ if there is a finite K such that for every $j \in \mathbb{N}$, and every increasing sequence $(v_1, \cdots, v_n)$ with $v_1 \geq \mathscr{L}$ and $v_n \leq \mathscr{U}$, we have

$$\sum_{i=1}^{j-1} |\phi(v_{i+1}) - \phi(v_i)| \leq K \ .$$

The total variation of ϕ on $[\mathscr{L}, \mathscr{U}]$ is the infimum of all K satisfying the above inequality.

Informally, suppose one could walk along the graph of v from $\mathscr{L}$ to $\mathscr{U}$. Then the total variation of a function can be thought of as the total height climbed in the trip added to the total height descended. It should be clear that functions with smaller total variation tend to be less "wiggly" than functions with larger total variation. As a result, we can expect function classes containing such functions to be less prone to overfitting. We can quantify this intuition with a bound on the fat-shattering dimension.

Theorem 5.41. *Let $\mathcal{V}$ be the class of all functions with total variation on $[0,1]$ not exceeding K. Then*

$$\mathcal{N}_{\infty,n}(\gamma, \mathcal{V}) < 2 \left(\frac{4n}{\gamma^2}\right)^{\left(1+\frac{2K}{\gamma}\right)\log_2 \frac{2en}{K}}$$

and

$$\text{fat}_{\mathcal{V}}(\gamma) = 1 + \left\lfloor \frac{K}{2\gamma} \right\rfloor \ .$$

This result is presented in Anthony and Bartlett (1999, Theorems 11.12 and 12.12), but the authors attribute the discovery to Simon (1997). This result is interesting in that the fat-shattering dimension is linear in $\frac{1}{\gamma}$. Thus it is clear that the class has infinite pseudodimension. Classes of functions with fat-shattering dimension growing as a polynomial in $\frac{1}{\gamma}$ are clearly better behaved than functions with faster growing dimensions. It turns out that polynomial growth has implications for the class to be able to tolerate noise. This application does not fall into our framework however, so we refer the interested reader to Bartlett et al. (1996), where some other implications are also discussed.

Other important functions are those which satisfy a Lipschitz condition. The following result relates the fat-shattering dimension of a class of Lipschitz functions to a covering number of the domain of the functions in the class.

Theorem 5.42 (Theorem 13 of Bartlett, 1998). *Let $(\mathcal{E}, d)$ be a totally bounded metric space. Let $\mathcal{V}$ be the class of all real-valued functions ϕ on $\mathcal{E}$ which are K-Lipschitz, i.e. for any $\eta_1, \eta_2 \in \mathcal{E}$, we have*

$$|\phi(\eta_1) - \phi(\eta_2)| \leq K d(\eta_1, \eta_2) \ .$$

Then $\mathrm{fat}_{\mathcal{V}}(\gamma) \leq \mathcal{N}\left(\mathcal{E}, \frac{\gamma}{K}, d\right)$.

Example 5.30. If $\mathcal{E}$ is a bounded subset of $\mathrm{I\!R}^N$, and d is the Euclidean distance, $\mathcal{N}(\mathcal{E}, \frac{\gamma}{K}, d)$ behaves like $(\frac{\gamma}{K})^{-n}$. Therefore, the class of functions which are K-Lipschitz on this space is Euclidean and thus has finite pseudodimension. □

We now present the well-known *radius-margin* bounds. The original results of this type were bounds on the VC dimension of a restricted class of half-spaces. These bounds are closely related to early work on the perceptron in the 1960's, but early proofs of the bound relied on a result which was thought to hold "by symmetry considerations" (see, for example, Vapnik, 1998). After the need for a more rigorous proof for this result was highlighted in Burges (1998) (and a correction to the result made), a proof for the result was finally presented in Hush and Scovel (2001), putting this class of bounds on a firm footing.

The following theorem is essentially the classical radius-margin bound on the VC dimension of restricted thresholded classifiers.

Theorem 5.43 (Radius-margin bound on VC dimension). *Let $\mathcal{E}$ be an inner product space, and let $\mathcal{E}' \subseteq \mathcal{E}$ be contained in some ball of radius c. For any $\eta \in \mathcal{E}$ and $s \in \mathrm{I\!R}$, we define the function $h_{\eta,s} : \mathcal{E} \to \mathrm{I\!R}$ by*

$$h_{\eta,s}(\eta') = \langle \eta, \eta' \rangle + s \ .$$

Consider the class

$$\mathcal{H} = \{h_{\eta,s} : \|\eta\| \leq K, s \in \mathrm{I\!R}\}$$

and its subclass

$$\mathcal{H}' = \left\{ h \in \mathcal{H} : \inf_{\eta' \in \mathcal{E}'} |h(\eta')| = 1 \right\} \ .$$

Consider the class $\mathcal{W}$ obtained by thresholding the elements of $\mathcal{H}'$ at zero,

$$\mathcal{W} = \operatorname{sgn}(\mathcal{H}') = \left\{ \operatorname{sgn} \circ h : h \in \mathcal{H}' \right\} \ .$$

The restriction of $\mathcal{W}$ to the points in $\mathcal{X}'$, $\mathcal{W}|_{\mathcal{E}'}$, satisfies

$$\operatorname{VC}\left(\mathcal{W}|_{\mathcal{E}'}\right) \leq \min\left\{ \left\lceil c^2 K^2 \right\rceil, N \right\} + 1 \ ,$$

where N is the dimension of $\mathcal{E}$ (possibly infinite).

While the above result has some theoretical appeal, the rise of margin bounds meant that it was more interesting to consider the fat-shattering dimension of the unthresholded functions which yielded the halfspaces. Indeed, in one of the first papers presenting margin bounds, Shawe-Taylor et al. (1998) gives such a fat-shattering dimension bound based on extending the result above by a γ-shattering argument. They obtain the following result.

Theorem 5.44 (Radius-margin bound on fat-shattering dimension). *Let $\mathcal{E}$ be an inner product space, and let $\mathcal{E}' \subseteq \mathcal{E}$ be contained in some ball of radius c. For any $\eta \in \mathcal{E}$ and $s \in \mathbb{R}$, define the function $h_{\eta,s} : \mathcal{E} \to \mathbb{R}$ by*

$$h_{\eta,s}(\eta') = \langle \eta, \eta' \rangle + s \ .$$

Consider the class

$$\mathcal{H} = \{ h_{\eta,s} : \|\eta\| = 1, |s| \leq c \} \ .$$

Then the restriction of $\mathcal{H}$ to $\mathcal{E}'$ satisfies

$$\operatorname{fat}_{\mathcal{H}|_{\mathcal{E}'}}(\gamma) \leq \min\left\{ \left\lceil \frac{9c^2}{\gamma^2} \right\rceil, N+1 \right\} + 1 \ ,$$

where N is the dimension of $\mathcal{E}$ (possibly infinite).

A more direct approach seems to have been taken in Gurvits (1997). This allowed tighter bounds to be obtained at the cost of some restrictions: notably, the ball containing $\mathcal{E}_0$ had to be centred at the origin, and s had to be zero. The following representative result is Shawe-Taylor and Cristianini (2000, Theorem III.6).

Theorem 5.45 (Radius-margin bound on fat-shattering dimension). *Let $\mathcal{E}$ be an inner product space. Let $\mathcal{H}$ be the class of linear functions with norm less than K. Let $\mathcal{E}_0$ be the ball of radius c about the origin in $\mathcal{E}$.*

Then

$$\mathrm{fat}_{\mathcal{H}|\mathcal{E}_0}(\gamma) \leq \left(\frac{Kc}{\gamma}\right)^2 \quad .$$

In principle, this result can be extended by adding an extra dimension to the space for the threshold, but then the size of the threshold is limited by the choice of c.

An important breakthrough in this area was the recent work of Hush and Scovel (2004). Their results provide exact values for the fat-shattering dimension of a restricted affine function class.

Theorem 5.46 (Fat-shattering dimension of restricted affine class). *Let $\mathcal{E}$ be an inner product space of finite dimension N, and $\mathcal{E}_0$ the ball of radius c about the origin of $\mathcal{E}$.*

For any $\eta \in \mathcal{E}$ and $s \in \mathrm{I\!R}$, we define the function $h_{\eta,s} : \mathcal{E} \to \mathrm{I\!R}$ by

$$h_{\eta,s}(\eta') = \langle \eta, \eta' \rangle + s \quad .$$

Define

$$\mathcal{H} = \{h_{\eta,s} : \|\eta\| \leq K, s \in \mathrm{I\!R}\} \quad ,$$

the class of affine transformations with the norm of the linear transformation component not exceeding K.

Denote, for $i \in \mathrm{I\!N}$,

$$\gamma_i = \begin{cases} \frac{1}{\sqrt{i-1}}, & i \text{ even;} \\ \frac{i}{(i-1)\sqrt{i+1}}, & i \text{ odd.} \end{cases}$$

Then

$$\mathrm{fat}_{\mathcal{H}|\mathcal{E}_0}(\gamma) = \begin{cases} N+1, & \frac{\gamma}{cK} \leq \gamma_{N+1} \\ i, & \gamma_{i+1} \leq \frac{\gamma}{cK} \leq \gamma_i, \quad 1 \leq i \leq N+1 \end{cases} \quad .$$

It follows that

$$\max\left\{\left(\frac{cK}{\gamma}\right)^2, 1\right\} \leq \mathrm{fat}_{\mathcal{H}|\mathcal{E}_0}(\gamma) \leq \min\left\{\left(\frac{cK}{\gamma}\right)^2 + \frac{5}{4}, N+1\right\} \quad .$$

The results in the theorem could be stated to apply to infinite dimensional spaces as well — this is only not done to avoid notational nuisances.

Note that this result combines the best of the previous two theorems, and in addition it applies for any threshold s.

We conclude this section with a consideration of convex hulls of hypothesis classes. Such convex hulls are relevant in a number of algorithms, notably ensemble classifiers, such as various voting classifiers and boosting, where a convex combination of *base classifiers* are employed to make a prediction.

Since the convex hull of a class $\mathcal{H}$ consists of real-valued functions, it would be useful to bound the pseudodimension or fat-shattering dimension of the class. Of course, it seems natural that such a dimension would depend on the capacity of the class $\mathcal{H}$.

The following specialization of a theorem from Bartlett (1998) is an example of such a result.

Theorem 5.47 (Fat-shattering dimension of a convex hull). *Let $\mathcal{H}$ be a class of functions mapping into $\left[\frac{-K}{2}, \frac{K}{2}\right]$. Suppose $\gamma \geq 0$ is such that $\varrho = \operatorname{fat}_{\mathcal{H}}(\frac{\gamma}{32}) \geq 1$.*

Then $\operatorname{fat}_{\operatorname{absconv}\mathcal{H}}(\gamma) \leq \frac{cK^2\varrho}{\gamma^2}\left(\ln\frac{K\varrho}{\gamma}\right)^2$ for some universal constant c.

Unfortunately, this result employs the unspecified constant c. However, the result is based on the following relationship between covering numbers, which does not depend on c.

Theorem 5.48 (Covering numbers of a convex hull). *Let $\mathcal{H}$ be a class of functions mapping into $\left[\frac{-K}{2}, \frac{K}{2}\right]$. Then*

$$\log_2\left(\mathcal{N}_{2,n}(\gamma, \operatorname{absconv}\mathcal{H})\right) \leq \frac{2K^2}{\gamma^2}\log_2\left(2\mathcal{N}_{2,n}\left(\frac{\gamma}{2}, \mathcal{H}\right) + 1\right) \quad .$$

5.9.5 Bounds from functional analysis

With the rise in popularity of the SV machine and related techniques, the idea of the *kernel trick* has become important. The kernel trick refers to a method for obtainining nonlinear techniques from linear techniques. The approach makes use of a *kernel function* $\mathscr{K}$. $\mathscr{K}(\eta_1, \eta_2)$ then typically corresponds to an inner product in some feature space:

$$\mathscr{K}(\eta_1, \eta_2) = \langle \Phi(\eta_1), \Phi(\eta_2) \rangle \quad ,$$

where Φ is a mapping from the input space to the feature space, which can typically be regarded as ℓ^2. More details on the kernel trick are available in the many books available treating support vector machines, e.g. Cristianini and Shawe-Taylor (2000), Schölkopf and Smola (2002).

Many kernel-based algorithms, and specifically support vector machines, select their decision rule from a class which can be expressed as (thresholded versions of) linear functions in feature space. As such, the results above can be applied to obtain bounds on the covering number of the hypothesis class. However, since the feature space typically has a much higher (often infinite) dimension, the transformation Φ employed by the kernel restricts the image of input space to a subset of the feature space.

Example 5.31. Consider the Gaussian kernel $\mathscr{K}(\eta_1, \eta_2) = \exp\left(\frac{-\|\eta_1 - \eta_2\|^2}{2\sigma^2}\right)$, where σ is a parameter controlling the bandwidth of the kernel.

In this case, for any η, we have $\mathscr{K}(\eta, \eta) = 1$, so that the kernel maps the input space onto the unit ball in ℓ^2. □

In what follows, we attempt to outline the approach used to obtain good bounds for kernel-based algorithms while glossing over technicalities. Full details can be found in Guo et al. (2002), Williamson et al. (1998a,b, 2000). In order to follow the argument, it is useful to introduce the idea of *entropy numbers*, which can be seen as a functional inverse of covering numbers.

Definition 5.8 (Entropy numbers). Let $j \in \mathbb{N}$. The j-th entropy number $\mathscr{N}_j(\mathcal{J})$ of a set $\mathcal{J} \subseteq \mathcal{E}$ is defined by

$$\mathscr{N}_j(\mathcal{J}) = \inf\{\gamma > 0 : \bar{\mathscr{N}}(\gamma, \mathcal{J}, d) \leq j\} \quad ,$$

where $(\mathcal{E}, d)$ is a pseudometric space.

The j-th entropy number $\mathscr{N}_j(T)$ of a bounded linear operator $T : \mathcal{E}_1 \to \mathcal{E}_2$ is defined by

$$\mathscr{N}_j(T) = \epsilon_j(T(B_{\mathcal{E}_1})) \enspace ,$$

where $B_{\mathcal{E}_1}$ denotes the unit ball in $\mathcal{E}_1$.

First, we note that $\mathscr{N}_1(T) = \|T\|$. Second, it is clear from the definition that $\mathscr{N}_j(T) \leq \gamma$ implies $\mathcal{N}(\gamma, T(U_E), d_{\mathcal{E}_\in}) \leq j$.

For the kernel $\mathscr{K}$, one can consider the integral operator $T_{\mathscr{K}}$ associated with $\mathscr{K}$. When $T_{\mathscr{K}}$ is non-negative and compact, the spectrum of $T_{\mathscr{K}}$ is countable, and one can express the map $\varPhi$ in terms of the eigenvalues (λ_i) and the associated eigenfunctions (ψ_i) of $T_{\mathscr{K}}$. Specifically, we have that $\varPhi : \mathcal{E} \to \ell^2$ satisfies

$$(\varPhi(\eta))^{(i)} = \sqrt{\lambda_i}\psi_i(\eta)$$

where the eigenvalues λ_i are sorted in non-increasing order. Furthermore the eigenfunctions are uniformly bounded by some constant $C_{\mathscr{K}}$. It follows that $\varPhi(\mathcal{E})$ is restricted to the axis-parallel parallelipiped[134]

$$[-\lambda_1 C_{\mathscr{K}}, \lambda_1 C_{\mathscr{K}}] \times [-\lambda_2 C_{\mathscr{K}}, \lambda_2 C_{\mathscr{K}}] \times \cdots \times [-\lambda_i C_{\mathscr{K}}, \lambda_i C_{\mathscr{K}}] \times \cdots \enspace .$$

The smallest ball containing this parallelipiped would intuitively have a radius equal to the "diagonal" of this infinite dimensional parallelipiped. However, $\varPhi(\mathcal{E})$ actually only occupies a thin slice of that ball. The key to improvement is to rather fit $\varPhi(\mathcal{E})$ into an ellipsoid. By appropriately scaling the axes of the ellipsoid, we can map the ellipsoid into the unit ball, and vice versa. If we denote the mapping from the unit ball to the ellipsoid by A, we can write $\varPhi = A(A^{-1}\varPhi)$. It turns out to be important that the scaling operation represented by A is a diagonal operator. We do not yet commit to a specific ellipsoid — we shall choose A to optimize the resulting bound.

Now, consider the class of linear functions in feature space represented by $\mathcal{H}_K = \{\langle v, \varPhi(\cdot)\rangle : \|v\| \leq K\}$. We shall provide bounds on $\mathcal{N}_{\infty,Q}(\gamma, \mathcal{H}_K)$.

[134] Technically, it is possible that this does not hold. However, similar results can still be obtained in this case. See the discussion in Guo et al. (2002, p. 241).

For an n-vector $\mathscr{V} = (v_1, \cdots, v_n) \in (\ell^2)^n$, define the *evaluation map* $S_{\mathscr{V}} : \ell^2 \to \ell^\infty_n$ by

$$S_{\mathscr{V}}(v) = (\langle v, v_1 \rangle, \langle v, v_2 \rangle, \cdots, \langle v, v_n \rangle) \ .$$

Let Q be an n-sample with $x_1, \cdots, x_n$ being the predictors for each data point. Denoting $(\Phi(x_1), \cdots, \Phi(x_n))$ by $\Phi(Q)$, $S_{\Phi(Q)}(v)$ is the vector

$$(\langle v, \Phi(x_1) \rangle, \cdots, \langle v, \Phi(x_n) \rangle) \ .$$

It follows that we are trying to find the covering number of

$$\{S_{\Phi(Q)}(v) : \|v\| \leq K\}$$

with the metric of $(\ell^\infty)^n$. We can rewrite this as

$$\left\{S_{A(A^{-1}\Phi(Q))}(Kv) : \|v\| \leq 1\right\} \ .$$

Now, because A is diagonal, it is self-adjoint, so

$$\left\langle Kw, AA^{-1}\Phi(x_i) \right\rangle = \left\langle AKw, A^{-1}\Phi(x_i) \right\rangle$$

and hence the set above is

$$\{S_{A^{-1}\Phi(Q)}(AKw) : \|w\| \leq 1\} = S_{A^{-1}\Phi(Q)} AK(B_{\ell^2}) \ .$$

The following results on entropy numbers of operators, taken from Williamson et al. (1998b), now allow us to obtain bounds. We write $\mathcal{L}(\mathcal{E}_1, \mathcal{E}_2)$ for the class of bounded linear operators from $\mathcal{E}_1$ into $\mathcal{E}_2$. The first result relates the entropy numbers of products of operators to those of each operator. The second result provides an entropy number result which is useful for the evaluation map. The third result bounds the entropy number of a diagonal operator, such as A.

Theorem 5.49 (Entropy numbers of composed operators). *Let $\mathcal{E}_1$, $\mathcal{E}_2$, and $\mathcal{E}_3$ be Banach spaces. Let $T_1 \in \mathcal{L}(\mathcal{E}_2, \mathcal{E}_3)$ and $T_2 \in \mathcal{L}(\mathcal{E}_1, \mathcal{E}_2)$. For any $j_1, j_2 \in \mathbb{N}$,*

$$\mathscr{N}_{j_1 j_2}(T_1 T_2) \leq \mathscr{N}_{j_1}(T_1) \mathscr{N}_{j_2}(T_2) \ . \tag{5.57}$$

Setting $j_1 = 1$, we have

$$\epsilon_j(T_1 T_2) \leq \|T_1\| \epsilon_j(S) \ .$$

A similar result is obtained by setting $j_2 = 1$.

Theorem 5.50 (Entropy number of an evaluation map). *Let $\mathcal{E}$ be a Hilbert space, $n \in \mathbb{N}$, and $T \in \mathcal{L}(\mathcal{E}, \ell_n^\infty)$. Then, for $c \geq 102.88$,*

$$\mathscr{N}_j(T) \leq c\|T\| \sqrt{(\ln j + 1)^{-1} \ln\left(1 + \frac{n}{\ln j + 1}\right)} \ .$$

It is conjectured that this theorem still holds with $c \geq 1.86$, but to my knowledge this problem remains open. The conjecture (and its motivation) appears with further improvements on this result in Williamson et al. (2000, Lemma 16).

Theorem 5.51 (Entropy number of a diagonal operator). *Let (η_i) be a non-increasing sequence defining a diagonal operator $T : \ell^p \to \ell^p$ by*

$$Tv = (\eta_1 v_1, \eta_2 v_2, \cdots, \eta_i v_i, \cdots) \ ,$$

for $1 \leq p \leq \infty$. Then, for all $j \in \mathbb{N}$,

$$\sup_{i \in \mathbb{N}} \left[j^{\frac{-1}{i}} (\sigma_1 \sigma_2 \cdots \sigma_i)^{\frac{1}{i}} \right] \leq \mathscr{N}_n(T) \leq 6 \sup_{i \in \mathbb{N}} \left[j^{\frac{-1}{i}} (\sigma_1 \sigma_2 \cdots \sigma_i)^{\frac{1}{i}} \right] \ .$$

We now have

$$\begin{aligned} \mathscr{N}_{j_1 j_2}(S_{A^{-1}\Phi(Q)} A K) &= K \mathscr{N}_{j_1 j_2}(S_{A^{-1}\Phi(Q)} A) \\ &\leq K \mathscr{N}_{j_1}(S_{A^{-1}\Phi(Q)}) \mathscr{N}_{j_2}(A) \end{aligned}$$

for any $j_1, j_2 \in \mathbb{N}$. For $j_1 = 1$, we note that by the construction of A, $\|S_{A^{-1}\Phi(Q)}\| \leq 1$, so that we obtain

$$\mathscr{N}_j\left(S_{A^{-1}\Phi(Q)} A K\right) \leq K \mathscr{N}_j(A) \ .$$

In this result, we only used Theorem 5.49. Alternatively, applying Theorem 5.50, one obtains

$$\mathscr{N}_{j_1 j_2}(S_{A^{-1}\Phi(Q)} A K) \leq cK \sqrt{(\ln j_1 + 1)^{-1} \ln\left(1 + \frac{n}{\ln j_1 + 1}\right)} \mathscr{N}_{j_2}(A) \ .$$

Finally, we need bounds on $\mathscr{N}_j(A)$. In order to ensure that A meets the requirements of the construction, we have that A is a diagonal operator defined by a sequence

$$\left(C_{\mathscr{K}} \left\| \left(\frac{\sqrt{\lambda_i}}{a_i} \right) \right\|_{\ell^2} a_i \right) \ ,$$

where the choice of (a_i) can be used to optimize Theorem 5.51 (under the restriction that the norm in the sequence is finite).

The optimal choice of (a_i) was found in Guo et al. (2002). Let

$$i^\star = \min \left\{ i \in \mathbb{N} : \lambda_{i+1} < \left(\frac{\lambda_1 \lambda_2 \cdots \lambda_i}{n^2} \right)^{\frac{1}{i}} \right\} \ .$$

Using $i^\star$, we define the optimal $(a_i^\star)$ by

$$a_i^\star = \begin{cases} \sqrt{\lambda_i}, & i \leq i^\star \\ \left(\frac{\sqrt{\lambda_1 \lambda_2 \cdots \lambda_{i^\star}}}{j} \right)^{\frac{1}{i^\star}} & i > i^\star \end{cases} \ .$$

For this choice of $(a_j^\star)$, Theorem 5.51 gives

$$\mathscr{N}_j(A) \leq 6 C_{\mathscr{K}} \sqrt{ i^\star \left(\frac{\lambda_1 \lambda_2 \cdots \lambda_{i^\star}}{j^2} \right)^{\frac{1}{i^\star}} + \sum_{i'=i^\star+1}^{\infty} \lambda_{i'} } \ .$$

We still have some problems with applying this result. The most obvious one is that it is not clear how to obtain the eigenvalues and eigenfunctions of the kernel. We will return to this problem in a moment. However, a more subtle problem is the assumption that $T_{\mathscr{K}}$ is compact. In most cases, the kernels employed in practice have non-compact support so that this assumption may be violated. The key to solving this problem is that the integral operator corresponding to a *ν-periodic extension* of a translation-invariant kernel is compact. Suppose the kernel $\mathscr{K}$ is defined for inputs in $\mathbb{R}^N$. Let $G_\nu = \{\eta \in \mathbb{R}_N : \big(\forall i \in [1:N] : \eta^{(i)} = j\nu, j \in \mathbb{Z}\big)\}$ be a regular grid of points with an interval of ν. Let $G_\nu(\eta) = \eta + G_\nu = \{\eta + \eta' : \eta' \in G_\nu\}$. Then the ν-periodic extension of $\mathscr{K}$ is defined by

$$\mathscr{K}_\nu(\eta, \eta') = \mathscr{K}'_\nu(\eta - \eta') = \sum_{\eta^\star \in G_\nu(\eta - \eta')} \mathscr{K}'(\eta^\star) \ ,$$

where $\mathscr{K}'$ is defined by

$$\mathscr{K}(\eta,\eta') = \mathscr{K}'(\eta - \eta') \ .$$

By selecting ν large enough for the expected application, the extended kernel $\mathscr{K}_\nu$ behaves exactly like $\mathscr{K}$ on the training sample and future points. On the other hand, the larger ν is, the worse the bounds obtained are.

The Fourier series representation of a kernel is important because the coefficients in the series correspond to the eigenvalues of the associated integral operator. If we denote the Fourier transform of $\mathscr{K}'$ by $\mathscr{F}[\mathscr{K}']$, and $\mathscr{K}'_\nu$ exists, then $\mathscr{K}'_\nu$ has an expansion as a Fourier series, since it is periodic. Furthermore, if $\mathscr{K}'$ is rotationally invariant, the Fourier series is related to $\mathscr{F}[\mathscr{K}'']$, where

$$\mathscr{K}'(\eta) = \mathscr{K}''(\|\eta\|) \ .$$

In N dimensions, the resulting Fourier series has eigenvalues indexed by $\mathbb{Z}^N$, and it is rotationally invariant. Specifically, we have (see Williamson et al., 1998b, Remark 13) that the eigenvalues of $T_{\mathscr{K}_\nu}$ can be represented by

$$\lambda_v = (2\pi)^{\frac{N}{2}} \mathscr{F}[\mathscr{K}''] \left(\frac{2\pi\|v\|}{\nu} \right) \ ,$$

where $v \in \mathbb{Z}^N$. Furthermore, we have that $C_{\mathscr{K}_\nu} = (\frac{2}{\nu})^{\frac{N}{2}}$. We see that the spectrum of $\mathscr{K}'_\nu$ is degenerate: if $\|v_1\| = \|v_2\|$, $\lambda_{v_1} = \lambda_{v_2}$. It turns out the result in Theorem 5.51 can be tightened slightly to take advantage of this, if we can calculate the multiplicity of each eigenvalue. An approach for doing this is presented in Williamson et al. (1998b).

Finally, the Fourier transform of $\mathscr{K}''$ can be obtained by using the Hankel transform:

Definition 5.9 (Hankel transform). The Hankel transform of order i of $\phi(v)$, $\mathscr{H}_i[\phi]$, is defined by

$$\mathscr{H}_i[\phi](\eta) = \int_0^\infty v\phi(v) \mathscr{J}_i(\eta v)\, dv \ ,$$

where $\mathscr{J}_i$ is the Bessel function of the first kind of order i,

$$\mathscr{J}_i(v) = \left(\frac{v}{2}\right)^i \sum_{j=0}^{\infty} \frac{(-1)^j v^{2j}}{2^{2j} j! \mathit{\Gamma}(j+i+1)} \ ,$$

and here $\mathit{\Gamma}$ is the Gamma function.

Generally, we have

$$F\left[\mathscr{K}''\right](v) = v^{-\left(\frac{N}{2}-1\right)}\mathscr{H}_{\frac{N}{2}-1}\left[\eta^{\frac{N}{2}-1}\mathscr{K}''(\eta)\right](v) \ .$$

Finally, one should sort the eigenvalues obtained using this method in decreasing order. However, in practice, the resulting Fourier transform exhibits (hopefully fast) decay with increasing $\|v\|$ so that we generally have $\lambda_{v_1} \leq \lambda_{v_2}$ when $\|v_1\| \geq \|v_2\|$.

Example 5.32. Let $\mathscr{K}$ be the N-dimensional Gaussian kernel,

$$\mathscr{K}(\eta, \eta') = \frac{\exp\left(\frac{-\|\eta-\eta'\|^2}{2\sigma^2}\right)}{\sigma^N} \ ,$$

so that

$$\mathscr{K}''(\eta) = \frac{\exp\left(\frac{-\eta^2}{2\sigma^2}\right)}{\sigma^N} \ .$$

Then

$$F[\mathscr{K}''](v) = v^{-\left(\frac{N}{2}-1\right)}\sigma^{-N}\mathscr{H}_{\frac{N}{2}-1}\left[\eta^{\frac{N}{2}-1}\exp\left(\frac{-\eta^2}{2\sigma^2}\right)\right](v) \ ,$$

where $\mathscr{H}_i[\phi]$ denotes the Hankel transform of order i of $\phi(\eta)$. The Fourier transform finally reduces to

$$\mathscr{F}[\mathscr{K}''](v) = \exp\left(\frac{-v^2\sigma^2}{2}\right) \ ,$$

which decays exponentially with $|v|$.

This yields the eigenvalues

$$\lambda_v = (2\pi)^{\frac{N}{2}}\exp\left(\frac{-\left(\frac{2\pi\|v\|}{\nu}\right)^2\sigma^2}{2}\right) \ .$$

□

We conclude this section with two comments. First, note that the bounds obtained are independent of the data, so effectively bound $\mathcal{N}_{\infty,n}$, as well. Second, we mention that one can also obtain bounds for covering numbers employing other metrics except the infinity metric. This is done by employing bounds on the entropy number of the identity operator between different spaces. For more details on this, the reader is referred to Williamson et al. (2000).

Chapter 6

Data-dependent bounds

The covering number approach discussed in the previous chapter is limited in that the double sample bound is applied uniformly over the cover of the loss class. This approach seems to disregard the potential benefits of the weighted approach employed using a “prior” in the countable case, as described in Section 5.4.

The chapter begins by presenting a method of sensibly introducing a “prior” in conjunction with covering number bounds. This method is closely related to methods proposed for model selection in the statistical learning community: structural risk minimization (SRM), complexity regularization, and the method of sieves.

Generalizing the method leads naturally to data-dependent bounds and related data-dependent methods for model selection. The first major data-dependent bounds, based on the concept of *sample compression*, were introduced as early as 1986 by Littlestone and Warmuth (1986).

The idea of using the unthresholded output of thresholded classifiers was the second source of data-dependent bounds. We present a framework, known as the luckiness framework, which motivates both of these types of bounds. However, note that these approaches were developed before the luckiness framework was formulated. As such, the direct arguments employed in the original work often provide better results than a direct application of the framework.

6.1 Combining a "prior" with covering number bounds

The covering number approach of the previous chapter generally provides a uniform bound on a measure of deviation over the entire loss class. On the other hand, the Occam's razor method of Section 5.4 suggests employing a "prior" to obtain better bounds for functions which are more likely to be of interest.

The Occam's razor method breaks down in the uncountable case because one can not sensibly define a "prior" over such a loss class.

A natural attempt to combine the Occam's razor method and covering number methods is the following: for an uncountable loss class, partition the class into a countable number of segments. Employ a "prior" over the segments, and obtain bounds for each segment employing the covering number method. This yields uniform bounds on the measure of deviation for each segment, but the bound will depend on which segment, and the "prior" probability associated with the segment.

Example 6.1. Consider the case of obtaining bounds for thresholded classifiers based on polynomials of arbitrary degree. Without a restriction on the degree of the polynomial, this class has infinite VC dimension.

Suppose one decomposes this class based on the degree of the polynomial being thresholded. Then, one can use covering number methods to obtain bounds for thresholded degree K polynomials (which has a finite VC dimension). Combining these bounds for all K yields bounds for all polynomials. Note, however, that for large enough K these bounds will generally be trivial. □

In the example above, generally a bound obtained for degree K polynomials will hold for all polynomials of degree K or less. As a result, instead of a partition of a loss class, one could consider any sequence of loss classes, and obtain bounds for each element of the sequence. Applying the Occam's razor method then yields bounds for any decision rule whose associated loss lies in the union of these loss classes. Thus, let $(\mathcal{F}_i)$ be a sequence of subsets of a loss class $\mathcal{F}$, with $\lim_{i\to\infty} \bigcup_{j=0}^{i} \mathcal{F}_i = \mathcal{F}$. Typically, but not necessarily, the $\mathcal{F}_i$ form a nested hierarchy of increasing complexity, according to some measure of complexity (e.g. VC dimension). This is a popular, but slightly weaker, formulation of this approach. Such a sequence of nested loss classes is sometimes called a *sieve*.

Example 6.2. Consider a loss class $\mathcal{F}$ corresponding to the class of all decision trees for binary classification. Define $\mathcal{F}_i$ as the loss class corresponding to those decision trees with i or less leaves.

In this case, we have

$$\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}_i \subseteq \cdots \subseteq \mathcal{F} \ .$$

If we can obtain a bound for each such $\mathcal{F}_i$, we can then combine them with the Occam's razor method. □

Even if a uniform "prior" is used with this approach, the resulting bounds will not generally be uniform over the entire loss class. This is because the covering numbers of each loss class will be different. In the case of a nested loss class sequence, the covering numbers will typically increase as the size of the class under consideration increases, resulting in looser bounds.

Thus, what distinguishes these training sample bounds from the bounds described earlier is that these bounds no longer have to depend on the complexity of the entire loss class, but merely on that of some $\mathcal{F}_i$ to which the loss corresponding to the decision rule under consideration belongs.

6.1.1 ERM and SRM

A classical approach to selecting a decision rule is empirical risk minimization (ERM). Simply put, the ERM algorithm proposes selecting a decision rule with the lowest training risk. Thus ERM and M-estimation (van de

Geer, 2000) are intimately related. An alternative view which shall be useful to us is that ERM minimizes an upper bound on true risk derived from a uniform bound on the upper absolute deviation over the entire loss class. For a specific confidence level, the bound on the error is simply the training risk plus some value obtained from the bound for each decision rule. Because the bound is uniform over the entire class, the value added is the same for each decision rule.

The examples above are useful when one expects to select a simpler classifier, such as one based on a low degree polynomial, or a tree with few leaves. These types of choices are very common, with most algorithms avoiding overly complex decision rules to avoid overfitting, particularly in the case of noisy data. In the case of the uniform "prior" over the sequence of loss subclasses ($\mathcal{F}_i$), the examples also provide a motivation to prefer algorithms selecting decision rules based on lower degree polynomials or trees with less leaves if good empirical results can be obtained: the smaller covering numbers corresponding to these smaller loss subclasses show that deviations between empirical and actual probabilities are smaller for the decision rules in these subclasses, so the empirical estimates are more reliable. This indicates a trade-off between empirical performance and the size of the covering number bounds of the loss subclasses. Indeed, this is the basis of structural risk minimization, first proposed as *ordered risk minimization* by Vapnik and Chervonenkis in 1974 (Devroye et al., 1996).

SRM is a generalization of ERM in that it also chooses a decision rule by minimizing an upper bound on true risk obtained from a bound on upper absolute deviation. However, the bound is no longer uniform over the entire loss class: instead, the bound is obtained using the Occam's razor method on a nested sequence of loss classes. A result of this approach is that simple decision rules with larger training risk may be preferred to more complex decision rules with smaller training risk. This can be seen as a form of *complexity regularization*, a general approach to selecting decision rules in which a trade-off is made between empirical performance and some penalty term related to the "complexity" of the decision rule (Devroye et al., 1996).

SRM thus attempts to find a good trade-off between empirical risk and a measure of decision rule complexity, based on risk bounds. However, for a decision class $\mathcal{W}$, the covering number bounds considered in the last chapter use the complexity of $\mathcal{F}_{\mathcal{W}}$ as a measure of decision rule complexity for every decision rule in $\mathcal{W}$. SRM overcomes this by refining the measure of decision rule complexity used for each decision rule by considering the complexity of various subsets of $\mathcal{F}_{\mathcal{W}}$.

We noted earlier that the sequence of nested hypothesis classes is sometimes called a sieve. This seems to be due to the parallels between the relationship of SRM to ERM and the relationship of the method of sieves (Grenander, 1981) to ML estimation. For a discussion of statistical aspects of the method of sieves and penalization/regularization methods, see Shen and Wang (1997).

Example 6.3. Consider the problem of finding a polynomial model for a regression problem, with no restrictions on the degree of the polynomial. How does one determine the degree of overfitting for one polynomial over another? Since higher degree polynomials are more complex and are more liable to overfit, making use of this knowledge in our selection of an hypothesis seems sensible.[135]

One approach to this example would be to set $\mathcal{F}_i$ as the loss class corresponding to the degree i polynomials of the decision class $\mathcal{W}$, and apply the SRM methodology with some "prior".[136]

Finally, one selects the hypothesis with the lowest bound on the risk. This selection then also implicitly provides a solution to the *model selection* problem of selecting the degree of the polynomial to fit. □

6.2 Generalizing the "prior"

Note that the "prior" used for the Occam's razor method needs to be specified before observing the data. That is, in traditional SRM, the collection of $\mathcal{F}_i$'s is fixed for all possible samples, allowing us to apply the training

[135] Another good example would be the number and degree of interaction terms to include in an analysis of variance model.

[136] The prior will need to decay in some way to sum to one, however.

sample bounds of the previous sections.

In this section, we introduce a generalized "prior" concept which has already been implicitly employed for realizable margin bounds, allowing us to circumvent this obstacle. This leads naturally to the concept of *data-dependent* SRM.

Recall that we employed the union bound (or margin unification lemma) in order to allow us to select a desired margin for the margin bounds we derived, after observing the training data. If we consider specifically the realizable margin bound, we note that for a specific choice of γ, the bound provides a bound only when $e_S(h) = 0$ and the margin on all the points in S are at least γ. Clearly, the hypotheses that this bound can be applied to can not generally be specified simply as a subset of $\mathcal{H}$ regardless of the sample. Instead, the subset $\mathcal{H}(S)$ the bound applies to is a function of the sample (and can thus be seen as a random variable). This bound is then made uniform by the margin unification lemma in Theorem 5.23, which can be seen as an application of the Occam's razor method.

As such, the realizable margin bound can be seen as an application of the Occam's razor method with a generalized "prior" to a margin bound with a fixed γ.

In the approach discussed in the previous section, we obtained bounds over $\mathcal{F}$ by considering various subsets of $\mathcal{F}$ and employing the union bound. Furthermore, if the union of the subsets considered was not $\mathcal{F}$ we did not obtain bounds for the elements of $\mathcal{F}$ not in the union.

The generalized "prior" takes this idea further: instead of considering subsets of $\mathcal{F}$, we consider subsets of $\mathcal{Z}^n \times \mathcal{F}$ for some n. If we can obtain a bound for each such subset, we can employ the Occam's razor method to obtain a bound which applies to all the subsets simultaneously.

Example 6.4. In the case of realizable margin bounds above, for any $\gamma \in \mathbb{R}$, we could consider the sets

$$R_i = \left\{ (S, g_h) \in \mathcal{Z}^m \times \mathcal{H} : (e_S(h, L_{\frac{1}{i}}) = 0) \right\} \quad .$$

Observe that here the R_i form a nested sequence of subsets in $\mathcal{Z}^m \times \mathcal{H}$.

Applying the Occam's razor method to bounds obtained for such R_i then provides a bound *for any hypothesis h which achieves a positive margin on all points in the sample S.* This approach does not provide bounds for other hypotheses, however. Furthermore, we note that an hypothesis can be bounded by this result on observing one sample, but not be bounded if a different sample is observed.

Note that this example presented a decomposition of the hypothesis class instead of the loss class. This is valid by the same reasoning used to derive margin bounds originally. □

Example 6.5. Consider the sets

$$R_i = \left\{ (S, w) \in \mathcal{Z}^m \times \mathcal{W} : r_S(w) \leq \frac{i}{m} \right\} \ .$$

Once again, the R_i form a nested sequence. One can obtain a realistic case bound for each R_i, including a realizable case bound for R_0. Combining these bounds with a "prior" on $[0 : m]$ yields a data-dependent bound *on every decision rule in $\mathcal{W}$.* □

6.2.1 Data-dependent SRM

In general, data-dependent SRM (DD-SRM) refers to an extension of SRM which optimises bounds based on this generalized "prior" approach, rather than traditional "priors". Such an algorithm is easily seen to be a generalization of SRM by noting that choosing the R_i independently of the sample with the generalized "prior" approach yields the traditional "prior" approach.

Note that in both examples above, the sets R_i with small i represented (typically small) subsets of $\mathcal{Z}^m \times \mathcal{F}$ with seemingly desirable properties (large margin/small empirical error). Other potentially desirable properties can be considered, but they should be formulated in such a way that bounds taking advantage of the property can be obtained. In general then, we are interested in constructing the sequences R_i such that the most desirable sample-decision rule couples are in the R_i with small i — in other words, the R_i form a hierarchy of decreasing desirability of sample-decision rule couples.

6.3 The luckiness framework

The first reasonably generic approach to data-dependent bounds was the so-called *luckiness framework*, which was developed in Shawe-Taylor et al. (1996) and Shawe-Taylor et al. (1998). The original luckiness framework had two major shortcomings: it was restricted to errors, and furthermore was restricted to the realizable case. However, both of these shortcomings were addressed in the later developments of the algorithmic luckiness framework in Herbrich and Williamson (2002). These frameworks make use of a concept called "probable smoothness" or "ω-smallness" to obtain bounds — functions which are probably smooth can be used as the basis for properties which can be exploited for deriving useful bounds.

Independent work in Gat (1999, 2000a,b) proposed a similar approach which removed the realizable restriction, and was algorithm-dependent. Cannon et al. (2002) provided an alternative version of these results which was not algorithm-dependent, and replaced the cardinality of a class by that of a corresponding cover in the resulting bounds.

The luckiness framework is named for its use of a *luckiness function* for quantifying the desirability of a decision rule w on a sample S. Generally the luckiness function is some mapping $\mathrm{Luck} : \bigcup_{i=1}^{\infty} \mathcal{Z}^i \times \mathcal{W} \to \mathbb{R}$, which should be larger for more desirable combinations of inputs and decision rules.

For an arbitrary n-sample Q, one can then define

$$\mathcal{W}(Q, w) = \{w' \in \mathcal{W} : \mathrm{Luck}(Q, w') \geq \mathrm{Luck}(Q, w)\} \ ,$$

the class of decision rules luckier than w on Q. Generally, the bounds we obtain employ covering numbers, so we consider the covering numbers of $\mathcal{W}(Q, w)$. For any metric d, and any loss function, we have that

$$\mathcal{N}(\gamma, \mathcal{F}_{\mathcal{W}(Q,w)}, d) \leq \mathcal{N}(\gamma, \mathcal{F}_{\mathcal{W}}, d)$$

for all γ. Consider some fixed $\gamma > 0$. Then for any i, we can define the sets

$$R_i(d,\gamma) = \left\{(Q,w) : \mathcal{N}(\gamma, \mathcal{F}_{\mathcal{W}(Q,w)}, d) \leq 2^i\right\} \quad .$$

We could replace the 2^i by any other increasing sequence if we desire. The intuition here is that R_i consists of those sample-decision rule couples for which there are very few luckier decision rules for the corresponding samples. Note that in order for the R_i to be deterministic, the metric employed should not depend on points in $\mathcal{Z}$ except through Q.

If we can obtain bounds for these R_i we can employ the margin bound to obtain data-dependent results. Alternatively, we can employ the margin unification lemma with a little ingenuity.

Let us review the approach we have taken to obtaining bounds on uncountable decision classes so far:

- employ a symmetrization lemma to reduce the problem to an argument over a dual sample;
- reduce this problem to a combinatorial one by a standard argument involving the symmetric group or swapping subgroup;
- reduce the problem to a finite cover, and make an adjustment for the approximation of the cover;
- bound deviations of individual decision rules using a dual sample bound;
- employ the uniform Occam's razor method over the cover to obtain bounds over the cover; and
- take expectations to make the probability statement unconditional.

If we wish to apply this methodology in the current setting, we will need to provide some modifications of previous results. The following modified symmetrization lemmas are useful in this context.[137]. They are based on Her-

[137] The reader may remember that we also considered modified symmetrization lemmas when discussing margin bounds.

brich and Williamson (2002, Lemmas 20 and 21)[138], and correspond to (a realizable version of) Theorem 5.6 and (5.7) respectively.

Theorem 6.1 (Data-dependent realizable symmetrization lemma). *Consider any distribution D, and any decision class $\mathcal{W}$. Let $\mathscr{E}$ be a predicate on $\bigcup_{i=1}^{\infty} \mathcal{Z}^i$. Then*

$$\begin{aligned}&\mathbb{P}_{S\sim D^m}\{\exists w \in \mathcal{W} : (e_D(w) > \epsilon) \wedge (e_S(w) = 0) \wedge \mathscr{E}(S)\}\\ &< \left[1 - e^{-\epsilon u}\right]^{-1} \mathbb{P}_{S\oplus P\sim D^{m+u}} \left\{\exists w \in \mathcal{W} : \left(e_P(w) > \frac{\epsilon}{2}\right) \wedge (e_S(w) = 0) \wedge \mathscr{E}(S)\right\} \ .\end{aligned}$$

Theorem 6.2 (Data-dependent symmetrization lemma for regular deviation). *Consider any distribution D, and any decision class $\mathcal{W}$. Let $\mathscr{E}$ be a predicate on $\bigcup_{i=1}^{\infty} \mathcal{Z}^i$. Suppose α, β satisfy*

$$\mathbb{P}_{P\sim D^u}\left\{r_D(w) - r_P(w) \leq \alpha\right\} \geq \beta \ .$$

Then

$$\begin{aligned}&\mathbb{P}_{S\sim D^m}\left\{\left(\sup_{w\in\mathcal{W}}\left[r_D(w) - r_S(w)\right] > \epsilon\right) \wedge \mathscr{E}(S)\right\}\\ &\qquad < \beta^{-1}\, \mathbb{P}_{S\oplus P\sim D^{m+u}}\left\{\left(\sup_{w\in\mathcal{W}}\left[r_P(w) - r_S(w)\right] > \epsilon - \alpha\right) \wedge \mathscr{E}(S)\right\} \ .\end{aligned}$$

The forms of these theorems appearing in Herbrich and Williamson (2002) employ $\alpha = \frac{\epsilon}{2}$, and $\beta = 2$, using α_H to obtain a sample size restriction for the results to hold. Theorem 6.1 has many similarities to Theorem 5.16; however, by considering only errors, tighter constants are obtained by employing properties of the binomial distribution. Thus it is not as straightforward to generalize this theorem, beyond the removal of the sample size restriction we have performed.

In order to apply these results usefully in our current context, we need a good way to choose $\mathscr{E}$. For a given choice of i, the idea is to try and specify $\mathscr{E}_i$ so that $\mathscr{E}_i(S)$ implies $(S \oplus P, w) \in R_i(d, \gamma)$, while not making $\mathscr{E}_i$ true unnecessarily. For certain metrics d, and in particular the one we will

[138] Note that the original proof of Lemma 20 was incorrect — the corrections can be found in Herbrich and Williamson (2004).

consider, this can not generally be done, and we must be satisfied with a probability statement. In what follows, suppose we have an $\mathscr{E}_i(S)$ such that

$$\mathbb{P}_{S\oplus P\sim D^{m+u}}\{((S\oplus P,w)\in R_i(d,\gamma))\wedge\overline{\mathscr{E}_i(S)}\}<\delta \ .$$

More generally, we can parametrize $\mathscr{E}_i$ with δ, so that $\mathscr{E}_i(S,\delta)$ implies $(S\oplus P,w)\in R_i(d,\gamma)$ with probability at least $1-\delta$. We will discuss finding such $\mathscr{E}_i$ in more detail later.

Next, we consider (for regular deviation)

$$\mathbb{P}_{S\oplus P\sim D^{m+u}}\left\{\exists w\in\mathcal{W}:\left(\sup_{w\in\mathcal{W}}[r_P(w)-r_S(w)]>\epsilon-\alpha\right)\wedge\mathscr{E}_i(S,\delta')\right\}$$
$$\leq\delta'+\mathbb{P}_{S\oplus P\sim D^{m+u}}\left\{\left(\sup_{w\in\mathcal{W}}[r_P(w)-r_S(w)]>\epsilon-\alpha\right)\wedge((S\oplus P,w)\in R_i(d,\gamma))\right\} \ .$$

Converting this to a probability statement on permutations is unchanged, and we obtain that the probability on the right hand side equals

$$\mathbb{E}_{Q\sim D^{m+u}}\,\mathbb{P}_{\tau\sim\mathrm{Unif}\,S_{m+u}}\left\{\mathscr{E}'(\tau(Q))|Q\right\} \ ,$$

where $\mathscr{E}'(S\oplus P)$ denotes

$$\left(\sup_{w\in\mathcal{W}}[r_P(w)-r_S(w)]>\epsilon-\alpha\right)\wedge((S\oplus P,w)\in R_i(d,\gamma)) \ .$$

By definition, when $(S\oplus P,w)\in R_i(d,\gamma)$ there is a γ-cover $\mathcal{F}'(S\oplus P,w))$ of $\mathcal{F}_{\mathcal{W}(S\oplus P,w)}$ of at most 2^i elements (using the metric d). This leads us to consider replacing $\mathscr{E}'(S\oplus P)$ by $\mathscr{E}^\star(S\oplus P)$, defined by

$$\left(\sup_{f\in\mathcal{F}'(S\oplus P,w)}[\mathbb{E}_{z\sim P}(f(z))-\mathbb{E}_{z\sim S}(f(z))]>\epsilon^\star-\alpha\right) \ .$$

Thus, it seems a sensible choice of the metric d is $d_{1,S\oplus P}$. A uniform bound of $\epsilon^\star-\alpha$ on this regular deviation over the cover elements then implies a corresponding bound on the elements of $\mathcal{W}(S\oplus P,w)$ of $\epsilon^\star-\alpha+\frac{(2m+u)(m+u)}{mu}\gamma$, as discussed just before Theorem 5.17. This leads one to the choice of $\epsilon^\star=\epsilon-\frac{(2m+u)(m+u)}{mu}\gamma$.

Noting that we have a bound on the cover sizes of 2^i, we obtain

$$\begin{aligned}&\mathbb{P}_{S\oplus P\sim D^{m+u}}\left\{\exists w\in\mathcal{W}:\left(\sup_{w\in\mathcal{W}}[r_P(w)-r_S(w)]>\epsilon-\alpha\right)\wedge\mathscr{E}(S)\right\}\\ &\leq\delta'+2^i\,\mathbb{E}_{Q\sim D^{m+u}}\sup_{w\in\mathcal{W}}\mathbb{P}_{\tau\sim\mathrm{Unif}\,S_{m+u}}\left\{\mathscr{E}'_w\left(\tau(Q),\epsilon-\alpha-\frac{(2m+u)(m+u)}{mu}\gamma\right)|Q\right\}\ ,\end{aligned}\tag{6.1}$$

where $\mathscr{E}'_w(S\oplus P,t)$ is defined by

$$r_P(w)-r_S(w)>t\ .$$

The interior probability can be bounded with standard dual sample bounds. We can also replace the symmetric group with the swapping subgroup at this stage and employ double sample bounds.

For instance, applying the dual sample bound for regular deviation of risk in Theorem 5.10, one obtains a bound on (6.1) of

$$\delta'+2^i\exp\left(-2m\left(\frac{\left(\epsilon-\alpha-\frac{(2m+u)(m+u)}{mu}\gamma\right)u}{m+u}\right)^2\right)\ .$$

Suppose that α and β satisfy the conditions of Theorem 6.2. In this case, we have

$$\begin{aligned}&\mathbb{P}_{S\sim D^m}\left\{\left(\sup_{w\in\mathcal{W}}[r_D(w)-r_S(w)]>\epsilon\right)\wedge\mathscr{E}_i(S,\delta')\right\}\\ &\qquad\leq\beta^{-1}\left[\delta'+2^i\exp\left(-2m\left(\frac{\left(\epsilon-\alpha-\frac{(2m+u)(m+u)}{mu}\gamma\right)u}{m+u}\right)^2\right)\right]\ .\end{aligned}$$

Setting this bound to δ and solving for ϵ, one obtains

$$\epsilon=\frac{(2m+u)(m+u)}{mu}\gamma+\alpha+\frac{m+u}{u}\sqrt{\frac{i\ln 2-\ln(\delta\beta-\delta')}{2m}}\ .\tag{6.2}$$

Note that for any choice of m and u, one can find a corresponding i_0 for which $i>i_0\Rightarrow\epsilon\geq 1$, so that the bounds are trivial. One can thus obtain a bound which applies for all $i\leq i_0$ by employing the Occam's razor method with a (typically uniform) "prior" over $[0:i_0]$, since larger values of i are irrelevant.

To summarize, we present the following theorem.

Theorem 6.3 (Data-dependent bound on regular deviation). *Suppose one can find an $\mathscr{E}_i$ such that for all non-negative integers i, and $\delta \in (0,1]$,*

$$\mathbb{P}_{S\oplus P\sim D^{m+u}}\{((S\oplus P,w)\in R_i(d_{1,S\oplus P},\gamma))\wedge \overline{\mathscr{E}_i(S,\delta)}\} < \delta \enspace .$$

Let $\alpha^\star(i)$ be a "prior" on the non-negative integers $\mathbb{N}_0$. Furthermore, let $\alpha(i)$ and $\beta(i)$ satisfy the conditions of Theorem 6.2 for every $i \in \mathbb{N}_0$, and let $0 < \delta'(i) < \delta$. Define

$$N = \{i \in \mathbb{N}_0 : \delta'(i) < \delta\beta(i)\} \enspace .$$

Then

$$\mathbb{P}_{S\sim D^m}\left\{\forall i \in N : \left(\left(\sup_{w\in\mathcal{W}}[r_D(w) - r_S(w)] > \epsilon_i\right) \wedge \mathscr{E}_i(S, \delta'\alpha^\star(i))\right)\right\} \le \delta \enspace ,$$

where

$$\epsilon_i = \frac{(2m+u)(m+u)}{mu}\gamma + \alpha(i) + \frac{m+u}{u}\sqrt{\frac{i\ln 2 - \ln(\alpha^\star(i)[\delta\beta(i) - \delta'(i)])}{2m}} \enspace .$$

6.3.1 ω-smallness and the choice of $\mathscr{E}$

The above theorem is useless unless we can find a useful form for $\mathscr{E}_i(S,\delta)$. To understand what the predicate $\mathscr{E}$ should encode, note that we need

$$\mathbb{P}_{S\oplus P\sim D^{m+u}}\{((S\oplus P,w)\in R_i(d_{1,S\oplus P},\gamma))\wedge \overline{\mathscr{E}_i(S,\delta)}\} < \delta \enspace ,$$

and that by definition $(S\oplus P, w) \in R_i(d_{1,S\oplus P},\gamma)$ when

$$\mathcal{N}_{1,S\oplus P}(\gamma, \mathcal{F}_{\mathcal{W}(S\oplus P,w)}) \le 2^i \enspace .$$

Thus, we desire a statement $\mathscr{E}_i(S,\delta)$ which will give us information about the covering numbers of a class of lucky decision rules on a dual sample $S \oplus P$ with first portion S. Furthermore, we need this statement to hold for all distributions in $\mathcal{S}$. This motivates the following definition.

Definition 6.1 (ω-smallness of a luckiness function). A luckiness function Luck is called ω-small w.r.t. L and $u(m)$ at scale γ if there are functions $\omega : \mathbb{R}\times\mathbb{N}\times(0,1] \to \mathbb{N}$ and $u : \mathbb{N}\to\mathbb{N}$ such that for all $m \in \mathbb{N}$, $\delta \in (0,1]$, and all $Q \in \mathcal{S}$,

$$\mathbb{P}_{S\oplus P\sim Q^{m+u(m)}}\left\{\exists w \in \mathcal{W} : \mathcal{N}_{1,S\oplus P}(\gamma, \mathcal{F}_{\mathcal{W}(S\oplus P,w)}) > \omega(\mathrm{Luck}(w,S), m, \delta)\right\} < \delta \enspace ,$$

where the loss class is induced by the loss function L.

If $u(m) = m$, we omit the expression "w.r.t. $u(m)$".

A study of this definition shows that if Luck is ω-small w.r.t. L and $u(m)$ at scale γ, the following choice of $\mathscr{E}_i(S,\delta)$ satisfies the requirements of Theorem 6.3:

$$\mathscr{E}_i(S,\delta) = [\omega(\text{Luck}(w,S),m,\delta) \leq 2^i] \ .$$

Further consideration of the definition reveals that a luckiness function will usually only be ω-small if luckiness of a decision rule on any portion of a sample is in some way indicative of its luckiness on the entire sample. In this sense, luckiness functions must be well-behaved.

A similar property was also referred to as *probable smoothness* in Shawe-Taylor et al. (1998). As pointed out in Herbrich (2002, Remark 4.23), the approach we have described above could be called "vanilla luckiness", because it is based on a restricted version of the probable smoothness concept in Shawe-Taylor et al. (1998). In the terminolgy of that paper, ω-smallness specifies probable smoothness with $\eta = 0$. As a result, some luckiness functions shown to be probably smooth in Shawe-Taylor et al. (1998) are not necessarily ω-small. Particularly, the luckiness function used to derive margin bounds in Shawe-Taylor et al. (1998) is shown to be probably smooth, but not shown to be ω-small. It is not clear to me whether the full probable smoothness concept from Shawe-Taylor et al. (1998) could be sensibly employed to obtain bounds in the non-realizable case, as has been done with ω-smallness.[139]

In order to combine Theorem 6.3 and Definition 6.1, a few restrictions are needed. First, we restrict the "prior" $\alpha^\star(i)$ to be uniform over some set A. In addition, we restrict the $\delta'(i)$ to all equal a constant δ'. This leads to the following result.

[139] Furthermore, we note that the margin bounds in Shawe-Taylor et al. (1998, Section 4) (employing the fat-shattering dimension), which are developed directly and served as an inspiration for the luckiness framework, do not fall into the results of the framework. A similar observation holds for the the margin bounds derived for linear classifiers employing covering numbers directly in Schölkopf et al. (1999a,b), Williamson et al. (1999), which seem to develop from those margin bounds. These bounds are interesting since they employ ideas similar to those in Section 5.9.5, but obtain bounds in terms of the eigenvalues of the Gram matrix obtained from the kernel $\mathscr{K}$ on the sample S. The resulting bounds on the covering numbers are then clearly data-dependent.

Theorem 6.4 (Luckiness bound). *Let* Luck *be ω-small w.r.t. L and $u(m)$ at scale γ. Let $\alpha^\star(i)$ be a uniform "prior" on a finite set $A \subset \mathbb{N}_0$. Furthermore, let $\alpha(i)$ and $\beta(i)$ satisfy the conditions of Theorem 6.2 for every $i \in A$, and let $0 < \delta' < \delta$. Let*

$$N = \{i \in A : \delta' < \delta\beta(i)\} \ .$$

Define

$$\phi(w) = \left\lceil \log_2 \omega\left(\mathrm{Luck}(w,S), m, \frac{\delta'}{|A|}\right)\right\rceil \ ,$$

and let $\mathcal{W}' = \{w \in \mathcal{W} : \phi(w) \in N\}$. Then,

$$\mathbb{P}_{S\sim D^m}\{\exists w \in \mathcal{W}' : r_D(w) - r_S(w) > \epsilon_w\} \leq \delta \ ,$$

where

$$\epsilon_w = \frac{(2m+u(m))(m+u(m))}{mu(m)}\gamma + \alpha(\phi(w)) + \frac{m+u(m)}{u(m)}\sqrt{\frac{\frac{\phi(w)}{\log_2 e} + \ln|A| - \ln(\delta\beta(\phi(w)) - \delta')}{2m}} \ .$$

A similar theorem can be obtained in the realizable case, based on the symmetrization lemma of Theorem 6.1 and the realizable dual sample bound in Theorem 5.14.

Let us briefly compare this result to those in Shawe-Taylor et al. (1998) and Herbrich and Williamson (2002). Shawe-Taylor et al. (1998, Theorem 6.4) is a bound on the realizable case, and employs the more general concept of probable smoothness. When η in that bound is set to zero, the result seems to correspond to the result that would be obtained here with further restrictions of a sufficiently small scale, $m = u$, and $\delta' = \frac{\delta}{4}$.

Herbrich and Williamson (2002, Theorems 8 and 9) (and the uniform versions below them on p. 164), on which these results were based, are algorithm-specific in that the resulting bounds only apply to the decision rule selected by the algorithm. We shall pay more attention to this approach in what follows. The bounds there are closely related to the ones we have here, and if the same approach taken there was taken to obtain algorithm-independent bounds, (essentially) a special case of our results would be obtained up to a constant factor of $\sqrt{\log_2 e}$. Specifically, their bounds are obtained with the following setting: $u(m) = m$, $\delta' = \frac{\delta}{4}$, $\alpha(w) = \frac{\epsilon_w}{2}$ (note that ϵ_w depends on w only through $\phi(w)$), $\beta(i) = 2$ and a uniform bound on $[0 : \frac{m}{2} - 1]$. The condition required on α and β is maintained by a sample size restriction, which is obtained using α_H. We note in passing that the uniform bound obtained from their Theorem 9 could be strengthened somewhat by restricting the uniform prior to $[0, \frac{m}{8} - 1]$, since for larger values of d the bound is trivial.

Besides a realizable version of the data-dependent bounds, there is much potential here for considering data-dependent bounds on other measures of deviation. In order to do this, appropriate modified symmetrization lemmas will be needed, and potentially new smallness concepts when working with the P-H ν-deviation, for example. The only work I am aware of in this regard is a data-dependent bound on relative deviation in Andonova Jaeger (2005), which employs other techniques instead of the luckiness framework. Furthermore, it is not clear how much benefit can be obtained by our formulation allowing $m \neq u$. Note however, that some of the improvements of standard covering number bounds, such as those in Devroye (1982), Shawe-Taylor et al. (1993), came by employing results with $m \neq u$.

An obvious candidate luckiness function is risk on a sample. However, a result on ω-smallness of risk essentially states we can control risk on the double sample in terms of the risk on the first sample. However, this has, in a way, been the problem under consideration for much of this work. Furthermore, applying such a result here would only lead to a potential slackening of the ω-smallness result originally employed.

Generally, then, the luckiness function can be seen as an encoding of our intuition of what behaviour on a sample is desirable for a decision rule. However, the ω-smallness requirement constrains our choices to those which are in some sense stable, preventing one from tailoring the bound too much towards the observed sample. Informally, one gets the sense that the luckiness function needs to be more stable than empirical risk based on the comments in the previous paragraph.

Example 6.6. We begin with a fairly trivial example of an ω-small luckiness function. Consider a decision class $\mathcal{W}$ with finite pseudodimension.

Consider the luckiness function $\mathrm{Luck}(w, Q) = -\operatorname{pdim}(\mathcal{W})$, and any choice of $u(m)$. Then, by Theorem 5.27,

$$\begin{aligned}\mathcal{N}_{1,S\oplus P}(\gamma, \mathcal{W}(S \oplus P, w)) &\leq \mathcal{N}_{1,S\oplus P}(\gamma, \mathcal{W}) \\ &\leq \mathcal{N}_{\infty,m+u(m)}(\gamma, \mathcal{W}) \\ &\leq \sum_{i=1}^{\operatorname{pdim}(\mathcal{W})} \binom{m+u(m)}{i} \left\lfloor \frac{1}{2\gamma} \right\rfloor^{i} .\end{aligned}$$

Since this always holds, we have that for all $\delta \in (0, 1]$,

$$\mathbb{P}_{S\oplus P\sim Q^{m+u(m)}} \left\{ \exists w \in \mathcal{W} : \begin{array}{r} \mathcal{N}_{1,S\oplus P}\left(\gamma, \mathcal{W}(S \oplus P, w)\right) \\ > \omega_u\left(-\operatorname{pdim}(\mathcal{W}), m, \delta\right) \end{array} \right\} < \delta \ ,$$

where we define

$$\omega_u(t, m, \delta) = \sum_{i=1}^{-t} \binom{m+u(m)}{i} \left\lfloor \frac{1}{2\gamma} \right\rfloor^{i}$$

for all δ.

Thus, for any choice of the function u, $\mathrm{Luck}(w, Q) = -\operatorname{pdim}(\mathcal{W})$ is ω_u-small w.r.t. the identity loss. However, the resulting bounds will not incorporate any improvements based on the hypothesis or data under consideration.

A similar analyis can be performed using the fat-shattering dimension by employing Theorem 5.26. □

The following example provides a data-dependent luckiness function.

Example 6.7. Consider the restriction of the zero-one functions in $\mathcal{W}$ to a sample Q, $\mathcal{W}_{|Q}$. Clearly, since the domain of these functions is Q, we have $\mathrm{VC}(\mathcal{W}_{|Q}) \leq |Q|$. $\mathrm{VC}(\mathcal{W}_{|Q})$ is known as the *empirical VC dimension on* Q.

Define the luckiness function $\mathrm{Luck}(w, Q) = -\mathrm{VC}(\mathcal{W}_{|Q})$. Then Shawe-Taylor et al. (1998, Proposition 7.7) show that Luck is ω-small w.r.t. the identity loss at sufficiently small[140] scale γ, where

$$\omega(t, m, \delta) = \left(\frac{-em}{1.54(t + \ln \delta)} \right)^{-3.08(t+\ln \delta)} .$$

□

[140] Since the functions are zero-one, for a small enough scale, the covering numbers involved equal the shatter coefficients.

Note that when the loss function is well-behaved, the results relating covering numbers of loss classes and decision classes can be used to infer ω'-smallness of a luckiness function with respect to such a loss function from ω-smallness with respect to the identity loss.

6.4 Algorithmic luckiness

If one intends to use a specific algorithm Θ to select a decision rule, it makes sense to try and tailor the generalized "prior" to suit Θ.

Ideally, for a specific sample Q, one would like to place all the generalized "prior" mass on the decision rule that will be selected based on Q. However, the technical conditions of the luckiness framework prevent one from doing this directly.

The algorithmic luckiness framework solves this in a similar manner to the luckiness framework we have outlined in the previous sections. A generalized "prior" is obtained by considering a set of hypotheses which is in some sense desirable, based on an *algorithmic luckiness function.* An ω-smallness constraint on the algorithmic luckiness function then leads to bounds similar to those of the luckiness framework.

There are two major differences between this framework and the luckiness framework. First, in this case, luck is assigned to samples, rather than to pairs of samples and decision rules. Secondly, the only decision rules which are assigned non-zero "prior" weight for a given sample Q are the decision rules which would be selected by Θ on a subsample of Q (of some fixed length).

We begin with the concept of an algorithmic luckiness function, which is a mapping $\mathrm{Aluck}_{\Theta} : \bigcup_{i=1}^{\infty} \mathcal{Z}^i \to \mathrm{I\!R}$. The analysis done here is for a generic algorithm Θ, which can be seen as a subscript to the sets and functions defined. We shall omit the subscript throughout for notational convenience, however.

Analogously to $\mathcal{W}(Q, w)$ in the luckiness framework, we can define $\mathcal{W}(Q, m)$ for an $(m+u)$-sample Q:

$$\mathcal{W}(Q, m) = \{\Theta(S_{\tau(Q)}) : \tau \in S_{m+u}, \mathrm{Aluck}(S_{\tau(Q)}) \geq \mathrm{Aluck}(S_Q)\} \ ,$$

where, as before, S_Q denotes the first m elements of the $(m+u)$-sample Q. Note that since many algorithms are sensitive to the order of the data presented to them, Aluck and these derived concepts are not generally permutation-invariant. Considering the $m+u$-sample Q, some of the $(m+u)!$ permutations are such that Aluck is larger on the first S points of the permuted sample than it is on the original sample. $\mathcal{W}(Q, w)$ then consists of the decision rules outputted by Θ on these truncated, permuted samples.

These sets $\mathcal{W}(Q, m)$ are again the basis of the sets R_i, defined by

$$R_i(d, \gamma, m) = \{Q : \mathcal{N}(\gamma, \mathcal{F}_{\mathcal{W}(Q,m)}, d) \leq 2^i\}$$

for some loss function L.

With this setup, if we can (with some probability) relate the truth of an expression $\mathscr{E}_i(S, \delta)$ to whether $S \oplus P$ is in $R_i(d, \gamma, m)$, we can apply Theorem 6.3 to obtain bounds which are dependent on Θ. The key to doing this is an analog of ω-smallness for algorithmic luckiness functions. If this condition on the algorithmic luckiness function holds, we can use the size of $\omega(\mathrm{Aluck}(\cdot))$ as the basis of such an $\mathscr{E}_i$.

Definition 6.2 (ω-smallness of an algorithmic luckiness function). An algorithmic luckiness function Aluck_Θ is called ω-small w.r.t. L and $u(m)$ at scale γ if there are functions $\omega : \mathbb{R} \times \mathbb{N} \times (0,1] \to \mathbb{N}$ and $u : \mathbb{N} \to \mathbb{N}$ such that for all $m \in \mathbb{N}$, $\delta \in (0,1]$, and all $Q \in \mathcal{S}$,

$$\mathbb{P}_{S \oplus P \sim Q^{m+u(m)}}\{\mathcal{N}_{1, S \oplus P}(\gamma, \mathcal{F}_{\mathcal{W}(S \oplus P, m)}) > \omega(\mathrm{Aluck}_\Theta(S), m, \delta)\} < \delta \ ,$$

where the loss class is induced by the loss function L.

If $u(m) = m$, we omit the expression "w.r.t. $u(m)$".

Thus, if Aluck_Θ is ω-small w.r.t. $u(m)$, $\omega(\mathrm{Aluck}_\Theta(S), m, \delta')$ controls the growth of the loss class corresponding to decision rules selected by "lucky"

m-subsamples of $S \oplus P$. As such, an appropriate choice for $\mathscr{E}_i(S, \delta)$ is

$$[\omega(\mathrm{Aluck}_{\Theta}(w, S), m, \delta) \leq 2^i] \ .$$

Finally, Theorem 6.3 gives the following algorithmic luckiness theorem, which is an algorithmic luckiness analog to the result in Theorem 6.4.

Theorem 6.5 (Algorithmic luckiness bound). *Let Θ be an algorithm. Let* Aluck_{Θ} *be ω-small w.r.t. L and $u(m)$ at scale γ. Let $\alpha^{\star}(i)$ be a uniform "prior" on a finite set $A \subset \mathbb{N}_0$. Furthermore, let $\alpha(i)$ and $\beta(i)$ satisfy the conditions of Theorem 6.2 (with $u = u(m)$) for every $i \in A$, and let $0 < \delta' < \delta$. Let*

$$N = \{i \in A : \delta' < \delta\beta(i)\} \ .$$

Define

$$\phi(w) = \left\lceil \log_2 \omega \left(\mathrm{Aluck}_{\Theta}(S), m, \frac{\delta'}{|A|} \right) \right\rceil \ ,$$

and let $\mathcal{W}' = \{w \in \mathcal{W} : \phi(w) \in N\}$.

Then,

$$\mathbb{P}_{S \sim D^m}\{(\Theta(S) \in \mathcal{W}') \wedge (r_D(\Theta(S)) - r_S(\Theta(S)) > \epsilon(S))\} \leq \delta \ ,$$

where

$$\begin{aligned} \epsilon(S) = {} & \frac{(2m + u(m))(m + u(m))}{mu(m)}\gamma + \alpha(\phi(w)) \\ & + \frac{m + u(m)}{u(m)} \sqrt{\frac{\frac{\phi(w)}{\log_2 e} + \ln|A| - \ln(\delta\beta(\phi(w)) - \delta')}{2m}} \ . \end{aligned}$$

An important difference between this result and Theorem 6.4 is that this bound does not apply to any decision rule other than that selected by the algorithm.[141] This loss of generality is the price paid to obtain the potentially tighter bounds by using algorithmic luckiness specific to Θ, rather than a more generic luckiness concept. This result is a (slightly stronger) more general form of the uniform version of the main result in Herbrich and Williamson (2002).

Once again, a realizable version can be formulated, and there is potential for future development with other measures of deviation.

[141] Strictly speaking, it can be formulated to apply to a handful of other decision rules, but this is not practically useful.

6.5 Sample compression bounds

In this section, we shall consider the application of data- and algorithm-dependent bounds to algorithms which can be framed as effective *compression schemes*. The pioneers in bounds for such algorithms were Nick Littlestone, Manfred Warmuth, and Sally Floyd: the two most influential sources seem to be Floyd and Warmuth (1995), Littlestone and Warmuth (1986).

In our context, a compression scheme for an algorithm Θ has two components: a compression function, and a (permutation-invariant) reconstruction function. The compression function is given a sample Q, and returns the compressed subsample Q'. The reconstruction function takes a labelled sample and returns a decision rule in $\mathcal{W}$. These two components are related in that the reconstruction function applied to Q' should yield $\Theta(Q)$. In other words, the compression function identifies the sample elements necessary to specify $\Theta(Q)$, and the reconstruction function can use those 'essential' sample elements to obtain $\Theta(Q)$.

Example 6.8. Consider a classification problem where the concept class corresponding to $\mathcal{W}$ consists of (axis-parallel) rectangles in $\mathbb{R}^2$.

Consider the algorithm Θ which selects the smallest rectangle $R(S)$ containing all the points in a sample S labelled one, and predicts $I(x \in R(S))$.

Consider the subsample S' consisting of the point labelled one with the smallest first feature, the point labelled one with the largest first feature, and the points labelled one with the smallest and largest second feature. Clearly S' consists of at most four points. Furthermore, it should also be clear that $\Theta(S') = \Theta(S)$.

Thus, the function choosing S' as described is a compression function corresponding to the reconstruction function Θ. □

Example 6.9. The previous example can be generalized to higher dimensions. In $\mathbb{R}^N$, the compressed subsample contains at most $2N$ elements.

It is interesting to note that the VC dimension of the class of axis-parallel parallelipipeds in $\mathbb{R}^N$ is also $2N$. □

It is very often the case that the reconstruction function in a compression scheme is the original algorithm. Note that for any permutation-invariant algorithm Θ, using the identity compression function and the algorithm Θ as a reconstruction function constitutes a (trivial) compression scheme.

In the examples above there a fixed upper limit K on the size of the compressed subsample available a priori. In such a case, the algorithm is said to have a compression scheme of size K.[142]

The basic intuition of sample compression bounds is that when the compression function returns a small subsample, the selected decision rule is in a sense simple, and thus less likely to exhibit a large deviation between empirical and actual error.

The original bounds in Littlestone and Warmuth (1986) were restricted to the realizable error case, that is, where $e_S(\Theta(S)) = 0$. The following is a classical result of this type.

Theorem 6.6 (Theorem 5 of Floyd and Warmuth, 1995). *Consider an algorithm Θ with an associated compression scheme. Let S' denote the compressed subsample associated with a sample S. Then*

$$\mathbb{P}_{S\sim D^m}\{(|S'| \leq K)\wedge(e_D(\Theta(S)) > \epsilon)\wedge(e_S(\Theta(S)) = 0)\} \leq \sum_{i=1}^{K}\binom{m}{i}(1-\epsilon)^{m-i} \ .$$

The authors further state that this can be extended to the realistic case using Chernoff bounds. Such an extended result is presented in Blum and Langford (2003, Theorem 11).

One shortcoming of this approach is that it can only be applied if $|S'| \leq K$ with K chosen a priori. This is not a problem if Θ has a compression scheme of size K, since then the bound will always apply (assuming a consistent decision rule can be found). Furthermore, it is desirable that the bound should be data-dependent in the sense that it is tighter when the compressed subsample is smaller.

For an algorithm Θ with an associated compression scheme, define the algorithmic luckiness function $\mathrm{Aluck}_\Theta(Q)$ as the negated size of the compressed

[142] K can generally be a function of m.

subsample, $-|Q'|$. The following result is a straightforward generalization of a result in Herbrich and Williamson (2002) to a more general choice of $u(m)$.

Theorem 6.7 (Algorithmic luckiness function for sample compression). *Let* Aluck *be defined as above. Then for any $\gamma > 0$, any $u(m)$, and any loss function L,* Aluck *is ω_u-small, where*

$$\omega_u(t, m, \delta) = \sum_{i=1}^{-t} \binom{m + u(m)}{i} \ .$$

Note the similarity of the form of ω_u to the coefficient in the bound of Theorem 6.6. One can use this result in conjunction with Theorem 6.5 to obtain bounds which are algorithm- and data-dependent. Note that the resulting bounds are no longer restricted to errors, and they hold generally, instead of only in the realizable or realistic cases.

6.5.1 MDL

Sample compression bounds are closely related to the ideas of algorithmic complexity, which were first proposed in the 1960's. Based on these ideas, the minimum message length (Wallace and Boulton, 1968) and minimum description length (MDL) (Rissanen, 1978) methodologies were proposed for selecting decision rules.

The MDL methodology suggests the following technique for selecting a decision rule from a class on the basis of a training sample: if one adds the minimal size (in bits) needed to code an algorithm to implement a decision rule (known as the *coding length* of the decision rule[143]), to the number of bits needed to store the points in the training sample inconsistent with the decision rule, one obtains a value for each decision rule. The MDL principle proposes selecting the decision rule minimizing this value.

[143] This length is defined as the length of a binary computer program — the type of computer is irrelevant. Solomonoff and Kolmogorov were working with these concepts before Rissanen proposed the MDL principle — more details are in Section 4.6 of Vapnik (1995).

A little reflection makes clear that the MDL principle represents a trade-off between decision rule complexity (the coding length) and accuracy on the training sample (storing errors). In fact, the MDL principle is an implementation of SRM, not as originally formulated by Vapnik, but in the more generic form presented in this work.[144] Note that the coding length of decision rules is independent of the data. The name derives from the fact that the sum described for a given decision rule is the amount of data that needs to be communicated to enable the receiver to calculate the responses from the predictors for each point in the training sample by employing that decision rule.[145] To calculate the responses, one evaluates the decision rule on the predictors of each training point. These outputs are then used, except for those points transmitted in the message, where we use the outputs provided in the message.

As is so often the case in practice, one doesn't usually know all the values desired — in this case, one generally does not know the coding length of the hypotheses, so one must employ upper bounds (such as the actual length of *some* program to implement an hypothesis). If the upper bound used is independent of the specific hypothesis (e.g. if the length of code implementing the entire hypothesis class were to be used), this approximation of the MDL principle would propose minimizing the number of errors, and one once again obtains ERM.

The sample compression approach we discussed above can be seen as using a data-dependent bound on the coding length of hypotheses in order to obtain bounds. This is because, for a given compression scheme, an implementation of the compression and reconstruction functions is common to all decision rules. For each decision rule, the same program can then be used, except that different compression subsamples must be provided. Using this upper bound on the coding length of each decision rule, the MDL methodology indicates that one should prefer decision rules with small compression subsamples, if their empirical risks are identical. As a result, employing the bounds

[144] Vapnik (1995, Section 4.6) discusses the relationship between MDL and Vapnik's SRM in more detail.

[145] Note that this approach assumes that training samples do not contain points with identical inputs but differing outputs.

obtained with the sample compression approach above for model selection can be seen as a kind of data-dependent MDL.

Chapter 7

Tightening bounds with concentration inequalities

Until now, we have only made use of the concentration inequalities for sums of random variables, detailed in the first half of Chapter 4, in obtaining our results.

Tight concentration inequalities allow one to use observations of variables instead of their underlying mean when making calculations, at some small penalty. The bounds we have studied until now used a form of union bound to combine such an argument for each function in the class. The resulting bounds are often expressed in terms of expected covering numbers, or in the case of chaining, a type of entropy integral. These values can then be upper bounded employing concepts such as the VC or fat-shattering dimension.

However, concentration inequalities provide an alternative approach. If it can be shown that the covering number for a given sample is concentrated about its mean, a suitable concentration inequality may be able to provide a good high-confidence bound on the expected covering number. This approach is presented in Section 7.1

An alternative approach manages to avoid the union bound argument entirely, by employing Rademacher complexities and concentration inequalities. The resulting bounds, known as Rademacher bounds, are discussed in

Section 7.2.

7.1 Covering numbers and concentration inequalities

Ledoux and Talagrand made significant advances in the understanding of concentration of measure in the late 1980s and throughout the 1990s — for the development of their theory, see Ledoux (1996, 1997, 2001), Ledoux and Talagrand (1991), Talagrand (1988, 1989, 1994, 1995, 1996b,c,d). Other players made significant contributions leading to the advanced concentration results provided in Sections 4.10 and 4.11.

In this section, we show that a concentration inequality for self-bounding functions allows us to obtain tight control of the expected effective class $\mathbb{E}_{Q\sim D^n} |Q_{\mathcal{W}}|$ simply by considering the size of the effective class for a *specific* n-sample Q, viz. $|Q_{\mathcal{W}}|$. The approach is based on Boucheron et al. (1999).

Theorem 7.1. *For $b \geq 2$, $H(Q) = \log_b |Q_{\mathcal{W}}|$ is a self-bounding function.*

To prove this, we will need to use Han's inequality (Boucheron et al., 1999, Han, 1978). Recall that the *Shannon entropy* (base b)[146] of a discrete random variable V with distribution P is

$$\mathrm{Ent}_b(V) = -\sum_{v\in\operatorname{supp} P} \mathbb{P}_{V\sim P}\{V = v\} \log_b \mathbb{P}_{V\sim P}\{V = v\} \ .$$

It is known (see, for example, MacKay, 2003, Section 2.4) that the uniform distribution maximizes the Shannon entropy. Han's inequality relates the entropy of a vector-valued variable to the entropy of the n subvectors each obtained by removing a component from the vector:

Lemma 7.1. *Let $V_1, \cdots, V_n$ be discrete random variables. Let $V_{\setminus i}$ denote $(V_1, \cdots, V_{i-1}, V_{i+1}, \cdots, V_n)$. Then*

$$\mathrm{Ent}_b(V_1, \cdots, V_n) \leq \frac{1}{n-1} \sum_{i=1}^{n} \mathrm{Ent}_b(V_{\setminus i}) \ .$$

[146] The constant b is a slight nuisance, but it allows us to use similar approaches for realizable bounds, where we may be interested in using $\log_2$. Furthermore, for polychotomous classification, we may want to use some other $b > 2$.

Note that we can rewrite this result as

$$\sum_{i=1}^{n}[\mathrm{Ent}_b(V_1,\cdots,V_n)-\mathrm{Ent}_b(V_{\setminus i})]\leq \mathrm{Ent}_b(V_1,\cdots,V_n)\ \ . \tag{7.1}$$

Proof. (of Theorem 7.1) Define $Q_{\setminus i}$ as Q with the i-th element removed. Clearly $\left|\left(Q_{\setminus i}\right)_{\mathcal{W}}\right| \leq |Q_{\mathcal{W}}| \leq 2\left|\left(Q_{\setminus i}\right)_{\mathcal{W}}\right|$, so that $0 \leq H(Q)-H(Q_{\setminus i}) \leq \log_b 2 \leq 1$.

Let Q have n elements. Consider a random variable $(V_1,\cdots,V_n)$ which has a uniform distribution over $Q_{\mathcal{W}}$. Now,[147]

$$\begin{aligned} H(Q) &= \log_b |Q_{\mathcal{W}}| \\ &= -\sum_{i=1}^{|Q_{\mathcal{W}}|}\frac{1}{|Q_{\mathcal{W}}|}\log_b\frac{1}{|Q_{\mathcal{W}}|} \\ &= \mathrm{Ent}_b(V_1,\cdots,V_n)\ \ . \end{aligned} \tag{7.2}$$

Defining $V_{\setminus i}$ from V as in Han's inequality, we note that $V_{\setminus i}$ is selected from some distribution on $\left(Q_{\setminus i}\right)_{\mathcal{W}}$ (not necessarily, uniform, however). But, since entropy is maximized on the uniform distribution, we have that

$$H(Q_{\setminus i}) \geq \mathrm{Ent}_b(V_{\setminus i})\ \ .$$

Thus, applying (7.1), we have

$$\begin{aligned}\sum_{i=1}^{n}[H(Q)-H(Q_{\setminus i})] &\leq \sum_{i=1}^{n}[\mathrm{Ent}_b(V_1,\cdots,V_n)-\mathrm{Ent}_b(V_{\setminus i})] \\ &\leq \mathrm{Ent}_b(V_1,\cdots,V_n) \\ &= H(Q)\ \ ,\end{aligned}$$

completing the proof. □

It follows from this theorem that one can apply (4.16) to bound $\mathbb{E}_{Q\sim D^n} H(Q)$. This yields the result, for any $\epsilon \geq 0$,

$$\mathbb{P}_{Q\sim D^n}\{H(Q)\leq \mathbb{E}_{Q\sim D^n} H(Q)-\epsilon\} \leq \exp\left(-\,\mathbb{E}_{Q\sim D^n} H(Q)h\left(\frac{-\epsilon}{\mathbb{E}_{Q\sim D^n} H(Q)}\right)\right)\ \ ,$$

and the right hand side can be further relaxed to

$$\exp\left(\frac{-\epsilon^2}{2\,\mathbb{E}_{Q\sim D^n} H(Q)}\right)$$

[147] The logarithm of a covering number is often referred to as the entropy of the class. This result serves to explain why.

— see Boucheron et al. (1999, Corollary 1). Setting this equal to δ_1 yields

$$\mathbb{P}_{Q\sim D^n}\left\{H(Q) \le \mathbb{E}_{Q\sim D^n} H(Q) - \sqrt{-2\,\mathbb{E}_{Q\sim D^n} H(Q) \ln \delta_1}\right\} \le \delta_1 \ ,$$

so that with probability at least $1-\delta_1$,

$$\mathbb{E}_{Q\sim D^n} H(Q) \le H(Q) + \sqrt{-2\,\mathbb{E}_{Q\sim D^n} H(Q) \ln \delta_1} \ .$$

This result is undesirable however, due to the occurence of $\mathbb{E}_{Q\sim D^n} H(Q)$ on the right hand side.

Boucheron et al. (1999) suggest the following approach to avoiding this impasse. Setting

$$\epsilon = \frac{1}{2}\,\mathbb{E}_{Q\sim D^n} H(Q) + v \ ,$$

one obtains

$$\mathbb{P}_{Q\sim D^n}\left\{\mathbb{E}_{Q\sim D^n} H(Q) > 2H(Q) + 2v\right\} \le \exp\left(\frac{-\left(\frac{1}{2}\,\mathbb{E}_{Q\sim D^n} H(Q) + v\right)^2}{2\,\mathbb{E}_{Q\sim D^n} H(Q)}\right) \ .$$

Expanding the square, and using

$$\frac{1}{2}\,\mathbb{E}_{Q\sim D^n} H(Q) + v \ge 0 \ ,$$

one can bound the right hand side by e^{-v}. We thus have

$$\mathbb{E}_{Q\sim D^n} H(Q) \le 2H(Q) - 2\ln \delta_1$$

with probability at least $1-\delta_1$.

We now present a generalization of this approach to allow smaller factors than 2 in front of $H(Q)$. Set $\epsilon = \kappa\,\mathbb{E}_{Q\sim D^n} H(Q) + v$, for $\kappa \in (0,1)$. This leads to

$$\mathbb{P}_{Q\sim D^n}\left\{\mathbb{E}_{Q\sim D^n} H(Q) > \frac{1}{1-\kappa}(H(Q)+v)\right\} \le \exp\left(\frac{-\left(\kappa\,\mathbb{E}_{Q\sim D^n} H(Q) + v\right)^2}{2\,\mathbb{E}_{Q\sim D^n} H(Q)}\right) \ .$$

Expanding the square, the exponent on the right is

$$-\frac{v^2}{2\,\mathbb{E}_{Q\sim D^n} H(Q)} - v\kappa - \frac{\kappa^2}{2}\,\mathbb{E}_{Q\sim D^n} H(Q) \ .$$

Our approach is to find a factor $\digamma(\kappa)$ such that this exponent never exceeds $-\digamma(\kappa)v$. This will be the case when the quadratic equation

$$\frac{v^2}{2\,\mathbb{E}_{Q\sim D^n} H(Q)} + v(\kappa - \digamma(\kappa)) + \frac{\kappa^2}{2}\,\mathbb{E}_{Q\sim D^n} H(Q) = 0$$

has zero or one solution. Examining the discriminant of the equation, we show that this requires

$$(\digamma(\kappa))^2 - 2\digamma(\kappa)\kappa \leq 0 \ .$$

The tightest bound results from the largest choice of $\digamma$, so we can use $\digamma(\kappa) = 2\kappa$. Note that when $\kappa = \frac{1}{2}$, $\digamma(\kappa) = 1$ so that this generalizes the result above. We thus obtain

$$\mathbb{P}_{Q\sim D^n}\left\{\mathbb{E}_{Q\sim D^n} H(Q) > \frac{1}{1-\kappa}[H(Q) + v]\right\} \leq \exp(-2\kappa v) \ ,$$

so that with probability at least $1 - \delta_1$,

$$\mathbb{E}_{Q\sim D^n} H(Q) \leq \frac{1}{1-\kappa}\left[H(Q) - \frac{\ln \delta_1}{2\kappa}\right] \ . \tag{7.3}$$

Note that this relaxation method applies for general self-bounding functions, not just $H(Q)$. The first part of the following theorem follows.

Theorem 7.2. *Let $\vartheta(Q)$ be a self-bounding function defining a statistic E, and $\kappa \in (0, 1)$. Then for $v > -\kappa\,\mathbb{E}\,E$,*

$$\mathbb{P}\left\{\mathbb{E}\,E > \frac{1}{1-\kappa}[E + v]\right\} \leq \exp(-2\kappa v) \ ,$$

so that for $\delta < \exp(2\kappa^2\,\mathbb{E}\,E)$,

$$\mathbb{E}_{Q\sim D^n} E \leq \frac{1}{1-\kappa}\left[E - \frac{\ln \delta}{2\kappa}\right] \ , \tag{7.4}$$

with probability at least $1 - \delta$.

Furthermore, if $\mathbb{E}\,E \geq 0$,

$$\mathbb{E}\,E \leq \inf_{\kappa\in(0,1)}\left(\frac{1}{1-\kappa}\left[E - \frac{\ln \delta}{2\kappa}\right]\right) \ . \tag{7.5}$$

The more useful second part is obtained from Bartlett et al. (2004, Lemma A.4).

We should compare the result for $H(Q)$ to what can be obtained from McDiarmid's inequality. It is easy to verify that $-H(Q)$ satisfies the bounded difference assumption with $c = 1$, so that from Theorem 4.6, we obtain

$$\mathbb{E}_{Q\sim D^n} H(Q) < H(Q) + \sqrt{\frac{-n \ln \delta}{2}}$$

with probability at least $1-\delta$. For small sample sizes, this may be a beneficial alternative. In what follows, we shall (somewhat arbitrarily) use the result from Theorem 7.2, however.

Next, we note that most bounds are in terms of $\ln \mathbb{E}_{Q\sim D^n} |Q_{\mathcal{W}}|$, instead of

$$\mathbb{E}_{Q\sim D^n} H(Q) = \mathbb{E}_{Q\sim D^n} \ln |Q_{\mathcal{W}}| \ .$$

Fortunately, these quantities can be related. In fact, due to (4.13) in Theorem 4.14, we have

$$\mathbb{E}_{Q\sim D^n}[e^{\lambda(H(Q)-\mathbb{E}_{Q\sim D^n} H(Q))}] \le \exp((e^\lambda - \lambda - 1)\,\mathbb{E}_{Q\sim D^n} H(Q)) \ ,$$

for all $\lambda \ge 0$. Setting $\lambda = \ln b$, we obtain

$$\mathbb{E}_{Q\sim D^n}[\exp(\ln b(H(Q) - \mathbb{E}_{Q\sim D^n} H(Q)))] \le \exp((b - \ln b - 1)\,\mathbb{E}_{Q\sim D^n} H(Q)) \ ,$$

which leads to

$$\mathbb{E}_{Q\sim D^n} b^{H(Q)} \le \exp((b-1)\,\mathbb{E}_{Q\sim D^n} H(Q)) \ .$$

Noting the definition of $H(Q)$, we then have that

$$\log_b \mathbb{E}_{Q\sim D^n} |Q_{\mathcal{W}}| \le \frac{b-1}{\ln b} \mathbb{E}_{Q\sim D^n} H(Q) \ .$$

As an example, we apply this result to the bound in (5.38) (for error). Taking the limit as $\gamma \to 0$, we obtain

$$\mathbb{P}_{S\sim D^m} \left\{ \sup_{w\in\mathcal{W}} [e_D(w) - e_S(w)] > 4 \left[\sqrt{\frac{2[\ln 4\,\mathbb{E}_{Q\sim D^m} |Q_{\mathcal{W}}| - \ln \delta_2]}{m}} \right] \right\} < \delta_2 \ .$$

In this case, we have $b = e$, so with probability at least $1 - (\delta_1 + \delta_2)$,

$$\sup_{w\in\mathcal{W}} [e_D(w) - e_S(w)]$$

does not exceed

$$4\left\lceil\sqrt{\frac{2\left[(e-1)\inf_{\kappa\in(0,1)}\left[\frac{1}{1-\kappa}\left(\ln|S_{\mathcal{W}}|-\frac{\ln\delta_1}{2\kappa}\right)\right]-\ln\frac{\delta_2}{4}\right]}{m}}\right\rceil \ .$$

Note that in this result, we can specifically use $|S_{\mathcal{W}}|$, the size of $\mathcal{W}$ restricted to the *actual sample observed*.

Using the bound of (5.38) was convenient because we had employed the random subsample lemma, so that the covering number we were considering was over an m-sample. More generally, however, we would like to be able to obtain a bound on $\mathbb{E}_{Q\sim D^{m+u}} H(Q)$ when Q is an $(m+u)$-sample. Since we don't actually observe the shadow sample, this problem is a little more difficult. Generally, it is not difficult to see that

$$|(S\oplus P)_{\mathcal{W}}|\le|S_{\mathcal{W}}||P_{\mathcal{W}}| \ .$$

Taking expectations and logarithms in the case $m=u$, one obtains that

$$\ln\mathbb{E}_{S\oplus P\sim D^{2m}}|(S\oplus P)_{\mathcal{W}}|\le 2\ln\mathbb{E}_{S\sim D^m}|S_{\mathcal{W}}| \ ,$$

which one can then combine with the results above in order to improve bounds obtained without employing the random subsample lemma.

The following result allows us to obtain bounds for $m\neq u$ — when $m=u$, applying it to $H(Q)$ yields a result similar to the one above.

Theorem 7.3. *Suppose a sequence of functions (ϑ_n) is such that $\vartheta_n:\mathcal{E}^n\to\mathbb{R}$ is symmetric and self-bounding with respect to ϑ_{n-1} for every n.*

Then, for any distribution τ on $\mathcal{E}$,

$$\mathbb{E}_{X\sim\tau^{m+u}}\vartheta_{m+u}(X)\le\frac{m+u}{m}\mathbb{E}_{X\sim\tau^m}\vartheta_m(X) \ .$$

The proof is a straightforward generalization of that used for Philips (2005, Lemma 3.14).

It also seems reasonable that *semi-supervised learning* approaches may allow one to use unlabelled examples to obtain even better empirical estimates.

Finally, we note that the results obtained above for the concentration of $H(Q)$ are based on a definition for a fixed decision class $\mathcal{W}$, so that the approach outlined here does not apply to covers of data-dependent classes.

However, the results obtained are attractive, since the traditional supremum bound on the size of the effective function class has been tightened, without requiring knowledge of the distribution D. However, an important question in practice is how feasible it is to calculate (or bound) the size of the empirically observed effective function class.

A potentially useful result in this direction is that $|Q_{\mathcal{W}}| \leq (n+1)^{\mathrm{VC}(Q_{\mathcal{W}})}$ — see, for example, Bartlett et al. (2002, p.15). However, finding the empirical VC dimension of $\mathcal{W}$ on Q is not generally easy.[148] This also limits its applicability in the luckiness framework — remember that the negated empirical VC dimension is an ω-small luckiness function (Example 6.7). However, a rather general method to bound it probabilistically (by using Bernstein's inequality) was presented in Williamson et al. (1999). The method is based on evaluating the capability of the decision class to model noise, so it is closely related to the Rademacher bounds we discuss in Section 7.2.

7.2 Rademacher bounds

The training sample bounds we have considered so far have provided a bound on the maximal deviation between empirical and true risks over a (potentially random) function class by effectively employing the union bound over a number of representative functions from the class. A function of the dimension measure or covering number of the class appeared as a complexity penalty to the empirical error when constructing upper bounds on the true error. This has an intuitive appeal, since the size of the cover of the hypothesis class gives an indication of the richness of the class, and thus, hopefully,

[148] As an aside, it is interesting to note that the empirical VC dimension of $\mathcal{W}$ is an example of a *configuration function*, a class of self-bounding functions, so that the empirical VC dimension is concentrated about its mean (Boucheron et al., 1999, Talagrand, 1995). The empirical fat-shattering dimension is also self-bounding. It is also clear that Theorem 7.3 applies to both of these concepts.

its likelihood of overfitting.

Rademacher bounds are the pre-eminent representative of an entirely different approach to bounding maximal deviation which has been developed in the past decade. Rademacher bounds generally make use of a number of concentration-of-measure results to derive bounds without employing the union bound over elements of the hypothesis class. The key to this approach is the realization that the maximal deviation between empirical and true risk exhibits strong concentration. Indeed, Example 4.2 shows that

$$\mathbb{P}_{S\sim D^m}\left\{V \geq \mathbb{E}_{S\sim D^m} V + \epsilon\right\} \leq e^{-2\epsilon^2 m} \quad ,$$

where $V = \vartheta(S) = \sup_{w\in\mathcal{W}}[r_S(w) - r_D(w)]$, thanks to the bounded difference inequality of Theorem 4.6.

Setting $\delta = \exp\left(-2\epsilon^2 m\right)$, it follows that with probability at least $1-\delta$,

$$\sup_{w\in\mathcal{W}}[r_D(w) - r_S(w)] < \mathbb{E}_{S\sim D^m} \sup_{w\in\mathcal{W}}[r_D(w) - r_S(w)] + \sqrt{\frac{\ln\frac{1}{\delta}}{2m}} \quad , \tag{7.6}$$

with similar bounds for the lower and two-sided regular deviation.

This leads one to search for bounds on the *expected* maximal deviation of empirical and true risk. The focus of Rademacher bounds and related bounds is thus to directly bound the expectation on the right hand side of this inequality.

7.2.1 The basic bound

The core of Rademacher bounds is the following classical symmetrization result employing Rademacher random variables. We quote it in the following convenient form from Bartlett et al. (2004, Lemma A.5).

Theorem 7.4 (Symmetrization inequality).

$$\max\left\{\mathbb{E}_{S\sim D^m} \sup_{w\in\mathcal{W}}[r_D(w) - r_S(w)], \mathbb{E}_{S\sim D^m} \sup_{w\in\mathcal{W}}[r_S(w) - r_D(w)]\right\} \\ \leq 2\,\mathbb{E}_{S\sim D^m, \zeta\sim\mathrm{Unif}\{-1,1\}^m}\, \mathcal{R}_S(\mathcal{W}) \quad .$$

Recall that $\mathcal{R}_S(\mathcal{W})$ is the Rademacher penalty of $\mathcal{W}$ for the sample S, which also appeared in the random subsample lemma (Theorem 5.19).

Combining the symmetrization inequality with (7.6) yields that with probability at least $1 - \delta_1$,

$$\sup_{w \in \mathcal{W}} [r_D(w) - r_S(w)] \leq 2\, \mathbb{E}_{S \sim D^m, \zeta \sim \text{Unif}\{-1,1\}^m}\, \mathcal{R}_S(\mathcal{W}) + \sqrt{\frac{\ln \frac{1}{\delta_1}}{2m}} \ , \tag{7.7}$$

with an analagous result holding for lower regular deviation, and a two-sided version holding with $\frac{1}{\delta_1}$ replaced by $\frac{2}{\delta_1}$ on the right hand side.

Our next step is bounding the expected Rademacher penalty. In practice, Rademacher bounds work well because the Rademacher penalty is also highly concentrated around its mean: we begin by noting that the Rademacher penalty satisfies the bounded difference assumption. Modifying any data point (x_i, y_i) and a Rademacher variable σ_i can result in a maximum change to the supremum of $\frac{2}{m}$. Applying the bounded difference inequality (Theorem 4.6) shows that

$$\mathbb{P}_{S \sim D^m, \zeta \sim \text{Unif}\{-1,1\}^m} \left\{ \mathbb{E}_{S \sim D^m, \zeta \sim \text{Unif}\{-1,1\}^m}\, \mathcal{R}_S(\mathcal{W}) - \mathcal{R}_S(\mathcal{W}) \geq \epsilon \right\} \\ \leq \exp\left(\frac{-\epsilon^2 m}{2} \right) \ ,$$

where the expectation of the Rademacher penalty is over all samples and Rademacher variables. A lower bound is identical, so a two-sided bound can be obtained by doubling the right-hand side.

Setting the right hand side here to δ_2, and employing the union bound to combine this with the result in (7.7), we obtain that with probability at least $1 - (\delta_1 + \delta_2)$,

$$\sup_{w \in \mathcal{W}} [r_D(w) - r_S(w)] < 2 \left(\mathcal{R}_S(\mathcal{W}) + \sqrt{\frac{2 \ln \frac{1}{\delta_2}}{m}} \right) + \sqrt{\frac{\ln \frac{1}{\delta_1}}{2m}} \ . \tag{7.8}$$

The same bound applies to

$$\sup_{w \in \mathcal{W}} [r_S(w) - r_D(w)]$$

and a two-sided bound applies with $\frac{1}{\delta_1}$ replaced by $\frac{2}{\delta_1}$.

An alternative approach is to note that if we consider

$$\bar{\mathcal{R}}_S(\mathcal{W}) = \mathbb{E}_{\zeta \sim \mathrm{Unif}\{-1,1\}^m}\, \mathcal{R}_S(\mathcal{W})$$

as a function of S, that it satisfies the bounded difference inequality with a maximum change of $\frac{1}{m}$. This yields

$$\mathbb{P}_{S\sim D^m}\left\{\mathbb{E}_{S\sim D^m, \zeta\sim \mathrm{Unif}\{-1,1\}^m}\, \mathcal{R}_S(\mathcal{W}) - \bar{\mathcal{R}}_S(\mathcal{W}) \geq \epsilon\right\} \leq \exp\left(-2\epsilon^2 m\right) \ .$$

Setting the right hand side to δ_2 and combining this with (7.7), we obtain that with probability at least $1 - (\delta_1 + \delta_2)$,

$$\sup_{w\in\mathcal{W}} [r_D(w) - r_S(w)] < 2\left(\bar{\mathcal{R}}_S(\mathcal{W}) + \sqrt{\frac{\ln \frac{1}{\delta_2}}{2m}}\right) + \sqrt{\frac{\ln \frac{1}{\delta_1}}{2m}} \ . \tag{7.9}$$

We call the average Rademacher penalty $\bar{\mathcal{R}}_S(\mathcal{W})$ the *Rademacher average* of $\mathcal{W}$ on S, and the mean Rademacher average

$$R_m(\mathcal{W}) = \mathbb{E}_{S\sim D^m}\, \bar{\mathcal{R}}_S(\mathcal{W})$$

is called the *Rademacher complexity* of $\mathcal{W}$.[149]

It is also common to present results using two alternative notions of a Rademacher penalty. In many cases, results are stated based on

$$\sup_{w\in\mathcal{W}} \left| \frac{1}{m}\sum_{i=1}^{m} \zeta_i w(x_i) \right| = \mathcal{R}'_S(\mathcal{W}) \ ,$$

which we shall call the *absolute Rademacher penalty* of $\mathcal{W}$ on S. It is clear that $\mathcal{R}'_S(\mathcal{W}) \geq \mathcal{R}_S(\mathcal{W})$. The absolute Rademacher average and absolute Rademacher complexity are defined analogously. In some cases, results are stated using scaled versions of these results, e.g. where the quantities are not normalized by the factor $\frac{1}{m}$ (see Mendelson and Philips, 2003), or the value is doubled (see Bartlett and Mendelson, 2002). Thus, great care must be taken when attempting to apply Rademacher bounds.

It is most common to set $\delta_1 = \delta_2 = \frac{\delta}{2}$ for a desired δ in these bounds. From these bounds, we see that a Rademacher penalty (or average) for a given

[149] Note that the Rademacher complexity is a function of the sample size.

sample can be used as an accurate estimate of the Rademacher complexity of the class, thus obtaining “calculable” bounds. The quotes here refer to the fact that efficiently evaluating the supremum in the definition of a Rademacher penalty is by no means a trivial task, if it can be performed at all. Koltchinskii (2001) illustrates that calculating a Rademacher penalty is in fact equivalent to performing ERM over the class.[150] A result of this is that most Rademacher bounds are formulated for ERM and SRM techniques. However, it is important to note that in general, the ERM problem is difficult, being an NP-hard problem[151]. In practice, efficient methods for ERM exist for a number of function classes. For these classes, Rademacher bounds are relatively easy to evaluate.

Also note that the bound of (7.8) is an example of a data-dependent bound: the bound on the deviation measure depends on the sample S through the Rademacher penalty.

The development of Rademacher bounds was due to research on two fronts: model selection, and attempts to improve the fat-shattering dimension. Specifically, use of the Rademacher penalties arose seemingly independently in Koltchinskii (2001), Bartlett et al. (2000), and Mendelson (2002), with the first of these references bearing the most explicit resemblance to the approach taken here.

Let us briefly consider the interpretation of the Rademacher penalty. We can view the Rademacher penalty as a measure of the ability of the decision class to model, or correlate with, noise: if the expression is very large, the decision class contains decision rules which can match high losses with those i for which $\zeta_i = 1$, and low losses for those i with $\zeta_i = -1$. If the expression is small, the decision class is unable to do this, so it is not as rich or complex. This viewpoint leads naturally to considering the Gaussian penalty of the hypothesis class, where the Rademacher variables are replaced by standard-

[150] The supremum corresponds to a decision rule selected by ERM when the labels are switched according to the Rademacher variables.

[151] This means that generally the problem of minimizing empirical risk is equivalent to a number of difficult problems in statistics and computer science which no-one yet knows how to solve in time polynomial in the sample size.

ized normal variables. It turns out that these quantities are closely related. We define (absolute) Gaussian penalties, averages, and complexities corresponding to the Rademacher concepts by replacing the Rademacher variable ζ_i with a standard normal r.v. g_i, and replacing $\mathcal{R}$ by $\mathscr{G}$.

Theorem 7.5 (Lemma 4 of Bartlett and Mendelson, 2002). *There are constants t_1, t_2 such that*

$$t_1 \mathcal{R}'_m(\mathcal{W}) \leq \mathscr{G}'_m(\mathcal{W}) \leq t_2 \ln m \mathcal{R}'_m(\mathcal{W})$$

for all function classes $\mathcal{W}$ and $m \in \mathbb{N}$.

It follows that bounds in terms of absolute Rademacher complexities imply bounds in terms of absolute Gaussian complexities, and vice versa. Bartlett et al. (2000) have also proposed a so-called *maximum discrepancy* average and corresponding complexity, but this also is intimately related to the absolute Rademacher complexity (Bartlett and Mendelson, 2002, Lemma 3). In what follows, we thus exclusively consider the Rademacher penalty.

7.2.2 Improvements on the basic bound

The first improvement to the basic bound is replacing the two separate applications of the bounded difference inequality, which are combined with the union bound, by a single application of the inequality. Specifically, we consider the statistic V defined by

$$\vartheta(S) = \sup_{w \in \mathcal{W}} [r_D(w) - r_S(w)] - 2\mathcal{R}_S(\mathcal{W}) \ .$$

Since $\mathcal{R}_S(\mathcal{W})$ can change by at most $\frac{2}{m}$ and $\sup_{w \in \mathcal{W}} |r_D(w) - r_S(w)|$ by at most $\frac{1}{m}$ with a change of a single data point, V satisfies the bounded difference assumption with $c = \frac{5}{m}$. This improvement was suggested in Hush and Scovel (1999), and also appears in Koltchinskii (2001). The resulting bound is

$$\mathbb{P}_{S \sim D^m} \left\{ \sup_{w \in \mathcal{W}} [r_D(w) - r_S(w)] < 2\mathcal{R}_S(\mathcal{W}) + \sqrt{\frac{25 \ln \frac{1}{\delta}}{2m}} \right\} \leq \delta \ .$$

Similar bounds can be obtained for lower and two-sided deviation. Note that this approach also removes the problem of selecting δ_1 and δ_2 in the basic bound.

A similar bound can be obtained in terms of $\bar{\mathcal{R}}_S(\mathcal{W})$. In this case,

$$\vartheta(S) = \sup_{w \in \mathcal{W}} [r_D(w) - r_S(w)] - 2\bar{\mathcal{R}}_S(\mathcal{W})$$

satisfies the bounded difference assumption with $c = \frac{3}{m}$, yielding a probabilistic bound on the maximal upper regular deviation of

$$2\bar{\mathcal{R}}_S(\mathcal{W}) + \sqrt{\frac{9 \ln \frac{1}{\delta}}{2m}} \ .$$

This bound is tighter, but it is more difficult in general to obtain $\bar{\mathcal{R}}_S(\mathcal{W})$ than $\mathcal{R}_S(\mathcal{W})$. However, if performing ERM to obtain $\mathcal{R}_S(\mathcal{W})$ is reasonably efficient, one can use Monte Carlo approximation to obtain $\bar{\mathcal{R}}_S(\mathcal{W})$ to any desired accuracy[152] (Bartlett et al., 2002).

We spent some time in Section 4.11 trying to obtain tighter bounds on the maximal deviation between empirical and true means. The final result was the functional Bennett's inequality due to Bousquet, which was a special case of Theorem 4.14.

We have mentioned that McDiarmid's inequality strictly extends Hoeffding's inequality, and that Bousquet's theorem extends Bennett's inequality. When we compared Hoeffding's inequality to Bernstein's and Bennett's inequality we noted that the latter inequalities were tighter when we could control the variance of the function under consideration well. Similarly, Bousquet's theorem for suprema of centred empirical processes only shows its full power when the index set of the process consists of functions for which we can control the variance well (uniformly).

For now, we thus consider bounding the uniform deviation of empirical and true risk over some subclass $\mathcal{W}'$ of $\mathcal{W}$. The idea would be that such bounds

[152] Note that one can also use concentration inequalities to control the inaccuracy of the Monte Carlo estimate.

on a number of classes could then be combined with the union bound. Furthermore, we assume we have some ς satisfying

$$\varsigma^2 \geq \sup_{w \in \mathcal{W}'} \mathbb{V}_{Z \sim D}\, w(Z) \ .$$

Writing $\mathscr{Y} = \sup_{w \in \mathcal{W}'}[r_D(w) - r_S(w)]$, with probability at least $1 - \delta_1$, we have

$$\begin{aligned} \mathscr{Y} &\leq \mathbb{E}_{S \sim D^m} \mathscr{Y} + \frac{\sqrt{-18 \ln \delta_1 [m\varsigma^2 + 2\, \mathbb{E}_{S \sim D^m} \mathscr{Y}]} - \ln \delta_1}{3m} \\ &\leq 2\mathcal{R}_m(\mathcal{W}') + \frac{\sqrt{-18 \ln \delta_1 [m\varsigma^2 + 4\mathcal{R}_m(\mathcal{W}')]} - \ln \delta_1}{3m} \ . \end{aligned} \tag{7.10}$$

This follows by applying (4.18) to $m\mathscr{Y}$, dividing by m, and applying Theorem 7.4.

We next show how to employ a variant of Theorem 4.14 to obtain a bound on $\mathcal{R}_m(\mathcal{W}')$.

Theorem 7.6 (p.25 of Lugosi, 2004). *Let $\mathcal{V}$ be a class of functions from $\mathcal{E}$ into $[-1, 1]$. Let $Q = (\eta_1, \eta_2, \cdots, \eta_n)$ be a sample from $\mathcal{E}$. Then the statistic defined by*

$$n\bar{\mathcal{R}}_Q(\mathcal{V}) = \mathbb{E}_{\zeta \sim \mathrm{Unif}\{-1,1\}^n} \sup_{\phi \in \mathcal{V}} \sum_{i=1}^{n} \zeta_i \phi(\eta_i)$$

is self-bounding.

Theorem 7.7. *Define $\vartheta(S) = 2m\bar{\mathcal{R}}_S(\mathcal{W}')$. Then ϑ is self-bounding.*

Proof. This follows by noting that

$$2\mathcal{W}' - 1 = \{2w - 1 : w \in \mathcal{W}'\}$$

is a class of functions into $[-1, 1]$, so that $m\bar{\mathcal{R}}_S(2\mathcal{W}' - 1) = 2m\bar{\mathcal{R}}_S(\mathcal{W}')$ is self-bounding. □

We obtain a bound by employing Theorem 7.2. Since

$$\begin{aligned} \mathcal{R}_m(\mathcal{W}') &= \mathbb{E}_{S \sim D^m} \bar{\mathcal{R}}_S(\mathcal{W}') \\ &\geq 0 \end{aligned}$$

(this is straightforward to verify), it follows from (7.5) that with probability at least $1 - \delta_2$,

$$2\mathcal{R}_m(\mathcal{W}') \leq \inf_{\kappa \in (0,1)} \left(\frac{1}{1-\kappa} \left[2\bar{\mathcal{R}}_S(\mathcal{W}') - \frac{\ln \delta_2}{2\kappa m} \right] \right) \quad , \tag{7.11}$$

for any $\delta_2, \kappa \in (0, 1)$.

As we did with $H(Q)$ in Section 7.1, we should compare this result to what can be obtained from McDiarmid's inequality. The bounded difference assumption for the Rademacher average holds with constant $\frac{1}{m}$, so that McDiarmid's inequality implies

$$\mathcal{R}_m(\mathcal{W}') \leq \bar{\mathcal{R}}_S(\mathcal{W}') + \sqrt{\frac{-\ln \delta}{2m}}$$

with probability at least $1 - \delta$. For certain combinations of m, δ and κ, as well as the unknown $\bar{\mathcal{R}}_S(\mathcal{W}')$, this bound may be more appropriate. In what follows we shall employ (7.11), however.

Combining this with (7.10), we obtain the following: for any $\kappa, \delta_1, \delta_2 \in (0, 1)$ with $\delta_1 + \delta_2 = \delta < 1$, with probability at least $1 - \delta$,

$$\mathscr{Y} \leq \phi(\kappa, \delta_2, \bar{\mathcal{R}}_S(\mathcal{W}')) + \frac{\sqrt{-18 \ln \delta_1 [m\varsigma^2 + 2\phi(\kappa, \delta_2, \bar{\mathcal{R}}_S(\mathcal{W}'))]} - \ln \delta_1}{3m} \quad , \tag{7.12}$$

where $\phi(\kappa, \delta_2, \bar{\mathcal{R}}_S(\mathcal{W}'))$ denotes the right hand side of (7.11). The same bound applies to the supremum of the lower regular deviation, due to the properties of Rademacher averages we shall establish later in Theorem 7.10. The result here should be compared to Bartlett et al. (2004, Theorem 2.1). Their result differs primarily in that they employ a result similar to Theorem 7.2 for upper bounds to ensure that the bound on $\mathscr{Y}$ only uses $\mathbb{E}_{S \sim D^m} \mathscr{Y}$ once. Furthermore, the result presented here does not employ all the relaxations employed there.

In deriving this result, we have tacitly assumed that the functional Bernstein's inequality in (4.18) employed to obtain (7.10), will outperform the bounded difference inequality if we have good control of the variance. If we compare the bounds for a fixed choice of δ_1, we see that the bounded differences inequality is only better when

$$\sqrt{\frac{-\ln\delta}{2m}} < \frac{\sqrt{-18\ln\delta[m\varsigma^2 + 2\,\mathbb{E}_{S\sim D^m}\mathscr{Y}]} - \ln\delta}{3m} \ .$$

This leads to the inequality

$$\frac{1-\sqrt{\frac{-2\ln\delta}{9m}}}{2} < \sqrt{\varsigma^2 + \frac{2\,\mathbb{E}_{S\sim D^m}\mathscr{Y}}{m}} \ .$$

For most combinations of sample size and confidence level, the left hand side will be slightly less than 0.5. The expression $\frac{2}{m}\,\mathbb{E}_{S\sim D^m}\mathscr{Y}$ is generally negligible, so that the functional Bernstein's inequality will be superior if one can bound ς^2 below 0.25, and will not usually be much worse even if the bound of 0.25 is used.

Ideally, we would like to obtain bounds of the form above for various classes $\mathcal{W}'$, and combine them using the Occam's razor method to obtain bounds on the combined class. It is natural to select the classes based on the control of ς available for the class. The resulting bounds will be tighter for those with tighter control of ς, facilitating the use of SRM with these bounds. In practice, however, it is not clear how to select such classes. Since we typically bound the variance of the loss of a decision rule by its risk, the natural choice of ς for a class $\mathcal{W}'$ is $\sup_{w\in\mathcal{W}'}[r_D(w)(1-r_D(w))]$.[153] However, we do not know the true risks, so that this approach is not feasible. At first sight, it thus seems that the best we can do with this approach is to obtain a uniform bound over $\mathcal{W}$ using $\varsigma = \frac{1}{4}$ (unless the loss function provides a better alternative). However, alternative approaches can indeed yield improved bounds in some situations.

Data-dependent Rademacher bounds

The concept of data-dependent Rademacher bounds has been developed by Petra Philips and Shahar Mendelson in recent years (Mendelson and Philips, 2003, 2004, Philips, 2005). We shall just give a flavour of the approach, but no details, since their results do not provide explicit constants.

[153] Tighter bounds might be available for some general (non-zero-one) loss functions.

Their approach employs data-dependent formulations of a symmetrization lemma and a random subsample lemma, together with a concept they call δ-symmetry. However, instead of employing a cover to bound the Rademacher penalty, they use concentration inequalities.

What is fascinating about their approach is that they show that many major data-dependent approaches to obtaining bounds developed earlier can be cast in their framework: essentially, for their framework to obtain good bounds, two assumptions must be satisfied. The authors show that the assumptions of other approaches are just alternative ways of effectively specifying these two assumptions. Their framework encompasses the traditional covering number bounds, (algorithmic) luckiness bounds, and bounds for compression schemes.

Local Rademacher bounds

The main reason Rademacher bounds have been under the spotlight in recent years is that they are well-suited to obtaining bounds for the ERM algorithm.

This is because the decision rule selected by ERM has certain desirable properties. These properties allow one to apply the Rademacher bound repeatedly to the decision rule, using the error bound from the i-th application to bound the variance for the $(i+1)$-th application. To do this, an Occam's razor method is employed over the various applications of the bound.

This approach seems to have been pioneered in Koltchinskii and Panchenko (2000), with later work summarized in Bartlett et al. (2004) and Koltchinskii (2006) (as well as the various discussions and the rejoinder to the second article).

In the process of developing results for ERM, Bartlett et al. (2004) also obtain results which apply to all functions in the decision class. These results are obtained by applying a bound similar to (7.12) to a modified decision class, which (roughly speaking) weights decision rules in order to control the variance of decision rules with a large variance of future losses.

An additional result then shows how a uniform bound on deviations over the weighted class can yield a bound on true risk for the original decision rules.

The proofs of these results are technically detailed, and the interested reader is referred to the original paper for the details. We will restrict ourselves to introducing the necessary concepts and presenting the most useful results (for more, and more general, such results, see the original article and references therein).

Definition 7.1 (Sub-root function). A function $\phi : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ is called *sub-root* if it is nondecreasing and $\frac{\phi(v)}{\sqrt{v}}$ is nonincreasing in v.

Thus, a sub-root function is a function which grows, but no faster than the square root function. All subroot functions except the zero function are called nontrivial subroot functions. It can be shown that any nontrivial subroot function ϕ is continuous, and the equation $\phi(v) = v$ has a unique solution. Denoting this solution by v_ϕ, one also has that $v \geq v_\phi$ exactly when $v \geq \phi(v)$. We call v_ϕ the *fixed point* of ϕ.

Approximating the fixed point of a sub-root function can in principle be done with a binary search, but it can be shown that iterating the function ϕ actually obtains quicker convergence to the fixed point. To perform this procedure we begin with some $v_0 \geq v_\phi$. Defining $v_{i+1} = \phi(v_i)$, we have that v_i converges very rapidly to v_ϕ: it can be shown that

$$v_\phi \leq v_i \leq \left(\frac{v_0}{v_\phi}\right)^{2^{-i}} v_\phi \ .$$

Definition 7.2 (Star hull). Let A be a set in a real vector space $\mathcal{E}$. Then the *star hull* of A about a point $\eta \in \mathcal{E}$ in the vector space is defined by

$$\text{star}(A, \eta) = \{\eta + v(\eta' - \eta) : v \in [0, 1], \eta' \in A\} \ .$$

When $b = 0$, we simply refer to the *star hull* of A, and write $\text{star}(A) = \text{star}(A, 0)$.

With this background, we can state the following result which flows from Bartlett et al. (2004, Theorem 3.3.2), in a similar fashion to their Corollary 3.5.

Theorem 7.8. *Let* $\varsigma : \mathrm{star}(\mathcal{W}) \to \mathbb{R}^+$ *satisfy*

$$\mathbb{V}_{Z\sim D}\, w(Z) \leq \varsigma(w) \leq \mathbb{E}_{Z\sim D}\, w(Z)$$

and $\varsigma(\digamma w) \leq \digamma^2 \varsigma(w)$ *for all* $w \in \mathrm{star}(\mathcal{W})$ *and* $\digamma \in [0,1]$.

Suppose ϕ *is a sub-root function satisfying*

$$\phi(v) \geq \mathcal{R}_m\left(\{w \in \mathrm{star}(\mathcal{W}) : \varsigma(w) \leq v\}\right)$$

for $v \geq v_\phi$.

Then, for any $K > 1$ *and* $\delta \in (0,1)$,

$$\mathbb{P}_{S\sim D^m}\left\{\exists w \in \mathcal{W} : r_D(w) > \frac{K}{K-1} r_S(w) + 6Kv_\phi - \frac{(\ln\delta)(11+5K)}{m}\right\} \leq \delta \;\; .$$

Thus this result bounds the true risk of a decision rule by a weighted empirical risk plus a complexity term which depends on the sample size, the confidence required, and v_ϕ, the fixed point of a sub-root function upper bounding the Rademacher complexity of low-variance decision rules in the star hull of $\mathcal{W}$. It can be shown that under the conditions of the theorem,

$$\mathcal{R}_m(\{w \in \mathrm{star}(\mathcal{W}) : \varsigma(w) \leq v\})$$

is itself a sub-root function of v, so that it is a desirable choice for ϕ, except that it is not clear how to evaluate it (note that the expression depends on the distribution D).

By showing that functions with low true variance of loss tend to have low empiricial variance of loss (i.e. variance on the sample S), Bartlett et al. (2004) present an alternative choice of ϕ which is a bound on the Rademacher average instead of the Rademacher complexity (for a specific choice of ς). Since this ϕ only upper bounds the Rademacher complexity with a certain probability, we have reduced confidence in the final bound.

Theorem 7.9. *Let* $\delta_1, \delta_2, \delta_3 \in (0,1)$ *satisfy* $\delta_1 + \delta_2 + \delta_3 \leq 1$.

Define $\varsigma_P(w) = \mathbb{E}_{Z\sim P}\, w^2(Z)$. *Then, with probability at least* $1 - \delta_2$,

$$\phi'(v) = 10\mathcal{R}_m(\{w \in \mathrm{star}(\mathcal{W}) : \varsigma_D(w) \leq v\}) + \frac{11\ln\delta_2}{m}$$

satisfies the conditions for ϕ in Theorem 7.8.

Let ϕ'' be a sub-root function satisfying

$$\phi''(v) \geq 20\bar{\mathcal{R}}_S(\{w \in \operatorname{star}(\mathcal{W}) : \varsigma_S(w) \leq 2v\}) + \frac{11 \ln \delta_2 + 20 \ln \delta_3}{m} \quad .$$

Then, with probability at least $1 - \delta_3$, $v_{\phi'} \leq v_{\phi''}$.

Applying Theorem 7.8, we thus have, for any $K > 1$,

$$\mathbb{P}_{S \sim D^m} \left\{ \exists w \in \mathcal{W} : r_D(w) > \frac{K}{K-1} r_S(w) + 6K v_{\phi''} - \frac{(\ln \delta_1)(11 + 5K)}{m} \right\} \\ \leq \delta_1 + \delta_2 + \delta_3 \quad .$$

This result, with all the δ_i equal, corresponds to Bartlett et al. (2004, Corollary 5.1).[154] Note that this result can be calculated in theory, since all the values involved can in principle be observed. However, obtaining such a Rademacher average is generally a daunting task. We now turn our attention to methods for bounding such Rademacher averages.

7.2.3 Bounding Rademacher and Gaussian penalties

We have mentioned that evaluating a Rademacher penalty is equivalent in difficulty to performing ERM on the function class. In the general case, performing ERM is not feasible, so that one must rely on upper bounds on the Rademacher penalty or average.

As with covering numbers and dimension measures, a number of results are available to aid obtaining bounds on the Rademacher penalty of a complex class from those of simpler classes. The results in the following theorem come from Bartlett and Mendelson (2002, Theorem 12). The results are stated for (absolute) Rademacher complexities, but in many cases, similar results can be obtained for Rademacher complexities, as well as (absolute) Rademacher averages and penalties, and their Gaussian counterparts.

Theorem 7.10 (Properties of (absolute) Rademacher complexities). *Let $\mathcal{V}_1, \cdots, \mathcal{V}_k$ be classes of real functions. Then*

[154] Note that their result uses the constant 13 instead of 31, which seems to be incorrect.

1. *If* $\mathcal{V}_1 \subseteq \mathcal{V}_2$,
$$\mathcal{R}'_m(\mathcal{V}_1) \leq \mathcal{R}'_m(\mathcal{V}_2) \enspace .$$
2. $$\mathcal{R}'_m(\operatorname{absconv} \mathcal{V}_1) = \mathcal{R}'_m(\operatorname{conv} \mathcal{V}_1) = \mathcal{R}'_m(\mathcal{V}_1) \enspace .$$
3. *For every* $v \in \mathrm{I\!R}$,
$$\mathcal{R}'_m(v\mathcal{V}_1) = |v|\mathcal{R}'_m(\mathcal{V}_1) \enspace .$$
4. *If* $v : \mathrm{I\!R} \to \mathrm{I\!R}$ *is* K*-Lipschitz and satisfies* $v(0) = 0$, *then*
$$\mathcal{R}'_m(v \circ \mathcal{V}_1) \leq 2K\mathcal{R}'_m(\mathcal{V}_1) \enspace .$$
5. *For any uniformly bounded function* v,
$$\mathcal{R}'_m(\mathcal{V}_1 + v) \leq \mathcal{R}'_m(\mathcal{V}_1) + \frac{\|v\|_\infty}{\sqrt{n}} \enspace .$$
6. *For any uniformly bounded function* v *and* $1 \leq p < \infty$, *let*
$$\mathcal{V}'_1(p) = \{|\phi - v|^p : \phi \in \mathcal{V}_1\} \enspace .$$
If $\|\phi - v\|_\infty \leq 1$ *for all* $\phi \in \mathcal{V}_1$, *then*
$$\mathcal{R}'_m(\mathcal{V}'_1(p)) \leq 2p\left(\mathcal{R}'_m(\mathcal{V}_1) + \frac{\|v\|_\infty}{\sqrt{n}}\right) \enspace .$$
7. $\mathcal{R}'_m\left(\sum_{i=1}^k \mathcal{V}_i\right) \leq \sum_{i=1}^k \mathcal{R}'_m(\mathcal{V}_i)$.

The same results hold for Rademacher complexities, with the following modifications:

- *In part (2), the first equality does not hold;*
- *part (3) requires* $v > 0$*;*
- *in parts (4) and (6), the constant* 2 *is not necessary.*

The modifications necessary to obtain bounds on Rademacher complexities are generally straightforward. The modification of part (4) (and hence part (6)) is an application of Ledoux and Talagrand (1991, Theorem 4.12).[155]

[155] Bartlett and Mendelson (2002) accidentally refer to Corollary 3.17 of the same book.

There is a close relationship between the empirical covering number and the Rademacher average of a class of zero-one functions on a sample S, as pointed out in the following theorem based on results in Bartlett and Mendelson (2002) and Kääriäinen (2004)[156].

Theorem 7.11 (Obtaining Rademacher penalties from shatter coefficients). *Let $\mathcal{W}$ be a class of functions into $\{0,1\}$. Then for all n-samples Q, we have*

$$\bar{\mathcal{R}}_Q(\mathcal{W}) = O\left(\sqrt{\frac{\mathrm{VC}(\mathcal{W}|_Q)}{n}}\right)$$

and

$$\bar{\mathcal{R}}_Q(\mathcal{W}) = O\left(\sqrt{\frac{\ln(|Q_{\mathcal{W}}|)}{n}}\right) \ .$$

Furthermore, we have that

$$\bar{\mathcal{R}}_Q(\mathcal{W}) \leq 2\sqrt{\frac{\ln|Q_{\mathcal{W}}|}{n}} + \frac{1}{|Q_{\mathcal{W}}|}$$

and

$$\mathbb{P}_{Q\sim D^n,\zeta\sim\mathrm{Unif}(\{-1,1\}^n)}\left\{\mathcal{R}_Q(\mathcal{W}) > \sqrt{\frac{2(\ln\mathcal{N}_{\mathcal{W}}(n) - \ln\delta)}{n}}\right\} \leq \delta \ .$$

For further useful results for calculating Rademacher averages for various function classes, the interested reader is referred to, inter alia, Bartlett et al. (2004), Bartlett and Mendelson (2002), Bousquet and Herrmann (2003), Koltchinskii and Panchenko (2002), Mendelson (2003), Shawe-Taylor and Cristianini (2004), von Luxburg and Bousquet (2004).

[156] The authors of these papers use these and similar results as a motivation to replace traditional bounds employing VC dimension of the class by Rademacher bounds: they note that the Rademacher averages are sample-based, allowing much lower values to be obtained, and in the worst case, the bound obtained using Rademacher averages is worse by a factor of $\sqrt{2}$.

Chapter 8

PAC-Bayesian bounds and Occam's hammer

In this chapter, we consider two novel methods for obtaining bounds for stochastic decision rules in specific scenarios. Firstly, the PAC-Bayesian approach provides bounds on the expected risk of the stochastic Gibbs strategy for selecting a decision rule. The very new Occam's hammer methodology, on the other hand, provides a bound for the risk of a single stochastic decision rule (among other applications). The common thread uniting these approaches is that both approaches make use of some "posterior" distribution in the Gibbs class $\mathcal{G}_{\mathcal{H}'}$ associated with some base hypothesis class $\mathcal{H}'$.

In Example 2.5, we discussed three common strategies when the hypothesis class was a Gibbs class. We will particularly focus on the Gibbs strategy in this chapter.

8.1 PAC-Bayesian bounds

In general, PAC-Bayesian bounds refer to any PAC-style bound obtained for a Bayesian approach to a problem. Typically, the Bayesian approach assumes that the prior in the problem under consideration is correct, thus validating the decision rule their approach selects. However, PAC-Bayesian bounds guarantee performance of the resulting decision rule regardless of the veracity of the prior. Thus PAC-Bayesian bounds can be seen as a method for validating a decision rule obtained by a Bayesian approach within the framework of classical statistics.

Perhaps the earliest PAC-Bayesian bound was an application of the luckiness framework to a Bayesian estimator in Shawe-Taylor and Williamson (1997).

However, the foundation of tight PAC-Bayesian bounds was laid by David McAllester with a theorem he proved in McAllester (1998) and generalized in McAllester (1999) and McAllester (2001). PAC-Bayesian bounds are novel in that they seem to extend the union bound to uncountable decision classes directly, without resorting to a cover or an advanced concentration inequality. In fact, we shall see later that these bounds actually can be seen as leveraging bounds obtained with the exponential moment method.

Every $Q \in \mathcal{Q}_{\mathcal{H}'}$ corresponds to an hypothesis $h_Q \in \mathcal{G}_{\mathcal{H}'}$. The Gibbs strategy then selects a stochastic decision rule w_Q by sampling an element of $\mathcal{H}'$ from Q.

Theorem 8.1 (McAllester's PAC-Bayesian bound). *Consider a deterministic hypothesis class $\mathcal{H}'$ and the associated Gibbs class $\mathcal{G}_{\mathcal{H}'}$. Let $\alpha \in \mathcal{Q}_{\mathcal{H}'}$ be a distribution over $\mathcal{H}'$, and $\delta \in (0, 1]$ a confidence level. Then, with probability at least $1 - \delta$, for all $Q \in \mathcal{Q}_{\mathcal{H}'}$,*

$$r_D(w_Q) \leq r_S(w_Q) + \sqrt{\frac{\mathrm{KL}(Q||\alpha) + \ln \frac{m}{\delta} + 2}{2m - 1}} \ .$$

In this theorem, α plays the role of a "prior" distribution over the base hypothesis class $\mathcal{H}'$. Investigating this bound, we see that the benefits it provide come at a price: the resulting bounds are generally trivial when the "prior" is a density over a continuous class, while the posterior Q is discrete (such as when we wish to select a single decision rule from $\mathcal{H}'$). This happens because the Kullback-Leibler divergence explodes when Q becomes highly concentrated on a few points.

It is instructive to compare this result to the following more direct approach.

Suppose $\mathcal{H}'$ is countable, and consider a "prior" α. Then, applying the Occam's razor method to Hoeffding's tail inequality (see Example 5.3) yields

$$r_D(w) \leq r_S(w) + \sqrt{\frac{\ln\frac{1}{\delta} + \ln\frac{1}{\alpha(h)}}{2m}} \tag{8.1}$$

for all $h \in \mathcal{H}'$. (Here we have used $\mathcal{H}'$ as the hypothesis class, along with the identity strategy $g = \mathrm{id}_{\mathcal{A}}$). Now consider some distribution $Q \in \mathcal{Q}_{\mathcal{H}'}$. From (8.1), we have

$$\mathbb{E}_{h\sim Q}\, r_D(h) \leq \mathbb{E}_{h\sim Q}\, r_S(h) + \mathbb{E}_{h\sim Q}\sqrt{\frac{\ln\frac{1}{\delta} + \ln\frac{1}{\alpha(h)}}{2m}}\ .$$

By Jensen's inequality, this yields

$$\begin{aligned} r_D(w_Q) &\leq r_S(w_Q) + \sqrt{\frac{\ln\frac{1}{\delta} + \mathbb{E}_{h\sim Q}\ln\frac{1}{\alpha(h)}}{2m}} \\ &= r_S(w_Q) + \sqrt{\frac{\ln\frac{1}{\delta} + \mathrm{KL}(Q||\alpha) + \mathrm{Ent}(Q)}{2m}}\ , \end{aligned} \tag{8.2}$$

where $\mathrm{Ent}(Q)$ denotes the entropy of Q.

Thus, the PAC-Bayesian bound improves this result roughly when $\mathrm{Ent}(Q) > \ln m$. Furthermore, the PAC-Bayesian bound also applies to uncountable $\mathcal{H}'$.

The form of the PAC-Bayesian bound above (McAllester, 2001, Theorem 1) was constructed using Hoeffding's tail inequality, so a two-sided version can easily be constructed. Seeger (2001) showed that the same proof technique employed for this result could be used for a variety of similar bounds, including a bound based on Hoeffding's r.e. bound. The result in this case is the most well-known PAC-Bayes bound, stated basically in the following form for the first time in Langford and Seeger (2001, Theorem 3):

Theorem 8.2 (Langford-Seeger PAC-Bayesian bound). *Consider any distribution α over the deterministic hypothesis class $\mathcal{H}'$, and a confidence level $\delta \in (0, 1]$. Then, with probability at least $1 - \delta$,*

$$\mathrm{KL}(r_S(w_Q)||r_D(w_Q)) \leq \frac{\mathrm{KL}(Q||\alpha) + \ln\frac{2m}{\delta}}{m-1}$$

holds simultaneously for all distributions $Q \in \mathcal{Q}_{\mathcal{H}'}$.

A one-sided version of this result also holds: with probability at least $1-\delta$,

$$r_D(w_Q) \leq \sup\left\{\epsilon : \mathrm{KL}(r_S(w_Q)||\epsilon) \leq \frac{\mathrm{KL}(Q||\alpha) + \ln\frac{m}{\delta}}{m-1}\right\} \quad . \tag{8.3}$$

Seeger (2001) showed that a slight tightening of this theorem is possible under zero-one loss, with the inequality in that case replaced by

$$\mathrm{KL}(e_S(w_Q)||e_D(w_Q)) \leq \frac{\mathrm{KL}(Q||\alpha) + \ln\frac{m+1}{\delta}}{m} \quad . \tag{8.4}$$

If we use the bound $\mathrm{KL}(v_1||v_2) \geq 2(v_2 - v_1)^2$ on the left hand side of the Langford-Seeger PAC-Bayesian bound, we see that with probability at least $1-\delta$, for all $\mathcal{H}_Q$,

$$r_D(w_Q) - r_S(w_Q) \leq \sqrt{\frac{\mathrm{KL}(Q||\alpha) + \ln\frac{2m}{\delta}}{2(m-1)}} \quad .$$

We see that this gives us a slight weakening of McAllester's PAC-Bayesian bound.

Although the Kullback-Leibler divergence between $r_S(w_Q)$ and $r_D(w_Q)$ is not analytically invertible, it can be inverted numerically: since, for a fixed v_1, $\mathrm{KL}(v_1||v_2)$ is monotonically increasing for $v_2 > v_1$, a simple line search for the smallest v such that

$$\mathrm{KL}(r_S(w_Q)||v) > \frac{\mathrm{KL}(Q||\alpha) + \ln\frac{2m}{\delta}}{m-1}$$

yields an upper bound $\mathscr{U}$ on $r_D(w_Q)$. A similar argument for $v_2 < v_1$ yields a lower bound $\mathscr{L}$, so that $[\mathscr{L}, \mathscr{U}]$ forms a $100(1-\delta)\%$ confidence interval for $r_D(w_Q)$. Note that the bound can be tightened by employing the improved bound for errors rather than risks. This approach first appeared in the literature in Seeger (2002), although it was foreshadowed in Langford and Seeger (2001).

In the bounds above, the role of the distribution α is very much like that of the "prior" α in Section 5.4, as it gives an indication of where it is felt more or less confidence should be placed on respective hypotheses in $\mathcal{H}'$. The KL divergence measures the difference between the original, "prior", assignment of confidence (the distribution α), and the final decision where to place confidence (represented by the distribution Q determining an element in the Gibbs class of $\mathcal{H}'$). In a way, Q can be seen as an analog of a Bayesian posterior distribution, although there is no requirement that Q and α are related. If Q and α are identical, the divergence is 0, and one obtains the tightest bound. The more the selected hypothesis differs from that specified by the "prior", the greater the KL divergence. So, in a way, the KL divergence can be seen as a penalty for incorrectly assigning "prior" confidence — a high divergence would then correspond to a lot of "lost confidence" in the countable case outlined above. Note further that if Q is not absolutely continuous with respect to α, then the KL divergence between them is infinite, and we obtain trivial bounds.

Both of the bounds above rely in essence on the following lemma, proved by an optimization argument involving the Kuhn-Tucker conditions for constrained optimization:

Lemma 8.1 (Lemma 4 of McAllester, 2001). *Let* $Q, \alpha, v \in \mathrm{I\!R}^N$, *with* $\alpha_i, Q_i > 0$ *and* $\sum_{i=1}^N Q^{(i)} = 1$. *If, for some* $c, K > 0$,

$$\sum_{i=1}^{N} \alpha^{(i)} e^{cv^{(i)}} < K \quad ,$$

then

$$\sum_{i=1}^{N} Q^{(i)} v^{(i)} \leq \frac{\mathrm{KL}(Q||\alpha) + \ln K}{c} \quad .$$

To use this lemma for the Langford-Seeger PAC-Bayesian bound above, we use Q as a discrete distribution over N hypotheses in an hypothesis class. α, corresponding to a discrete α above, represents a prior assignment of confidence to those N hypotheses, while $v^{(i)} = \mathrm{KL}(r_S(h_i)||r_D(h_i))$, where the h_i are the N hypotheses under consideration. Making these substitutions, we obtain a bound on the KL divergence between the true and expected error rate of the Gibbs classifier obtained by sampling from h_i according to Q. This bound is in terms of the KL divergence between Q and α, as well as K and c.

Appropriate values for K and c in the lemma above are obtained by bounding the mean of an exponential function of $\mathrm{KL}(r_S(h)||r_D(h))$ w.r.t. $h \sim \alpha$. Full details of that derivation are in Langford and Seeger (2001, Lemma 2). Finally, a limit argument can be employed to show that this bound still holds when N tends to infinity. In this case, the limits of the distributions α and Q can be used, and the Langford-Seeger PAC-Bayesian theorem above results.

8.1.1 Links with concentration inequalities

The choice $v^{(i)} = \mathrm{KL}(r_S(h_i)||r_D(h_i))$ in Lemma 8.1 used to prove Theorem 8.2 is motivated by the form of Hoeffding's r.e. bound, where the probability of deviation between true and empirical risk for a single hypothesis h on an n-sample is bounded by

$$\exp(-n\,\mathrm{KL}(r_D(h) - \epsilon||r_D(h))) \ .$$

Seeger (2001) shows that other concentration inequalities showing exponential decay can be used instead of Hoeffding's r.e. bound. In general, the choice of $v^{(i)}$ is then linked to the form of the concentration inequality employed in a similar fashion.

In this light, it is natural to wonder whether a PAC-Bayes bound built on Bennett's or Bernstein's inequality can be obtained. Unfortunately, technical restrictions employed in Seeger's approach seems to prevent such a result being obtained in his framework. However, such a result can indeed be obtained based on Bernstein's inequality.

In this section, we shall present such a result using a rather different approach to that presented above. This approach, which was pioneered by Olivier Catoni (Catoni, 2003, 2004b), will hopefully shed light on the source of the the improvements of Theorem 8.1 over (8.2).

When deriving a concentration inequality such as Hoeffding's inequality or Bernstein's inequality, one uses the exponential moment method by bounding the right hand side of

$$\mathbb{P}\{V > \epsilon\} \leq e^{\lambda V - \lambda\epsilon} \ .$$

For the obvious choice $V(h') = r_D(h') - r_S(h')$, we are interested in bounds on the deviations of $\mathbb{E}_{h'\sim Q} V(h')$. The approach taken to obtain (8.2) was to bound V for every h' with at least a certain probability $1-\delta$, and then take this expectation over both sides, to obtain a bound on $\mathbb{E}_{h'\sim Q} V(h')$ which holds with probability at least $1-\delta$.

A more direct approach is to consider

$$\mathbb{P}_{S\sim D^m}\{\mathbb{E}_{h'\sim Q}[V(h') - \epsilon(h')] > 0\}$$

directly. Instead of using the bound obtained for V by a concentration inequality, the idea is now to obtain a concentration inequality for the expected deviation directly. By the exponential moment method, we have that

$$\mathbb{P}_{S\sim D^m}\{\mathbb{E}_{h'\sim Q}[V(h') - \epsilon(h')] > 0\} \leq \mathbb{E}_{S\sim D^m} \exp(\lambda\, \mathbb{E}_{h'\sim Q}[V(h') - \lambda\epsilon(h')]) \ .$$

Furthermore, we wish to obtain a bound that holds for all Q[157]. Thus we would like to work with

$$\mathbb{P}_{S\sim D^m}\{\exists Q \in \mathcal{Q}_{\mathcal{H}'} : \mathbb{E}_{h'\sim Q}[V(h') - \epsilon(h')] > 0\} \ . \tag{8.5}$$

If ϵ is independent of h', this leads us to

$$\mathbb{P}_{S\sim D^m}\left\{\sup_{Q\in\mathcal{Q}_{\mathcal{H}'}} \mathbb{E}_{h'\sim Q} V(h') > \epsilon\right\} \leq \mathbb{E}_{S\sim D^m} \exp\left(\lambda \sup_{Q\in\mathcal{Q}_{\mathcal{H}'}} \mathbb{E}_{h'\sim Q} V(h') - \lambda\epsilon\right) \ . \tag{8.6}$$

The problem with this, however, is that the supremum includes distributions which can be concentrated on h' exhibiting "arbitrarily bad" behaviour (where the KL divergence can become infinite). Thus, it seems ϵ can not be held constant. If we regard the "prior" α as encoding a kind of desirable behaviour of the hypotheses in $\mathcal{H}'$, the KL divergence $\mathrm{KL}(Q||\alpha)$ might be interpreted as a "deviation from desirability" of Q. Fortunately, a link can be established between $\mathbb{E}_{h'\sim Q} V(h')$ and $\mathrm{KL}(Q||\alpha)$. If we consider the KL divergence $\mathrm{KL}(Q||\alpha)$ as a convex function of Q, the *convex conjugate* or *Legendre-Fenchel transform* of $\mathrm{KL}(\cdot||\alpha)$ is

$$\mathrm{Leg}[\mathrm{KL}(\cdot||\alpha)](\phi) = \sup_{Q\in\mathcal{Q}_{\mathcal{H}'}} [\mathbb{E}_{h'\sim Q}\, \phi(h') - \mathrm{KL}(Q||\alpha)] \ .$$

[157] This step corresponds to the transition from a test sample to a training sample bound.

Theorem 8.3 (Lemma 1.4.2 of Catoni, 2004b).

$$\mathrm{Leg}[\mathrm{KL}(\cdot||\alpha)](\phi) = \ln \mathbb{E}_{h'\sim\alpha} \exp(\phi(h')) \ .$$

In addition, if ϕ is upper bounded, we have

$$\ln \mathbb{E}_{h'\sim\alpha} \exp(\phi(h')) = \mathbb{E}_{h'\sim Q}\, \phi(h') - \mathrm{KL}(Q||\alpha) + \mathrm{KL}(Q||\nu) \ ,$$

where ν is the Gibbs distribution derived from α based on ϕ, i.e. (assuming α is a continuous distribution),

$$\nu(h') = \frac{\exp(\phi(h'))}{\mathbb{E}_{h''\sim\alpha} \exp(\phi(h''))}\alpha(h') \ .$$

Thus, this result tells us that the deviation of $\mathbb{E}_{h'\sim Q}\, \phi(h')$ from $\mathrm{KL}(Q||\alpha)$ is uniformly controlled[158] by the Legendre-Fenchel transform of $\mathrm{KL}(\cdot||\alpha)$ at ϕ. The next step thus involves bounding this Legendre-Fenchel transform. We consider, for any $\lambda > 0$,

$$\mathbb{P}_{S\sim D^m}\left\{ \sup_{Q\in\mathcal{Q}_{\mathcal{H}'}} \big[\mathbb{E}_{h'\sim Q}[\lambda V(h') - \epsilon(\lambda, h')] - \mathrm{KL}(Q||\alpha)\big] > 0 \right\} \ . \qquad (8.7)$$

By considering the function $\phi(\cdot) = \lambda V(\cdot) - \epsilon(\lambda, \cdot)$, and employing Theorem 8.3, this probability equals

$$\begin{aligned} & \mathbb{P}_{S\sim D^m}\{\ln \mathbb{E}_{h'\sim\alpha} \exp(\lambda V(h') - \epsilon(\lambda, h')) > 0\} \\ = \ & \mathbb{P}_{S\sim D^m}\{\mathbb{E}_{h'\sim\alpha} \exp(\lambda V(h') - \epsilon(\lambda, h')) > 1\} \\ = \ & \mathbb{E}_{S\sim D^m}\, I\left(\mathbb{E}_{h'\sim\alpha} \exp(\lambda V(h') - \epsilon(\lambda, h')) > 1\right) \\ \leq \ & \mathbb{E}_{S\sim D^m}\, \mathbb{E}_{h'\sim\alpha} \exp(\lambda V(h') - \epsilon(\lambda, h')) \\ = \ & \mathbb{E}_{h'\sim\alpha}\, \mathbb{E}_{S\sim D^m} \exp(\lambda V(h') - \epsilon(\lambda, h')) \ , \qquad (8.8) \end{aligned}$$

where we used the fact that the exponential function is always positive.

Finally, we note that if we have derived a concentration inequality for V using the exponential moment method, we have a bound on $\mathbb{E}_{S\sim D^m} \exp(\lambda V(h'))$ for any h'. Thus, in principle, every quantity in this bound can be calculated.

[158] Some technical restrictions apply to the Theorem — see Catoni (2004b) for details.

Example 8.1. Theorem 4.1 is the basis of Hoeffding's tail inequality. From this theorem, one can obtain, for any $h' \in \mathcal{H}'$,

$$\mathbb{E}_{S\sim D^m} \exp(\lambda V(h')) \leq \exp\left(\frac{\lambda^2}{8m}\right) \ .$$

Applying this to (8.8) with ϵ independent of h', one obtains a bound of

$$\exp\left(\frac{\lambda^2}{8m} - \epsilon(\lambda)\right) \ .$$

A convenient choice of ϵ, which helps cancel terms, is

$$\epsilon(\lambda) = \frac{\lambda^2}{8m} - \ln\delta \ .$$

In this case, the bound reduces to δ, and we obtain the following probability statement: for any $\lambda > 0$,

$$\mathbb{P}_{S\sim D^m}\left\{\exists Q \in \mathcal{Q}_{\mathcal{H}'} : \mathbb{E}_{h'\sim Q} V(h') > \frac{1}{\lambda}\left[\left(\frac{\lambda^2}{8m} - \ln\delta\right) + \mathrm{KL}(Q||\alpha)\right]\right\} \leq \delta \ . \tag{8.9}$$

Ideally, we would now like to select a λ minimizing the right hand side. However, the derivative of the right hand side is a function of Q, through the KL divergence, which may be infinite. Thus, we have to content ourselves with selecting a λ which may not be optimal. Here, we shall consider the choice $\lambda = \sqrt{2m}$. In this case, we have

$$\mathbb{P}_{S\sim D^m}\left\{\exists Q \in \mathcal{Q}_{\mathcal{H}'} : r_D(w_Q) > r_S(w_Q) + \frac{\mathrm{KL}(Q||\alpha) + \frac{1}{4} - \ln\delta)}{\sqrt{2m}}\right\} \leq \delta \ . \tag{8.10}$$

Comparison with the result in Theorem 8.1 shows that this approach provides a bound which converges at a rate of $O\left(\frac{1}{\sqrt{m}}\right)$, rather than $O\left(\sqrt{\frac{\ln m}{m}}\right)$ for any choice of δ and Q. This improvement must, however, be balanced with the loss of the square root in the numerator[159]. □

Example 8.2. To compare this approach to that of Langford and Seeger, we recover (8.4) for zero-one loss with this approach. In this case, we use $V(h') = \mathrm{KL}(r_D(h')||r_S(h'))$.

[159] Audibert and Bousquet (2007, Section B.3.6) derive another bound with the same $O\left(\frac{1}{\sqrt{m}}\right)$ behaviour, which exhibits a clearer relationship to Theorem 8.1, by employing a choice of V more similar to that used in the original result.

From the analysis above, we have

$$\mathbb{P}_{S\sim D^m}\left\{\sup_{Q\in\mathcal{Q}_{\mathcal{H}'}}\left[\mathbb{E}_{h'\sim Q}[\lambda V(h')-\epsilon(\lambda,h')]-\mathrm{KL}(Q||\alpha)\right]>0\right\}$$
$$\leq \mathbb{E}_{h'\sim\alpha}\,\mathbb{E}_{S\sim D^m}\exp(\lambda V(h')-\epsilon(\lambda,h')) \ .$$

Seeger (2001) shows that for zero-one loss,

$$\mathbb{E}_{S\sim D^m}\exp\left(m\,\mathrm{KL}\left(r_D(h')||r_S(h')\right)\right)\leq m+1$$

for any h'. Thus, choosing $\lambda = m$, we obtain

$$\mathbb{P}_{S\sim D^m}\left\{\sup_{Q\in\mathcal{Q}_{\mathcal{H}'}}\left[\mathbb{E}_{h'\sim Q}[mV(h')-\epsilon(m,h')]-\mathrm{KL}(Q||\alpha)\right]>0\right\}$$
$$\leq (m+1)\,\mathbb{E}_{h'\sim\alpha}\exp(-\epsilon(m,h')) \ .$$

Restricting ourselves to a choice of ϵ independent of h', setting the bound to δ and solving for ϵ yields

$$\epsilon(m)=\ln\frac{m+1}{\delta} \ .$$

Thus, with probability at least $1-\delta$, for all $Q\in\mathcal{Q}_{\mathcal{H}'}$,

$$\mathbb{E}_{h'\sim Q}\,\mathrm{KL}(r_D(h')||r_S(h'))\leq\frac{\mathrm{KL}(Q||\alpha)+\ln\frac{m+1}{\delta}}{m} \ .$$

It is easy to verify that $\mathrm{KL}(v_1||v_2)$ is convex w.r.t. v_1 and v_2, so that applying Jensen's inequality to

$$\mathbb{E}_{h'\sim Q}\,\mathrm{KL}(r_D(h')||r_S(h'))$$

allows us to recover the result in (8.4).

Note that while the choice of $\lambda = m$ is convenient in this case, there is no clear reason to expect it to be a good choice. Unfortunately, it is not clear how to find good bounds on $\mathbb{E}_{S\sim D^m}\exp(\lambda\,\mathrm{KL}(r_D(h')||r_S(h')))$ for other choices of λ. □

We now return to obtaining a bound based on Bernstein's inequality. First, we present the bound on the m.g.f. of V employed in deriving Bernstein's inequality. The following result is obtained from the proof of Catoni (2004b, Theorem 1.4.1):

Theorem 8.4. *Let V_i, c, σ^2 and ϵ be as in Theorem 4.3. Then the m.g.f. of $V = \frac{1}{n}\sum_{i=1}^n V_i$ is bounded by*

$$\exp\left(\phi\left(\frac{c\lambda}{n}\right)\frac{\sigma^2}{n}\lambda^2\right) \ ,$$

where

$$\phi(v) = \frac{e^v - v - 1}{v^2} \ .$$

Furthermore, choosing

$$\lambda = \frac{n}{c}\ln\left(1 + \frac{c\epsilon}{\sigma^2}\right)$$

is sufficient to derive the simple form of Bernstein's inequality (Theorem 4.4) using the exponential moment method.

By using $V_i(h') = r_D(h') - L(h'(x_i), y_i)$, where (x_i, y_i) denote the i-th element of S, we have $V = r_D(h') - r_S(h')$, with $c = 1$. It follows that

$$\begin{aligned}
&\mathbb{P}_{S\sim D^m}\left\{\sup_{Q\in\mathcal{Q}_{\mathcal{H}'}}\left[\mathbb{E}_{h'\sim Q}[\lambda V(h') - \epsilon(\lambda, h')] - \mathrm{KL}(Q||\alpha)\right] > 0\right\} \\
\leq\ & \mathbb{E}_{h'\sim\alpha}\,\mathbb{E}_{S\sim D^m}\exp(\lambda V(h') - \epsilon(\lambda, h')) \\
\leq\ & \mathbb{E}_{h'\sim\alpha}\exp(-\epsilon(\lambda, h'))\exp\left(\phi\left(\frac{\lambda}{m}\right)\frac{(\varsigma(h'))^2}{m}\lambda^2\right) \ , \qquad (8.11)
\end{aligned}$$

for some appropriate variance bound $(\varsigma(h'))^2$.

At this stage, we are faced with a number of issues: we need a bound for each $(\varsigma(h'))^2$, we need a way to choose the $\epsilon(\lambda, h')$; and we need to make these selections in a way permitting us to invert the resulting bound.

Suppose for now that we have appropriate values for $(\varsigma(h'))^2$. For a desired confidence level $1-\delta$, the following choice of $\epsilon(\lambda, h')$ is convenient:

$$\epsilon(\lambda, h') = \phi\left(\frac{\lambda}{m}\right)\frac{(\sigma(h'))^2}{m}\lambda^2 - \ln\delta \ . \qquad (8.12)$$

With this choice, all the difficult factors in the bound of (8.11) cancel, and the bound reduces to δ.

It follows that with probability at least $1-\delta$, for all $Q \in \mathcal{Q}_{\mathcal{H}'}$,

$$r_D(w_Q) \leq r_S(w_Q) + \frac{1}{\lambda}\left[\mathrm{KL}(Q||\alpha) + \phi\left(\frac{\lambda}{m}\right)\frac{(\varsigma(h'))^2}{m}\lambda^2 - \ln\delta\right] \ .$$

One simple choice for $(\varsigma(h'))^2$ is $r_D(h')$. In this case,

$$\mathbb{E}_{h'\sim\alpha}(\varsigma(h'))^2 = r_D(w_Q) \ ,$$

leading to the bound

$$r_D(w_Q) \leq \left[1 - \frac{\lambda}{m}\phi\left(\frac{\lambda}{m}\right)\right]^{-1}\left[r_S(w_Q) + \frac{1}{\lambda}[\mathrm{KL}(Q||\alpha) - \ln\delta]\right] \ , \qquad (8.13)$$

which only holds (with probability at least $1-\delta$) when

$$1 - \frac{\lambda}{m}\phi\left(\frac{\lambda}{m}\right) > 0 \ .$$

This bound appears in Catoni (2004b, Corollary 1.5.1).

The choice $(\varsigma(h'))^2 = r_D(h')[1 - r_D(h')]$ is somewhat more difficult to handle, and will not provide much benefit when $r_D(h')$ is small. We shall not consider this choice further here.

If we consider the result in (8.13), we see that the best choice of λ once again depends on the sample S, this time through the empirical risk $r_S(w_Q)$ and the KL divergence $\mathrm{KL}(Q||\alpha)$. We now briefly present a modification to the above bounds allowing us to select λ after observing S. Two approaches are feasible: the most simple is employing the Occam's razor method over a grid of potential choices of λ, while the second is to construct another PAC-Bayesian argument for the choice of λ. Both of these approaches employ a "prior" over potential choices of λ. These approaches are described in Catoni (2004b, Section 1.5.3). We shall content ourselves with stating the following result, which is similar to Catoni (2004b, Corollary 1.5.6)[160]:

Theorem 8.5. *For some $b > 1$, let G be the grid*

$$G = \{2mb^{-i} : i \in [0 : \log_b 2m]\} \ .$$

Employing the uniform "prior" over G, and the Occam's razor method, we obtain the following: with probability at least $1-\delta$,

$$r_D(w_Q) \leq \inf_{\{\lambda\in G: \frac{\lambda}{m}\phi\left(\frac{\lambda}{m}\right)<1\}}\left[\left[1 - \frac{\lambda}{m}\phi\left(\frac{\lambda}{m}\right)\right]^{-1}\left[r_S(w_Q) + \frac{1}{\lambda}[\mathrm{KL}(Q||\alpha) - \ln\frac{\delta}{\log_b 2m + 1}]\right]\right] \ .$$

[160] Their corollary can be seen as utilizing the margin unification lemma for this problem.

8.1.2 PAC-Bayesian margin bounds

The first application of the margin with PAC-Bayesian bounds was in Herbrich et al. (1999), where the authors used the margin achieved by a SV machine to obtain a bound on the size of the largest ball in *version space*[161]. This bound was then applied to an early version of the PAC-Bayesian bound from McAllester (1998) applicable when the posterior Q is only positive for hypotheses in the version space. Later work in Herbrich (2002), Herbrich and Graepel (2001) focused on obtaining larger sets in version space to tighten the bounds, and noted that the results were applicable to linear classifiers in general.

The second development in this direction involved the application of a PAC-Bayesian bound to a margin bound. As noted earlier, the earliest agnostic margin bound was that developed specifically for voting classifiers in Schapire et al. (1997). By replacing an application of the Hoeffding tail inequality with a smart application of McAllester's PAC-Bayesian bound (Theorem 8.1), which was then state-of-the-art, a tighter bound could be obtained. Subsequent improved PAC-Bayesian bounds implied improvements to this bound in Langford and Seeger (2001). Notably, these results hold for the voting classifier based on the posterior, not the Gibbs classifier.

The most well-known PAC-Bayesian margin bounds, however, are those derived in Langford and Shawe-Taylor (2002), McAllester (2003). These results employed a Gaussian "prior"[162] for classifiers based on thresholding a weighted sum of basis functions. In order to obtain a bound, they derive a result relating the true risk and the empirical margin risk.

Specifically, let $\eta_1, \eta_2, \eta_3, \cdots, \eta_M$ be an orthonormal basis of an inner product space of functions $\mathcal{H}'$, and let $\alpha \sim N(0, I_M)$, where I_M denotes the $M \times M$ identity matrix. In this situation, an hypothesis h' can be written uniquely as a basis expansion $h' = \sum_{i=1}^{M} h_i \eta_i$, and we use the distribution α

[161] The version space for a given sample S is the set of all hypotheses consistent on S. As such, it is only defined for zero-one loss functions.

[162] This choice is almost completely determined by properties needed by the "prior" in the derivation.

to define this canonical "prior":

$$\alpha(h') = \alpha((h_1, h_2, \cdots, h_M)) \ .$$

Consider any $h^\star \in \mathcal{H}'$ and $v > 0$. Define the posterior distribution $Q(h^\star, v)$ as the following renormalized subset of α:

$$Q(h^\star, v)(h') = \frac{I(\langle h^\star, h' \rangle \geq v)}{\mathbb{P}_{h'' \sim \alpha}\{\langle h^\star, h'' \rangle \geq v\}} \alpha(h^\star) \ .$$

Note that this can be seen as the Gibbs distribution derived from α based on[163] $\phi(h') = \ln I(\langle h^\star, h' \rangle \geq u)$.

We will be interested in $\mathrm{KL}(Q(h^\star, v) || \alpha)$:

$$\begin{aligned} \mathrm{KL}(Q(h^\star, v)||\alpha) &= \mathbb{E}_{h' \sim Q(h^\star, v)} \ln \frac{Q(h^\star, v)(h')}{\alpha(h')} \\ &= \mathbb{E}_{h' \sim Q(h^\star, v)} \ln \frac{1}{\mathbb{P}_{h'' \sim \alpha}\{h^\star \cdot h'' \geq v\}} \\ &= -\ln \mathbb{P}_{h'' \sim \alpha}\{\langle h^\star, h'' \rangle \geq u\} \ . \end{aligned}$$

Expanding $h^\star = \sum_{i=1}^M h_i^\star$ and $h'' = \sum_{i=1}^M h_i$, we have that

$$\langle h^\star, h'' \rangle = \sum_{i=1}^M h_i h_i^\star \ ,$$

a linear combination of zero-mean, independent, Gaussian r.v.'s, with variance $\sum_{i=1}^M (h_i^\star)^2$. Thus, we have that

$$\begin{aligned} \mathrm{KL}(Q(h^\star, v)||\alpha) &= -\ln \mathbb{P}_{Z \sim N(0, \|h^\star\|^2)}\{Z \geq v\} \\ &= \ln \frac{1}{1 - \Phi\left(\frac{v}{\|h^\star\|}\right)} \ . \end{aligned}$$

In what follows, we consider only $h^\star$ with $\|h^\star\| = 1$.

In this setting, we can obtain the following results (see Lemma 4 and Corollary 2 of McAllester, 2003): for any distribution P on $\mathcal{Z}$, and any $\gamma > 0$,

$$e_P\left(h'_{Q(h^\star, v)}, L_{\frac{\gamma}{2}}\right) \leq e_P\left(h^\star, L_\gamma\right) + \left(1 - \Phi\left(\frac{\gamma v}{2}\right)\right) \qquad (8.14)$$

[163] This interpretation assumes sensible methods are used to handle undefined terms in terms of limits.

and

$$e_P\left(h^\star, L_0\right) \leq e_P\left(h'_{Q(h^\star,v)}, L_{\frac{\gamma}{2}}\right) + \left(1 - \varPhi\left(\frac{\gamma v}{2}\right)\right) \quad . \tag{8.15}$$

These formulae use some unusual notation which is necessitated by the fact that in the stochastic case, we are dealing with a composite strategy: thresholding the function obtained by the Gibbs strategy. L_γ denotes the margin loss at margin γ (see Section 5.6.1). The composite strategy necessitates the use of L_0 rather than the underlying zero-one loss function. Furthermore, while $h^\star$ is not strictly a stochastic decision rule in this setting, it is treated as such for the purpose of defining e_P. Note that if we write $w^\star$ for the thresholded version of $h^\star$, we have $e_P(h^\star, L_0) = e_P(w^\star, L)$.

The idea of the PAC-Bayesian margin bound we present is now to use $P = S$ in (8.14) and $P = D$ in (8.15), allowing us to employ the regular PAC-Bayesian bound with loss function $L_{\frac{\gamma}{2}}$. Specifically, we obtain, for any $\gamma > 0$,

$$\begin{aligned} e_D\left(w^\star, L\right) \leq e_S\left(h^\star, L_\gamma\right) &+ \left[e_D\left(h'_{Q(h^\star,v)}, L_{\frac{\gamma}{2}}\right) - e_S\left(h'_{Q(h^\star,v)}, L_{\frac{\gamma}{2}}\right)\right] \\ &+ 2\left(1 - \varPhi\left(\frac{\gamma v}{2}\right)\right) \quad . \end{aligned}$$

We now apply (8.3), to $e_D\left(h'_{Q(h^\star,v)}, L_{\frac{\gamma}{2}}\right)$, yielding, with probability at least $1 - \delta$,

$$\begin{aligned} e_D(w^\star, L) \leq e_S\left(h^\star, L_\gamma\right) - e_S\left(h'_{Q(h^\star,v)}, L_{\frac{\gamma}{2}}\right) + 2\left(1 - \varPhi\left(\frac{\gamma v}{2}\right)\right) \\ + \sup\left\{\epsilon : \mathrm{KL}\left(e_S\left(h'_{Q(h^\star,v)}, L_{\frac{\gamma}{2}}\right) || \epsilon\right) \leq \frac{\mathrm{KL}(Q(h^\star,v)||\alpha) + \ln\frac{m}{\delta}}{m-1}\right\} \quad . \end{aligned}$$

A weaker form, which eliminates the need to calculate $e_S\left(h'_{Q(h^\star,v)}, L_{\frac{\gamma}{2}}\right)$ is

$$e_D(w^\star, L) \leq \sup\left\{\epsilon : \begin{array}{c} \mathrm{KL}\left(e_S\left(h^\star, L_\gamma\right) + \left[1 - \varPhi\left(\frac{\gamma v}{2}\right)\right] || \epsilon - \left[1 - \varPhi\left(\frac{\gamma v}{2}\right)\right]\right) \\ \leq \frac{\ln\frac{1}{1-\varPhi(v)} + \ln\frac{m}{\delta}}{m-1} \end{array}\right\} \quad ,$$

where we have also replaced $\mathrm{KL}(Q(h^\star, v)||\alpha)$ by $\ln\frac{1}{1-\varPhi(v)}$.

To calculate this bound, we need to select γ and v. McAllester (2003) proposes using $v = \frac{\sqrt{8\ln\frac{m\gamma^2}{4}}}{\gamma}$, and selecting γ after observing the data by employing an Occam's razor bound over the set

$$G = \left\{ \sqrt{\frac{i}{m^2}} : i \in [1 : m^2] \right\} \ .$$

The choice of u allows us to obtain appropriate bounds on $\Phi(u)$ and $\Phi\left(\frac{\gamma u}{2}\right)$, yielding the following result.

Theorem 8.6 (McAllester's PAC-Bayesian margin bound). *With probability at least* $1-\delta$*, for all* $h^\star$ *with* $\|h^\star\| = 1$,

$$e_D(w^\star, L) \leq \inf_{\gamma \in G} \sup \left\{ \epsilon : \begin{array}{c} \mathrm{KL}\left(e_S(h^\star, L_\gamma) + \frac{4}{m\gamma^2} || \epsilon - \frac{4}{m\gamma^2}\right) \\ \leq \frac{\frac{4\left(\ln\left(\frac{m\gamma^2}{4}\right)\right)_+}{\gamma^2} + \frac{7}{2}\ln m + \ln\frac{1}{\delta} + 3}{m-1} \end{array} \right\} \ .$$

8.1.3 Data-dependent PAC-Bayesian bounds

Two shortcoming of the PAC-Bayesian bounds we have considered so far is that the bounds generally explode when the posterior is concentrated on a single decision rule; and the bounds are data-independent. Catoni (2004b) presents a method which employs a double sample approach to attempt to remedy these issues. We begin by showing how his approach can be used to obtain data-independent results analogous to traditional covering number results.

Due to symmetrization lemmas (such as (5.7)), a bound on

$$\mathbb{P}_{S \oplus P \sim D^{m+u}} \left\{ \sup_{Q \in \mathcal{Q}_{\mathcal{H}'}} [r_P(w_Q) - r_S(w_Q)] > \epsilon^\star \right\} \tag{8.16}$$

for an appropriate $\epsilon^\star$ can be used to obtain a bound on

$$\mathbb{P}_{S \sim D^m} \left\{ \sup_{Q \in \mathcal{Q}_{\mathcal{H}'}} [r_D(w_Q) - r_S(w_Q)] > \epsilon \right\} \ .$$

By using the conditioning argument of (5.13), (8.16) can be expressed as

$$\mathbb{E}_{M \sim D^{m+u}} \mathbb{P}_{\tau \sim \mathrm{Unif}\, S_{m+u}} \{\mathscr{E}(\tau(M)) | M\} \ ,$$

where

$$\mathscr{E}(S \oplus P) = \left[\sup_{Q \in \mathcal{Q}_{\mathcal{H}'}} [r_P(w_Q) - r_S(w_Q)] > \epsilon^\star\right] \ . \tag{8.17}$$

Catoni's idea is to bound

$$\mathbb{P}_{\tau \sim \mathrm{Unif}\, S_{m+u}}\{\mathscr{E}(\tau(M))|M\}$$

using a "prior" which depends on M. However, to make the approach feasible, the "priors" must be well-behaved in some sense[164]. This is done by using an exchangeable mapping α from the space of $m+u$-samples $\mathcal{Z}^{m+u}$ to $\mathcal{Q}_{\mathcal{H}'}$. The mapping α is exchangeable if, for any $\tau \in S_{m+u}$, and any $M \in \mathcal{Z}^{m+u}$,

$$\alpha(M) = \alpha(\tau(M)) \ ,$$

i.e. the "prior" is insensitive to reordering the sample M. Furthermore, his approach permits one to let ϵ be sensitive to M.

Unfortunately, Catoni's results are limited to the case $m = u$. This is particularly disappointing because his focus is on transductive learning (Vapnik, 1995, 1998), where one is specifically interested in the error rate on a given future test sample, which in practice is unlikely to be the same size as the training sample. In what follows, we present a novel approach combining Catoni's idea with ideas inspired by Devroye's dual sample bound in Devroye (1982), which allows us to obtain results for the case $m \neq u$. While his results only hold for zero-one loss functions, we show that our results apply to general loss functions.

For a given $(m+u)$-sample M, we are interested in

$$\mathbb{P}_{\tau \sim \mathrm{Unif}\, S_{m+u}}\{\mathscr{E}(\tau(M))|M\} \ .$$

Thanks to (5.14), we can rewrite $r_P(h') - r_S(h')$ as $\frac{m+u}{u}[r_{S\oplus P}(h') - r_S(h')]$, so that

$$\mathscr{E}(S \oplus P) = \sup_{Q \in \mathcal{Q}_{\mathcal{H}'}} \mathbb{E}_{h' \sim Q}\left[r_{S\oplus P}(h') - r_S(h') - \frac{u}{m+u}\epsilon^\star\right] > 0 \ .$$

[164] This idea is comparable to that of ω-smallness of a luckiness function.

By the same reasoning that led to (8.7) from (8.5), we shall consider for an arbitrary $M \in \mathcal{Z}^{m+u}$, any exchangeable "prior" $\alpha(M) \in \mathcal{Q}_{(\mathcal{H}')|_M}$, and any $\lambda > 0$,

$$\mathbb{P}_{\tau \sim \mathrm{Unif}\, S_{m+u}} \left\{ \sup_{Q \in \mathcal{Q}_{(\mathcal{H}')|_M}} \left[\begin{array}{c} \mathbb{E}_{h' \sim Q} \left[\lambda V(\tau(M), h') - \epsilon(\tau(M), \lambda, h')\right] \\ -\,\mathrm{KL}(Q||\alpha(M)) \end{array} \right] > 0 \right\} , \tag{8.18}$$

where

$$\begin{aligned} V(S \oplus P, h') &= r_{S \oplus P}(h') - r_S(h') \\ &= r_M(h') - r_S(h') \ . \end{aligned}$$

A very important point to note is that Q and $\alpha(M)$ are distributions over the *restricted* function class $(\mathcal{H}')|_M$. This is what will prevent the KL divergence from exploding when Q is concentrated on a specific decision rule (at least in the case of zero-one loss functions).

Employing Theorem 8.3 as with the derivation of (8.8), this probability does not exceed[165]

$$\mathbb{E}_{h' \sim \alpha(M)}\, \mathbb{E}_{\tau \sim \mathrm{Unif}\, S_{m+u}} \exp\left(\lambda V\left(\tau(M), h'\right) - \epsilon\left(\tau(M), \lambda, h'\right)\right) \ . \tag{8.19}$$

If we require that ϵ be exchangeable, then $t(\tau(M), \lambda, h')$ is independent of τ, and the bound equals

$$\mathbb{E}_{h' \sim \alpha(M)} \exp(-\epsilon(M, \lambda, h'))\, \mathbb{E}_{\tau \sim \mathrm{Unif}\, S_{m+u}} \exp(\lambda V(\tau(M), h')) \ . \tag{8.20}$$

The following result is central to further progress.

Theorem 8.7 (Theorem 4 in Hoeffding, 1963). *Let $E_1, E_2, \cdots, E_n$ be a random n-sample taken from a population without replacement, while $E_1', E_2', \cdots, E_n'$ is a random n-sample taken from the same population with replacement.*

Then, for any convex, continuous function ϕ,

$$\mathbb{E}\, \phi\left(\sum_{i=1}^n E_i\right) \leq \mathbb{E}\, \phi\left(\sum_{i=1}^n E_i'\right) \ .$$

[165] In the derivation of this expression, the final step exchanges expectations. This is only possible because of the assumption that the "prior" $\alpha(M)$ is exchangeable.

In our current scenario we can view the choice of τ as selecting a random sample without replacement, where E_i is the loss on the i-th element of S_τ. Applying this theorem with the function $\phi(v) = e^{-\lambda v}$ to this situation, we obtain

$$\mathbb{E}_{\tau \sim \mathrm{Unif}\, S_{m+u}} \exp(-\lambda r_{S_\tau}(h')) \leq \mathbb{E}_{S' \sim M^m} \exp(-\lambda r_{S'}(h')) \ .$$

Expanding V and employing this result yields a bound of

$$\mathbb{E}_{h' \sim \alpha(M)} \exp(-\epsilon(M, \lambda, h'))\, \mathbb{E}_{S \sim M^m} \exp(\lambda[r_M(h') - r_S(h')]) \ .$$

Note that in this form, we are considering the m.g.f. of the difference between the empirical and "true" risk, conditional on the "true" distribution on $\mathcal{Z}$ being M. As a result, we can use the results derived earlier for Hoeffding and Bernstein's inequalities.

A bound for error

We first consider bounds for an easier case, when ϵ is independent of h'. Using the approach in Example 8.1, we have the bound

$$\begin{aligned}\mathbb{E}_{h' \sim \alpha(M)} \exp(-\epsilon(M, \lambda, h'))\, \mathbb{E}_{S \sim M^m} \exp(\lambda[r_M(h') - r_S(h')]) \\ \leq \exp\left(\frac{\lambda^2}{8m} - \epsilon(M, \lambda)\right) \ .\end{aligned}$$

Previously, it was convenient to choose ϵ to help us cancel terms in the bound. In this case, the situation is somewhat different. We wish to obtain a bound on $\mathscr{E}(M)$ as in (8.17) for a given M, so that we can take an expectation w.r.t. M. Thus we would like to choose ϵ to cancel the KL divergence in probability statement (8.18). However, the KL divergence can vary over each $Q(M) \in \mathcal{Q}_{(\mathcal{H}')|_M}$, while ϵ is constant (since we assume ϵ is independent of h'). The following observations help us address this problem.

In the case of a zero-one loss function, the restricted hypothesis class $(\mathcal{H}')|_M$ is discrete. We begin by considering the most simple choice of $\alpha(M)$, $\alpha(M) = \mathrm{Unif}((\mathcal{H}')|_M)$. Clearly $\alpha(M)$ is exchangeable. Notably, in this case, if we select Q entirely concentrated on a single element of $(\mathcal{H}')|_M$, we obtain

$$\begin{aligned}\mathrm{KL}(Q||\alpha(M)) &= 1\ln\frac{1}{\frac{1}{|(\mathcal{H}')|_M|}} \\ &= \ln|(\mathcal{H}')|_M| \ .\end{aligned}$$

Furthermore, such a Q maximizes the KL divergence from the uniform $\alpha(M)$: it is not difficult to show, by a similar argument, that for any $Q \in \mathcal{Q}_{(\mathcal{H}')|_M}$,

$$\mathrm{KL}\left(Q||\alpha(M)\right) = \ln\left|(\mathcal{H}')|_M\right| - \mathrm{Ent}\left(Q\right) \ .$$

These results are special cases of similar results which hold for arbitrary finite sets, not just $(\mathcal{H}')|_M$.

Thus we have

$$\begin{aligned} & \mathbb{P}_{\tau\sim\mathrm{Unif}\,S_{m+u}}\left\{\sup_{Q\in\mathcal{Q}_{(\mathcal{H}')|_M}}\left[\begin{array}{r} r_M(w_Q)-r_{S_\tau}(w_Q) \\ -\frac{1}{\lambda}\left[\epsilon(M,\lambda)+\ln|(\mathcal{H}')|_M|\right]\end{array}\right]>0\right\} \quad (8.21)\\ \leq\ & \mathbb{P}_{\tau\sim\mathrm{Unif}\,S_{m+u}}\left\{\sup_{Q\in\mathcal{Q}_{(\mathcal{H}')|_M}}\left[\begin{array}{r} r_M(w_Q)-r_{S_\tau}(w_Q) \\ -\frac{1}{\lambda}\left[\epsilon(M,\lambda)+\mathrm{KL}(Q||\alpha(M))\right]\end{array}\right]>0\right\} \ .\end{aligned}$$

In (8.21), we select $\epsilon(M,\lambda)$ such that

$$\frac{1}{\lambda}[\epsilon(M,\lambda)+\ln|(\mathcal{H}')|_M|] = \frac{u}{m+u}\epsilon^\star \ ,$$

obtaining

$$\epsilon(M,\lambda) = \frac{\lambda u\epsilon^\star}{m+u} - \ln|(\mathcal{H}')|_M| \ . \tag{8.22}$$

For any distribution $Q \in \mathcal{Q}_{\mathcal{H}'}$, consider a corresponding distribution $Q^\star \in \mathcal{Q}_{(\mathcal{H}')|_M}$ where, for any $h^\star \in (\mathcal{H}')|_M$ we define

$$Q^\star(h^\star) = Q(\{h' \in \mathcal{H}' : h'|_M = h^\star\}) \ .$$

Clearly,

$$r_M(w_Q) - r_S(w_Q) = r_M(w_{Q^\star}) - r_S(w_{Q^\star}) \ ,$$

so that (8.21) with ϵ as in (8.22), equals

$$\begin{aligned}
\mathbb{P}_{\tau\sim\mathrm{Unif}\,S_{m+u}}\{\mathscr{E}(\tau(M))|M\} &\leq \exp\left(\frac{\lambda^2}{8m}-\epsilon(M,\lambda)\right)\\
&= \exp\left(\frac{\lambda^2}{8m}-\frac{\lambda u\epsilon^\star}{m+u}+\ln|(\mathcal{H}')|_M|\right)\\
&= |(\mathcal{H}')|_M|\exp\left(\frac{\lambda^2}{8m}-\frac{\lambda u\epsilon^\star}{m+u}\right) \quad .
\end{aligned}$$

The optimal choice of λ here is $\frac{4mu\epsilon^\star}{m+u}$, yielding a bound of

$$|(\mathcal{H}')|_M|\exp\left(\frac{-2mu^2(\epsilon^\star)^2}{(m+u)^2}\right) \quad .$$

This result provides a transductive bound for the case $m \neq u$ in terms of the size of the effective hypothesis class on M.

Taking the expectation of this bound w.r.t. $M \sim D^{m+u}$, we obtain that

$$\begin{aligned}
\mathbb{P}_{S\oplus P\sim D^{m+u}}\{\mathscr{E}(S\oplus P)\} &\leq \mathbb{E}_{M\sim D^{m+u}}\,|(\mathcal{H}')|_M|\exp\left(\frac{-2mu^2(\epsilon^\star)^2}{(m+u)^2}\right)\\
&= \mathbb{E}_{M\sim D^{m+u}}\,|M_{\mathcal{H}'}|\exp\left(\frac{-2mu^2(\epsilon^\star)^2}{(m+u)^2}\right) \quad .
\end{aligned}$$

This is a dual sample bound, and can now be combined with a symmetrization lemma for regular deviation.

Example 8.3. Consider the case $m = u$. In this case, we combine the result with the Vapnik symmetrization lemma for regular deviations (Theorem 5.6). This yields

$$\begin{aligned}
&\mathbb{P}_{S\sim D^m}\left\{\sup_{Q\in\mathcal{Q}_{\mathcal{H}'}}[e_D(w_Q)-e_S(w_Q)]>\epsilon\right\}\\
\leq\;& 2\,\mathbb{P}_{S\oplus P\sim D^{2m}}\left\{\sup_{Q\in\mathcal{Q}_{\mathcal{H}'}}[e_P(w_Q)-e_S(w_Q)]>\epsilon-\frac{1}{m}\right\}\\
\leq\;& 2\,\mathbb{E}_{M\sim D^{2m}}\,|M_{\mathcal{H}'}|\exp\left(-\frac{m}{2}\left(\epsilon-\frac{1}{m}\right)^2\right) \quad .
\end{aligned}$$

This result then holds for arbitrary distributions over $\mathcal{H}'$. If we consider those Q concentrated on single hypotheses, rather than exploding as previous PAC-Bayesian bounds do, we obtain a rather weaker version of (5.16). This weakness comes predominantly from the approach we used here employing Hoeffding's inequality, which applies for general loss functions, while (5.16) is based on a result specifically for errors. □

A bound for risk

In this section, we consider general loss functions. In this case, the restricted function class is typically infinite, so that the approach used for error will not work. The solution is to work with covers of the restricted loss class.

For this purpose, let $\mathcal{H}'_\gamma(M,p)$ be any subset of $\mathcal{H}'$ of cardinality $|\mathcal{N}_{p,M}(\gamma,\mathcal{F}_{\mathcal{H}'})|$ such that $\mathcal{F}_{\mathcal{H}'_\gamma(M,p)}$ is a γ-cover of $\mathcal{F}_{\mathcal{H}'}$ w.r.t. $d_{p,M}$.

In the case of errors, we considered distributions over the restricted class $(\mathcal{H}')|_M$, and then showed that the results for these distributions could be applied to arbitrary distributions over $\mathcal{H}'$. Now we use a similar approach, using distributions over $\mathcal{H}'_\gamma(M,p)$ instead. Unfortunately, in this case, not all the functions in $\mathcal{H}'$ correspond exactly to elements of $\mathcal{H}'_\gamma(M,p)$, so an adjustment term is necessary, as is the case with covering number bounds for risk.

We begin with a variant of (8.18): for an arbitrary $M \in \mathcal{Z}^{m+u}$, any exchangeable "prior" $\alpha(M,p) \in \mathcal{Q}_{\mathcal{H}'_\gamma(M,p)}$, and any $\lambda > 0$, we bound

$$\mathbb{P}_{\tau\sim\mathrm{Unif}\,S_{m+u}}\left\{\sup_{Q\in\mathcal{Q}_{\mathcal{H}'_\gamma(M,p)}}\left[\begin{array}{c}\mathbb{E}_{h'\sim Q}[\lambda V(\tau(M),h') - \epsilon(\tau(M),\lambda,h')]\\ -\,\mathrm{KL}(Q||\alpha(M))\end{array}\right] > 0\right\} .$$

As before, if we let $\alpha(M,p) = \mathrm{Unif}(\mathcal{H}'_\gamma(M,p))$, we obtain for all $Q \in \mathcal{Q}_{\mathcal{H}'_\gamma(M,p)}$ that

$$\begin{aligned}\mathrm{KL}(Q||\alpha(M,p)) &= \ln\mathcal{N}_{p,M}(\gamma,\mathcal{F}_{\mathcal{H}'}) - \mathrm{Ent}(Q)\\ &\le \ln\mathcal{N}_{p,M}(\gamma,\mathcal{F}_{\mathcal{H}'}) \ .\end{aligned}$$

Following the same route by which we obtained (8.21) in the case of errors, we obtain that for any $\lambda > 0$,

$$\begin{aligned}\mathbb{P}_{\tau\sim\mathrm{Unif}\,S_{m+u}}\left\{\sup_{Q\in\mathcal{Q}_{\mathcal{H}'_\gamma(M,p)}}\left[\begin{array}{l} r_M(w_Q) - r_{S_\tau}(w_Q)\\ \qquad -\frac{1}{\lambda}\left[\epsilon(M,\lambda) + \ln\mathcal{N}_{p,M}\left(\gamma,\mathcal{F}_{\mathcal{H}'}\right)\right]\end{array}\right] > 0\right\}\\ \le \exp\left(\frac{\lambda^2}{8m} - \epsilon(M,\lambda)\right) \ .\end{aligned}$$

Again, it is convenient to choose $\epsilon(M,\lambda)$ such that

$$\frac{1}{\lambda}[\epsilon(M,\lambda)+\ln\mathcal{N}_{p,M}(\gamma,\mathcal{F}_{\mathcal{H}'})]=\frac{u}{m+u}\epsilon^{\star} \ ,$$

obtaining

$$\epsilon(M,\lambda)=\frac{\lambda u\epsilon^{\star}}{m+u}-\ln\mathcal{N}_{p,M}(\gamma,\mathcal{F}_{\mathcal{H}'}) \ . \tag{8.23}$$

This yields a bound of

$$\mathcal{N}_{p,M}(\gamma,\mathcal{F}_{\mathcal{H}'})\exp\left(\frac{\lambda^2}{8m}-\frac{\lambda u\epsilon^{\star}}{m+u}\right) \ ,$$

which is optimal when $\lambda=\frac{4mu\epsilon^{\star}}{m+u}$. In that case, we obtain a bound of

$$\mathcal{N}_{p,M}(\gamma,\mathcal{F}_{\mathcal{H}'})\exp\left(\frac{-2mu^2(\epsilon^{\star})^2}{(m+u)^2}\right) \ .$$

We now adjust this result to apply to all of $\mathcal{Q}_{\mathcal{H}'}$. For an arbitrary distribution $Q\in\mathcal{Q}_{\mathcal{H}'}$, we can define an associated distribution $Q^{\star}\in\mathcal{Q}_{\mathcal{H}'_{\gamma}(M,p)}$ by

$$Q^{\star}(h^{\star})=Q(h':f'_{h'}=f_{h^{\star}}) \ ,$$

where $f'_{h'}$ denotes the closest element of the cover $\mathcal{F}_{\mathcal{H}'_{\gamma}(M,p)}$ to $f_{h'}$. Thus the mass of $Q^{\star}$ at $h^{\star}$ is the probability (w.r.t. Q) of selecting an h' so that $f_{h^{\star}}$ is the closest cover element to $f_{h'}$. In a way $Q^{\star}$ can thus be seen as the projection of Q onto $\mathcal{H}'_{\gamma}(M)$ (Audibert and Bousquet, 2007, Section 3).

Consider $p=1$. Then, by construction, for any $h'\in\mathcal{H}'$, there is an $h^{\star}\in\mathcal{H}'_{\gamma}(M,1)$ such that $r_M(h^{\star})-r_M(h')\leq\gamma$. By an analysis similar to that preceding Theorem 5.17, we can show that in this case

$$[r_{P_{\tau(M)}}(h')-r_{S_{\tau(M)}}(h')]-[r_{P_{\tau(M)}}(h^{\star})-r_{S_{\tau(M)}}(h^{\star})]\leq\frac{(2m+u)(m+u)}{mu}\gamma \ ,$$

so that

$$\begin{aligned}
&\mathbb{E}_{h'\sim Q}[r_{P_{\tau(M)}}(h')-r_{S_{\tau(M)}}(h')]\\
\leq\ &\mathbb{E}_{h'\sim Q}[r_{P_{\tau(M)}}(h^{\star}(h'))-r_{S_{\tau(M)}}(h^{\star}(h'))]+\frac{(2m+u)(m+u)}{mu}\gamma\\
=\ &\mathbb{E}_{h^{\star}\sim Q^{\star}}[r_{P_{\tau(M)}}(h^{\star})-r_{S_{\tau(M)}}(h^{\star})]+\frac{(2m+u)(m+u)}{mu}\gamma \ .
\end{aligned}$$

It follows that

$$\begin{aligned}
&\mathbb{P}_{\tau\sim\mathrm{Unif}\,S_{m+u}}\left\{\sup_{Q\in\mathcal{Q}_{\mathcal{H}'}}\left[r_{P_{\tau(M)}}(w_Q)-r_{S_{\tau(M)}}(w_Q)\right]>\epsilon^\star|M\right\}\\
\le\ &\mathbb{P}_{\tau\sim\mathrm{Unif}\,S_{m+u}}\left\{\sup_{Q\in\mathcal{Q}_{\mathcal{H}'_\gamma(M,1)}}[r_{P_{\tau(M)}}(w_Q)-r_{S_{\tau(M)}}(w_Q)]>\epsilon^\star-\frac{(2m+u)(m+u)}{mu}\gamma|M\right\}\\
\le\ &\mathcal{N}_{1,M}\left(\gamma,\mathcal{F}_{\mathcal{H}'}\right)\exp\left(\frac{-2mu^2\left[\epsilon^\star-\frac{(2m+u)(m+u)}{mu}\gamma\right]^2}{(m+u)^2}\right)\ . \qquad (8.24)
\end{aligned}$$

The other choice of p we consider is $p=\infty$. In this case, as noted after Theorem 5.17, we obtain

$$[r_{P_{\tau(M)}}(h')-r_{S_{\tau(M)}}(h')]-[r_{P_{\tau(M)}}(h^\star)-r_{S_{\tau(M)}}(h^\star)]\le 2\gamma\ ,$$

where $f_{h^\star}$ is the closest cover element to $f_{h'}$. A similar analysis to above yields a bound of

$$\mathcal{N}_{\infty,M}\left(\gamma,\mathcal{F}_{\mathcal{H}'}\right)\exp\left(\frac{-2mu^2[\epsilon^\star-2\gamma]^2}{(m+u)^2}\right)\ . \qquad (8.25)$$

We can obtain dual sample bounds by taking the expectation of (8.24) and (8.25) w.r.t. $M\sim D^{m+u}$. From (8.24), we obtain

$$\begin{aligned}
&\mathbb{P}_{S\oplus P\sim D^{m+u}}\left\{\sup_{Q\in\mathcal{Q}_{\mathcal{H}'}}[r_P(w_Q)-r_S(w_Q)]>\epsilon^\star\right\}\\
&\le\mathbb{E}_{S\oplus P\sim D^{m+u}}\mathcal{N}_{1,S\oplus P}\left(\gamma,\mathcal{F}_{\mathcal{H}'}\right)\exp\left(\frac{-2mu^2\left[\epsilon^\star-\frac{(2m+u)(m+u)}{mu}\gamma\right]^2}{(m+u)^2}\right)\ ,
\end{aligned} \qquad (8.26)$$

while the probability bound based on (8.25) is

$$\mathbb{E}_{S\oplus P\sim D^{m+u}}\mathcal{N}_{\infty,S\oplus P}(\gamma,\mathcal{F}_{\mathcal{H}'})\exp\left(\frac{-2mu^2[\epsilon^\star-2\gamma]^2}{(m+u)^2}\right)\ .$$

Example 8.4. We now obtain a bound by applying the symmetrization lemma for regular deviation of risk in (5.7) to the dual sample bound of (8.26): for

$0 < \beta \leq 1$ and α a suitable function as determined by the symmetrization lemma,

$$\begin{aligned}&\mathbb{P}_{S\sim D^m}\left\{\sup_{Q\in\mathcal{Q}_{\mathcal{H}'}}\left[r_D(w_Q)-r_S(w_Q)\right]>\epsilon\right\}\\ &\leq \beta^{-1}\,\mathbb{E}_{S\oplus P\sim D^{m+u}}\,\mathcal{N}_{1,S\oplus P}(\gamma,\mathcal{F}_{\mathcal{H}'})\exp\left(\frac{-2mu^2\left[\epsilon-\alpha(u,\beta)-\frac{(2m+u)(m+u)}{mu}\gamma\right]^2}{(m+u)^2}\right)\ .\end{aligned} \tag{8.27}$$

This result is a generalization of Theorem 5.17: we recover that result when the Q are concentrated on single hypotheses. □

A potential improvement of the results above is to utilise the Bernstein-type PAC-Bayesian bounds presented in Section 8.1.1, instead of the Hoeffding-type bounds used here.

The data-dependent approach

In the above results, we have employed a uniform "prior" over a cover of $\mathcal{F}_{\mathcal{H}'}$ in deriving our results. While this is certainly exchangeable, the power of being able to select the prior based on the dual sample M is not being utilized fully.

In what follows, we shall present an algorithm- and data-dependent bound. General data-dependent bounds can be obtained using similar techniques. It will be convenient to consider a deterministic algorithm Θ[166]. The idea, quite simply, is to choose $\alpha(M)$ as a uniform "prior", not on a cover of $\mathcal{F}_{\mathcal{H}'}$, but instead on a cover of the loss class associated with

$$\mathcal{H}'(\Theta, M) = \{\Theta(S_{\tau(M)}) : \tau \in S_{m+u}\}\ .$$

Applying this approach makes it difficult, if not impossible, to apply the techniques above of generalizing to arbitrary distributions or arbitrary decision rules, however. This should not be a problem, though, since this

[166] The assumption that the algorithm is deterministic is convenient, but similar results can be obtained for stochastic algorithms. The resulting bounds will employ covers of larger function classes, however.

approach caters for a specific algorithm, where we wish to bound the performance of $\varTheta(S)$.

Following a similar approach to the derivation of the bound on risk in (8.27), one obtains the following result:

$$\begin{aligned}&\mathbb{P}_{S\sim D^m}\{r_D(\varTheta(S)) - r_S(\varTheta(S)) > \epsilon\}\\ &\quad\leq \begin{array}{c}\left[1-\exp\left(\frac{-\epsilon^2 u}{2}\right)\right]^{-1}\mathbb{E}_{S\oplus P\sim D^{m+u}}\,\mathcal{N}_{1,S\oplus P}\left(\gamma,\mathcal{F}_{\mathcal{H}'(\varTheta,S\oplus P)}\right)\\ \exp\left(-2m\left[\frac{\epsilon u}{2(m+u)}-\frac{2m+u}{m}\gamma\right]^2\right)\end{array} .\end{aligned}$$

Most of the proof involves simply replacing $\mathcal{H}'$ by $\mathcal{H}'(\varTheta, M)$. At the end of the proof, however, one must use a data-dependent symmetrization lemma because $\mathcal{H}'(\varTheta, M)$ is data-dependent. The result given here employs Theorem 6.2 for this purpose, where $\mathscr{E}(S) = I(h' = \varTheta(S))$. This result is rather similar to the algorithmic luckiness result of Theorem 6.5. The main difference between these bounds is that the result we give here uses $\mathbb{E}_{S\oplus P\sim D^{m+u}}\,\mathcal{N}_{1,S\oplus P}(\gamma,\mathcal{F}_{\mathcal{H}'(\varTheta,S\oplus P)})$ as a measure of capacity (which we see is data-independent), while the algorithmic luckiness result employs the algorithmic luckiness function on S to obtain an algorithm- and data-dependent capacity measure.

In order to get a data-dependent result in this case, we follow the same approach as that used in the luckiness framework: we include a condition on the size of $\mathcal{N}_{1,S\oplus P}(\gamma,\mathcal{F}_{\mathcal{H}'(\varTheta,S\oplus P)})$ in the predicate $\mathscr{E}$ used in the data-dependent symmetrization lemma of Theorem 6.2. We then obtain a bound subject to this condition for various bounds on the size, and apply an Occam's razor bound over the various size conditions.

The most natural condition on the size of $\mathcal{N}_{1,S\oplus P}(\gamma,\mathcal{F}_{\mathcal{H}'(\varTheta,S\oplus P)})$ is

$$\mathscr{E}'_i(S\oplus P) = [\ln \mathcal{N}_{1,S\oplus P}(\gamma,\mathcal{F}_{\mathcal{H}'(\varTheta,S\oplus P)}) \leq i]$$

for $i \in \mathbb{N}$. A problem with this choice is that it is dependent on the double sample $S \oplus P$. We thus consider the following alternative:

$$\mathscr{E}^\star_i(S) = [\ln \mathbb{E}_{P\sim D^u}\,\mathcal{N}_{1,S\oplus P}(\gamma,\mathcal{F}_{\mathcal{H}'(\varTheta,S\oplus P)}) \leq i] \enspace .$$

With this choice, we obtain[167]

$$\mathbb{P}_{S\oplus P\sim D^{m+u}}\left\{(r_D(\Theta(S))-r_S(\Theta(S))>\epsilon)\wedge\mathscr{E}_i^{\star}(S)\right\} \\ \leq\left[1-\exp\left(\frac{-\epsilon^2u}{2}\right)\right]^{-1}\exp\left(i-2m\left[\frac{\epsilon u}{2(m+u)}-\frac{2m+u}{m}\gamma\right]^2\right) \quad (8.28)$$

for any $i\in\mathbb{N}$.

In order to apply the Occam's razor method w.r.t. a "prior" α over $\mathbb{N}$, one must set the right hand side to $\delta\alpha(i)$ and solve for ϵ (as a function of S) for each i. To make this feasible, we restrict $\epsilon(S)$ to obtain an upper bound of 2 on

$$\left[1-\exp\left(\frac{-[\epsilon(S)]^2u}{2}\right)\right]^{-1}$$

and invert the weaker bound instead. This occurs when $[\epsilon(S)]^2u\geq 2\ln 2$. In this case, we solve

$$\delta\alpha(i)=2\exp\left(i-2m\left[\frac{\epsilon(S)u}{2(m+u)}-\frac{2m+u}{m}\gamma\right]^2\right)$$

for $\epsilon(S)$, obtaining

$$\epsilon(S)=\frac{2(m+u)}{u}\left[\frac{2m+u}{m}\gamma+\sqrt{\frac{i-\ln\frac{\delta\alpha(i(S))}{2}}{2m}}\right] ,$$

where for any sample S, the appropriate choice of i is

$$i(S)=\lceil\ln\mathbb{E}_{P\sim D^u}\mathcal{N}_{1,S\oplus P}(\gamma,\mathcal{F}_{\mathcal{H}'(\Theta,S\oplus P)})\rceil .$$

We thus have, with probability at least $1-\delta$, for all S satisfying

$$[\epsilon(S)]^2u\geq 2\ln 2 ,$$

that

$$\mathbb{P}_{S\oplus P\sim D^{m+u}}\left\{r_D(\Theta(S))-r_S(\Theta(S))>\epsilon(S)\right\}\leq\delta , \quad (8.29)$$

which is an algorithm- and data-dependent bound.

[167] Note that here the $\mathscr{E}$ employed for applying Theorem 6.2 is $I(h'=\Theta(S))\wedge\mathscr{E}_i^{\star}(S)$.

Comparing the bound of (8.29) to the algorithmic luckiness bound of Theorem 6.5, we see that the results are basically identical: the algorithmic luckiness function in the algorithmic luckiness bound is used to probabilistically bound the expected covering number on the double sample which appears explicitly in this bound.

8.1.4 Discussion

Since the discovery of PAC-Bayesian bounds, a mini-industry has arisen investigating various methods of employing them. We begin by briefly mentioning two other developments in the field, besides the bounds presented above.

First, PAC-Bayesian compression bounds (Catoni, 2004b, Graepel et al., 2005) have been developed which use the size of a compression set to obtain data-dependent bounds using methods like those discussed in the previous section. A second approach in Audibert and Bousquet (2007) is PAC-Bayesian generic chaining. This approach uses the ideas of generic chaining to obtain data-dependent bounds, using a sequence of exchangeable "priors" over successively finer-grained partitions of $\mathcal{H}'$.

An interesting question is why PAC-Bayesian bounds perform as well as they do. One interesting theory by Catoni (2004b) involves an interpretation of PAC-Bayesian bounds as investigating the *quantiles* of an empirical process, rather than the supremum of the process, which earlier bounds studied. As a result, previous bounds had to cater for highly unusual hypotheses, while averaging bounds can avoid this if the "prior" and the posterior avoid (or at least do not unduly emphasize) these hypotheses.

Next, we turn to the issue of evaluating these PAC-Bayesian bounds. Most of the bounds we discussed above bounded $r_D(w_Q)$ in terms of $r_S(w_Q)$. In general, it is not easy to actually calculate $r_S(w_Q)$. Generally, however, this value can be approximated arbitrarily closely using Monte Carlo techniques. Results in this direction are presented in Langford (2002, Section 6.3).

Finally, we raise an important point in practice. For general PAC-Bayesian bounds, the "prior" must be specified before observing the data. If this is not controlled, an α can be selected which matches the posterior distribution Q well, yielding artificially low bounds. Thus, for reporting results, some way of verifying that the "prior" was selected without reference to the data is needed.

8.2 Shell decomposition of the union bound

The shell decomposition of the union bound can be seen as a theoretically well-motivated data-dependent approach to selecting the "prior" used in the application of the Occam's razor method for a countable decision class. A practical version of this bound was first developed in Langford and McAllester (2000), and was primarily motivated by the theoretical work of Haussler et al. (1996) based on models of learning used in statistical mechanics.

The idea of this approach is to divide the decision class into layers, or shells, depending on the (unknown) true risk of each decision rule. Put another way, we bin the decision rules, grouping those with similar true risk. We then divide the confidence equally among the bins, and within each bin, the confidence allocated to that bin is again split equally. Both of these equal splits can also be split according to a between-bin or within-bin "prior" if desired, but this is usually not done.

Note that the approach just described can not be performed explicitly, since the true risks of the decision rules are not known. The shell decomposition approach uses the training risk to approximate these shells, allowing one to obtain a bound.

In what follows, although we follow the spirit of Langford and McAllester (2004), we generalize their results in a few directions: first, we employ arbitrary "priors" (allowing one to obtain bounds on countably infinite classes); second, our results cater for general loss functions rather than zero-one loss functions; third, we allow for the case where the number of bins need not equal the size of the sample. The results we derive are related to Hoeffding's tail inequality instead of Hoeffding's r.e. inequality, as employed in Langford and McAllester (2004). As a result, we can optimize our bound over a parameter at the end.

Consider a sequence $0 = v_0 < v_1 < v_2 < \cdots < v_{K-1} < v_K = 1$. This sequence can be used to define subclasses (bins) of a countable decision class $\mathcal{W}$ by

$$\mathcal{W}_j = \{w \in \mathcal{W} : v_{j-1} < r_D(w) \leq v_j\}$$

for $j \in [2 : K]$, and

$$\mathcal{W}_1 = \{w \in \mathcal{W} : r_D(w) \leq v_1\} \ .$$

In general, it is most useful to have the i_j equally spaced, i.e. $v_j = \frac{j}{K}$. This is the case we consider.

The idea is now to divide δ up among the $\mathcal{W}_j$, typically equally, and obtain confidence intervals for the modified confidence levels — applying the union bound will then yield a bound over $\mathcal{W}$. Generally, within each $\mathcal{W}_j$, the union bound is also applied, so that another division of confidence is obtained.

If $\mathcal{W}_j$ is allocated $\frac{\delta}{K}$ confidence, and the decision rules in $\mathcal{W}_j$ are allocated proportionally to some general "prior" α over $\mathcal{W}$, this approach can be seen as updating the "prior" α by taking into account the true errors of the decision rules. Let α' be a discrete distribution on $[1 : K]$ reflecting the weights used for the union bound over the classes $\mathcal{W}_j$. Then, assuming we allocate confidence within $\mathcal{W}_j$ proportionally to α, the confidence we wish to assign to $w \in \mathcal{W}_j$ in order to obtain a $100(1-\delta)\%$ confidence interval is $\delta(w) = \delta\alpha(\{w\})\frac{\alpha'(\{j\})}{\alpha(\mathcal{W}_j)}$. However, we can not calculate this, because we do not know the true error of w, and hence we do not know the appropriate value of j. In addition, we do not know the $\alpha(\mathcal{W}_j)$.

The solution we present to the second problem is to use the concentration of $r_D(w) - r_S(w)$ to bound $\alpha(\mathcal{W}_j)$. We begin with the m.g.f. of $r_D(w) - r_S(w)$,

$$\mathbb{E}_{S\sim D^m} \exp(\lambda[r_D(w) - r_S(w)]) \leq \exp\left(\frac{\lambda^2}{8m}\right) \ .$$

Thus, for any distribution $Q \in \mathcal{Q}_{\mathcal{W}}$, we have

$$\mathbb{E}_{w \sim Q}\, \mathbb{E}_{S \sim D^m} \exp(\lambda[r_D(w) - r_S(w)]) \leq \exp\left(\frac{\lambda^2}{8m}\right) \ ,$$

so that for any $\lambda > 0$,

$$\mathbb{P}_{S \sim D^m}\left\{\mathbb{E}_{w \sim Q} \exp(\lambda[r_D(w) - r_S(w)]) \leq \frac{\exp\left(\frac{\lambda^2}{8m}\right)}{\delta_1}\right\} > 1 - \delta_1$$

for all $\delta_1 > 0$ (by Markov's inequality). A similar argument shows that

$$\mathbb{P}_{S \sim D^m}\left\{\mathbb{E}_{w \sim Q} \exp(\lambda[r_S(w) - r_D(w)]) \leq \frac{\exp\left(\frac{\lambda^2}{8m}\right)}{\delta_1}\right\} > 1 - \delta_1$$

for all $\delta_1 > 0$.

We now assume that S satisfies

$$\mathbb{E}_{w \sim Q} \exp(\lambda|r_D(w) - r_S(w)|) \leq \frac{\exp(\frac{\lambda^2}{8m})}{\delta_1} \ ,$$

which from the two results above occurs with probability at least $1 - 2\delta_1$. Again from Markov's inequality, we have for this S that

$$\mathbb{P}_{w \sim Q}\left\{\exp(\lambda|r_D(w) - r_S(w)|) \geq \frac{2\exp\left(\frac{\lambda^2}{8m}\right)}{\delta_1}\right\} \leq \frac{1}{2} \ .$$

Let us write Q for the renormalized version of the "prior" α over $\mathcal{W}_j$, i.e.

$$Q(w) = \frac{\alpha(w)}{\alpha(\mathcal{W}_j)} \ .$$

It follows that

$$\alpha\left(\left\{w \in \mathcal{W}_j : |r_D(w) - r_S(w)| \leq \frac{1}{\lambda} \ln \frac{2\exp\left(\frac{\lambda^2}{8m}\right)}{\delta_1}\right\}\right) \geq \frac{1}{2}\alpha(\mathcal{W}_j) \ .$$

Since we are considering $w \in \mathcal{W}_j$, we have

$$\begin{aligned}\left|\frac{2j-1}{2K} - r_S(w)\right| &\leq \left|\frac{2j-1}{2K} - r_D(w)\right| + |r_D(w) - r_S(w)| \\ &\leq |r_D(w) - r_S(w)| + \frac{1}{2K}\end{aligned} \ ,$$

so that

$$\alpha\left(\mathcal{W}_j\right) \leq 2\alpha\left(\left\{w \in \mathcal{W}_j : \left|\frac{2j-1}{2K} - r_S(w)\right| \leq \frac{1}{2K} + \frac{1}{\lambda}\ln\frac{2\exp\left(\frac{\lambda^2}{8m}\right)}{\delta_1}\right\}\right) .$$

The minimal value of

$$\frac{1}{\lambda}\ln\frac{2\exp\left(\frac{\lambda^2}{8m}\right)}{\delta_1}$$

can be shown to occur at $\lambda = \sqrt{8m\ln\frac{\delta_1}{2}}$. With this choice of λ we obtain that with probability at least $1-2\delta_1$ (over the selection of S),

$$\begin{aligned}\alpha(\mathcal{W}_j) &\leq 2\alpha\left(\left\{w \in \mathcal{W}_j : \left|\frac{2j-1}{2K} - r_S(w)\right| \leq \frac{1}{2K} + \frac{\ln\frac{4}{\delta_1^2}}{\sqrt{8m\ln\frac{\delta_1}{2}}}\right\}\right)\\ &\leq 2\alpha\left(\widehat{\mathcal{W}_j}(\delta_1)\right) ,\end{aligned}$$

where we define

$$\widehat{\mathcal{W}_j}(v) = \left\{w \in \mathcal{W} : \left|\frac{2j-1}{2K} - r_S(w)\right| \leq \frac{1}{2K} + \frac{\ln\frac{4}{v^2}}{\sqrt{8m\ln\frac{v}{2}}}\right\} .$$

By employing a "prior" α' over $[1:K]$, one has that with probability at least $1-2\delta_1$,

$$\forall j \in [1:K] : \alpha(\mathcal{W}_j) \leq 2\alpha\left(\widehat{\mathcal{W}_j}(\delta_1\alpha'(\{j\}))\right) .$$

We note that one can obtain each $\alpha\left(\widehat{\mathcal{W}_j}(\delta_1)\right)$ in theory, since all the values needed are known. Now that we have probabilistically upper bounded each $\alpha(\mathcal{W}_j)$, we would like to apply the Occam's razor method. However, we still do not know the appropriate choice of j for a given w. Despite this, we are still able to obtain bounds in certain cases using this "prior".

To do this, we need to go back to basics, however. One general approach to constructing a confidence region of coverage at least $1-\delta$ is to begin with a probability statement involving the parameter of interest w.r.t. the sample space which holds with probability at least $1-\delta$. The confidence region is then obtained by selecting any values of the parameter of interest for which the predicate inside the probability statement holds.

As an example, we shall employ Hoeffding's tail inequality to obtain a confidence region for $r_D(w)$ using this approach. Employing the Occam's razor method with Hoeffding's tail inequality implies that

$$\mathbb{P}_{S\sim D^m}\left\{\forall j \in [1:K] : \left(\forall w \in \mathcal{W}_j : r_D(w) \le r_S(w) + \sqrt{\frac{\ln \frac{\alpha(\mathcal{W}_j)}{\delta_2 \alpha(w)\alpha'(\{j\})}}{2m}}\right)\right\} > 1 - \delta_2 \ .$$

Upper bounding the $\alpha(\mathcal{W}_j)$ by $\alpha\left(\widehat{\mathcal{W}_j}\left(\delta_1\alpha'\left(\{j\}\right)\right)\right)$, we have that

$$\mathbb{P}_{S\sim D^m}\left\{\forall j \in [1:K] : \left(\forall w \in \mathcal{W}_j : r_D(w) \le r_S(w) + \sqrt{\frac{\ln \frac{\alpha\left(\widehat{\mathcal{W}_j}(\delta_1\alpha'(\{j\}))\right)}{\delta_2 \alpha(w)\alpha'(\{j\})}}{2m}}\right)\right\} > 1 - 2\delta_1 - \delta_2 \ .$$

In this case, the true error rate influences the right hand side, so that direct inversion can not be done. Instead, we obtain that

$$\left\{p : p \le r_S(w_S) + \sqrt{\frac{\ln \frac{\alpha\left(\widehat{\mathcal{W}_j}(\delta_1\alpha'(\{j(p)\}))\right)}{\delta_2 \alpha(w_S)\alpha'(\{j(p)\})}}{2m}}\right\} \ ,$$

where $j(p) = \max(1, \lceil np \rceil)$, is a $100(1 - 2\delta_1 - \delta_2)\%$ confidence region for $r_D(w_S)$.

A similar bound employing Hoeffding's r.e. inequality (using uniform "priors" throughout) is presented for zero-one loss functions and a finite decision class in Langford and McAllester (2004). However, for zero-one loss functions, it is even better to use a direct bound on the binomial distribution, such as the binomial tail bound. Similarly for general loss functions, it may be profitable to investigate the application of Bennett's or Bernstein's inequality instead of the inequalities due to Hoeffding, as these inequalities can incorporate additional information on the variance of $r_D(w)$, so that it may be possible to obtain tighter bounds for those w with small risk or variance.

The largest problem with these bounds is still their application. In order to calculate the bounds, the empirical error of every decision rule in $\mathcal{W}$ needs to be found. For some decision classes this is feasible, but in general this can not be done. Langford (2002, Section 8.2) presents a so-called *sampling shell bound* which attempts to circumvent this problem by replacing the "prior" measure of the empirical shells above by an empirical estimate of the same amount. This is done by sampling from $\mathcal{W}$ according to α and using the resulting empirical measure. This approach however, suffers a dramatic loss in performance compared to the regular shell bound outlined above.

Finally, we briefly discuss why we expect such a bound to work well. For most decision classes, we expect very few functions in the class to perform very well on the problem, while most of the functions exhibit mediocre performance. As a result, the $\mathcal{W}_j$ corresponding to small j tend to be very small, so that the corresponding confidence allocated to each decision rule in $\mathcal{W}_j$ is much higher than that placed on the typical decision rule in those $\mathcal{W}_j$ corresponding to larger j, thus resulting in tighter confidence intervals for those w with small risk. Note that the key to this approach was that we could effectively replace the true shells with empirical shells. To do this we used the m.g.f. of the deviation between true and empirical error. Thus this result relies on the concentration of the empirical error.

8.3 Occam's hammer

In Example 2.5, we discussed three strategies when the hypothesis class is a Gibbs class. The PAC-Bayesian bounds discussed earlier in this chapter present bounds for the Gibbs strategy.

In this section, we consider another approach which lies somewhere between the Gibbs strategy and the Bayes strategy. This approach selects a single base hypothesis according to the distribution Q corresponding to an element of the Gibbs class, and then uses the resulting hypothesis for all future data. As a result, if this approach is used, the decision class actually consists of the base hypotheses, i.e. $\mathcal{W} = \mathcal{H}'$. In contrast, the Gibbs decision rule is stochastic, with the decision on each future point being made by a potentially different base hypothesis.

Suppose that this approach, which we shall call the *Blanchard strategy* [168] selects an hypothesis h'_S based on sampling from $Q(S)$. Thus, in a sense, we are interested in a confidence interval for $r_D(h'_S)$. From this viewpoint, we can employ standard bounds which apply to the entire base hypothesis class.

An alternative viewpoint, which we pursue here, is to provide a confidence interval w.r.t. the selection of S and the sampling from Q. Specifically, let $U \sim \mathrm{Unif}[0,1]$, and let $h'_S = h'_u$, the $100u$-th percentile of the distribution Q (again, we assume a sensible definition of percentile). Then we might desire a statement of the form

$$\mathbb{P}_{S \sim D^m, U \sim \mathrm{Unif}[0,1]}\{r_D(h'_S) \in A(S,U)\} \geq 1 - \delta$$

for some set-valued function A. This can be rewritten as

$$\mathbb{P}_{S \sim D^m, h' \sim Q(S)}\{r_D(h') \in A(S,h')\} \ .$$

Bounds on statements of this form can be constructed from regular confidence intervals on individual base hypotheses by a new construction known as *Occam's hammer* (Blanchard and Fleuret, 2007). This paper presents the following result, restated for our purposes, which can be seen as a generalization of the Occam's razor method.

Theorem 8.8 (Occam's hammer). *Suppose for every $h' \in \mathcal{H}'$ and $\delta \in [0,1]$ we have a set $R(h',\delta)$ such that*

$$\mathbb{P}_{S \sim D^m}\{S \in R(h',\delta)\} \leq \delta$$

and the function $I(S \in R(h',\delta))$ is jointly measurable in (S,h',δ). Furthermore, we assume that for all h',

$$\emptyset = R(h',0) \subseteq R(h',\delta_1) \subseteq R(h',\delta_2)$$

[168] According to the definitions given in Section 2.2, the approach discussed here is not strictly a strategy, mainly because it requires knowledge of the distribution Q — see footnote 10.

for $\delta_1 \leq \delta_2$.

Let τ be a nonnegative measure on $\mathcal{H}'$. Let α_1 be a "prior" measure on $\mathcal{H}'$ which is absolutely continuous w.r.t. τ, and denote the Radon-Nikodym derivative $\frac{d\alpha_1}{d\tau}$ by α_1'. Let α_2 be a "prior" on $(0, \infty)$ (the inverse density prior), and denote the c.d.f. of α_2 by F_{α_2}. For some $\delta' \in [0,1]$, define

$$\varDelta(h', v) = \min\left(\delta' \alpha_1'(h') F_{\alpha_2}(v), 1\right) \quad ,$$

the level function of h' at v.

Consider any algorithm $\varTheta$ which returns an hypothesis $\varTheta(S) \in \mathcal{G}_{\mathcal{H}'}$ corresponding to a distribution $Q(S) \in \mathcal{Q}_{\mathcal{H}'}$, and employs the Blanchard strategy.

If, for all S, $Q(S)$ is absolutely continuous w.r.t. τ, with Radon-Nikodym derivative $Q'(S)$, and $(Q'(S))(h')$ is a jointly measurable function of (S, h'), then

$$\mathbb{P}_{S \sim D^m, h' \sim Q(S)} \left\{ S \in R\left(h', \varDelta\left(h', \left[\left(Q'(S)\right)(h')\right]^{-1}\right)\right)\right\} \leq \delta' \quad . \tag{8.30}$$

Our application of this theorem links the sets $R(h', \delta)$ to confidence intervals. Specifically, for some approach to deriving confidence intervals for $r_D(h')$, we set $R(h', \delta)$ to be the set of samples for which the realization of the confidence interval does not contain $r_D(h')$. Furthermore, we need these confidence intervals to satisfy a monotonicity condition. The measure τ is a general measure of volume on the base hypothesis class — for example, one could choose the Lebesgue measure on the parameter space.

In this context, we note that (8.30) implicitly provides a $100(1 - \delta')\%$ confidence interval on $r_D(h'_S)$, which is dependent on the "priors" α_1 and α_2. The "prior" α_1 is already familiar, but the "prior" α_2 needs some explanation. The term inverse density prior derives from the fact that F_{α_2} is applied to the inverse of $(Q'(S))(h')$ when the level function is employed in (8.30).

To clarify the role of α_2, suppose for a moment that $Q'(S)$ is constant over a set $R \subset \mathcal{H}'$ with finite volume $\tau(R)$, and zero elsewhere. Then the inverse of the density at h' is exactly $\tau(R)$. In this case, α_2 can be seen as a "prior" over the volume of the set R. When $Q'(S)$ is not constant, the inverse density at h' is no longer constant. In this case, note that the inverse density is exactly the volume a set R would need to be such that a constant density on R would integrate to one. Thus, in a sense, the inverse density at h' is an "effective volume" measure, which caters for the density at h'. Thus, base hypotheses with higher densities, which are more likely to be selected, correspond to smaller "effective volumes" than those with lower densities.

We now provide an example of the application of this theorem, employing the binomial tail bound. This example is based on Blanchard and Fleuret (2007, Proposition 3.1), which uses the same "priors" to obtain bounds using confidence intervals from Hoeffding's r.e. inequality.

Example 8.5. Consider a volume measure τ over $\mathcal{H}'$. We set the "prior" measure α_1 equal to τ, so that $\alpha_1' = 1$. For any $K > 0$, we consider the inverse density prior α^2 having density $\frac{1}{K}v^{\frac{1}{K}-1}$ on $v \in (0,1]$, and zero elsewhere. Thus, we have

$$F_{\alpha_2}(v) = \frac{1}{K+1}\min\left(v^{\frac{1}{K}+1}, 1\right) \quad .$$

Basing our approach on confidence intervals obtained from the binomial tail bound, we define $R(h',\delta)$ as

$$R(h',\delta) = \left\{S \in \mathcal{Z}^m : e_D(h') > \mathrm{UBT}(e_S(h'), m, 1-\delta)\right\} \quad ,$$

the samples for which the empirical error is misleadingly low (see Section 3.2.4). It is not difficult to see that this choice satisfies the monotonicity requirements of Theorem 8.8.

With these choices, if we let $\varTheta$ be an algorithm satisfying the conditions of Theorem 8.8, it follows that

$$\mathbb{P}_{S\sim D^m, h'\sim Q(S)}\left\{S \in R\left(h', \varDelta\left(h', \left[\left(Q'(S)\right)(h')\right]^{-1}\right)\right)\right\} \le \delta' \quad . \tag{8.31}$$

Now, in this case we have

$$\begin{aligned}\varDelta\left(h', \left[\left(Q'(S)\right)(h')\right]^{-1}\right) &= \min\left(\delta'\alpha_1'(h')F_{\alpha_2}\left(\left[\left(Q'(S)\right)(h')\right]^{-1}\right), 1\right) \\ &= \min\left(\frac{\delta'}{K+1}\min\left(\left[\left(Q'(S)\right)(h')\right]^{\frac{-(K+1)}{K}}, 1\right), 1\right) \quad .\end{aligned}$$

If we write $\varDelta^\star$ as shorthand for this expression, we have $S \notin R(h', \varDelta^\star)$ with probability at least $1-\delta'$ (over joint selection of S and h'). When $S \notin R(h', \varDelta^\star)$, we have

$$e_D(h') \le \mathrm{UBT}(e_S(h'), m, 1-\varDelta^\star) \quad ,$$

so that $[0, \mathrm{UBT}(e_S(h'_S), m, 1 - \varDelta^\star)]$ is a $100(1 - \delta')\%$ confidence interval for the true error of a base hypothesis h'_S selected by sampling from $Q(S)$ using the Blanchard strategy. □

Chapter 9

Practical application of bounds

In this chapter, we calculate a number of the bounds discussed in the thesis, evaluating their relative performance and the relative effects of various bound parameters.

9.1 Introduction

This chapter aims to concretize a number of the estimators presented in the course of this thesis. We do this by evaluating the estimators on a benchmark data set, employing two algorithms. Our focus will be on upper confidence intervals for risk based on a training sample. For such intervals, we investigate the influence of various parameters on the upper bounds obtained.

All our investigations were performed using the R statistical computing language (R Development Core Team, 2006).

Description	*R object*	*Genuine*	*Spam*	*Total*
Full data set	`spam100`	2788	1813	4601
Training sample	`spam80`	2237	1444	3681
Test sample	`spam20`	551	369	920

Table 9.1: The various samples and their composition

9.2 Benchmark data: the spam data set

For our analysis, we employed a data-set for a two-class classification problem, the "spam" data set. This set is a benchmark data set in machine learning, and consists of 4601 data points, each representing an email message. Each data point consists of 57 predictors, together with a classification of the point into spam (labelled 1) or genuine email (labelled 0). The first 54 predictors are percentages: real values between 0 and 100. The next is a positive real number, and the last 2 predictors are natural numbers. For the purposes of our calculations, we treat all of the predictors as real numbers, although this is not optimal (we will return to this in the discussion).

The data set is publicly available from the UCI (University of California, Irvine) repository of machine learning databases and domain theories (Asuncion and Newman, 2007), at `http://mlearn.ics.uci.edu/databases/spambase/` .

The data set was divided randomly into a training and test sample using an 80/20 split. The various samples and their composition are summarized in Table 9.1.

9.2.1 Loss functions

We considered two loss functions for this classification problem. The simpler of the two was simply the misclassification rate, using the zero-one loss function $L_m(y_1, y_2) = I(y_1 \neq y_2)$. However, the true loss in such a problem is inherently asymmetric if an algorithm attempts to filter out spam from genuine e-mail: filtering out a genuine e-mail message by misclassifying it as spam is substantially worse (i.e. higher loss) than allowing a spam e-mail to slip through as a genuine e-mail. We modelled this using the asymmetric

Description	*Fitted model*	*Test risk*	*Training risk*
Decision tree, loss=L_m	`dt80_symm`	0.1021739	0.1016028
Decision tree, loss=L_a	`dt80_asymm`	0.02347826	0.01586525
Boosting, loss=L_m	`boost80s`	0.07608696	0.07171964

Table 9.2: Training and test risks of fitted models

loss function

$$L_a(y_1, y_2) = \begin{cases} 0 & \text{if } y_1 = y_2 \\ 1 & \text{if } y_1 > y_2 \\ 0.1 & \text{if } y_1 < y_2 \end{cases}$$

9.2.2 Algorithms applied

We applied two related algorithms to the training sample. First, we considered decision trees, as implemented in the `rpart` R package (Therneau and Atkinson, 2007). We applied this algorithm using default values, for both loss functions. Secondly, we considered boosting. Our approach here was to boost stumps using Adaboost. This analysis was performed using the `ada` R package (Culp et al., 2006), which employs `rpart` to construct the stumps. In this case, we only considered the misclassification rate. We used mostly default values for the algorithm, along with the recommended settings for selecting stumps as the base learner. The only non-default value selected was that we elected to use 200 iterations of the base learner, rather than 50.

Tables 9.2 and 9.3[169] summarize the fitted models we will discuss.

Comparing the training and test risks of these models, it seems that if overfitting is taking place, it is not serious in the case of the misclassification loss. For the decision trees, this is because the algorithm incorporates pruning, and in the case of boosting, this indicates that the number of iterations is not too high (in fact, more iterations would likely improve the fitted model).

[169] In the confusion matrices presented, the rows denote the true class member, and the columns the prediction: the first row/column corresponds to genuine e-mail, and the second to spam. It is perhaps worth mentioning that the `confusion` function in the `mda` R package reverses this traditional presentation.

Fitted model	*Training sample*	*Test sample*
`dt80_symm`	$\begin{bmatrix} 2135 & 102 \\ 272 & 1172 \end{bmatrix}$	$\begin{bmatrix} 522 & 29 \\ 65 & 304 \end{bmatrix}$
`dt80_asymm`	$\begin{bmatrix} 2231 & 6 \\ 524 & 920 \end{bmatrix}$	$\begin{bmatrix} 544 & 7 \\ 146 & 223 \end{bmatrix}$
`boost80s`	$\begin{bmatrix} 2157 & 80 \\ 184 & 1260 \end{bmatrix}$	$\begin{bmatrix} 525 & 26 \\ 44 & 325 \end{bmatrix}$

Table 9.3: Confusion matrices of fitted models on training and test samples

9.3 Test sample estimators

Table 9.4 contains the test sample estimators obtained for the three models. There are five broad groupings in the table. The first grouping consists of point estimators, and the other four list the upper end of a one-sided 95% confidence interval for the risk. The second group provides results employing exact distributions, the third group lists various results based on asymptotic normality, the fourth group presents bootstrap intervals, and the final group presents intervals based on various concentration inequalities. Many of the estimators are specific to the binomial distribution arising from a zero-one loss function, and hence are not applicable to `dt80_asymm`.

The ME prior employed is a Beta$(1, 1)$ distribution, and the minimax prior is a Beta$(\sqrt{920}, \sqrt{920})$ distribution. The PPE estimates are obtained by employing heuristics:

- for the decision tree, the probability employed is based on the split of training points reaching the node used to classify a given test point;
- for boosting, probability estimates are based on a transformation of the unthresholded values. These probability estimates are obtained from the `ada` package.

The Bayesian estimators shrink the error estimates towards 0.5, with the minimax performing a stronger shrinkage.

With respect to `dt80_symm` and `boost80s`, it seems reasonable to consider the max-P bound as a gold standard. The mid-P bound provides a slightly

Estimator	dt80_symm	dt80_asymm	boost80s
Test risk	0.1021739	0.02347826	0.07608696
Maximum entropy	0.1030369	-	0.0770065
Minimax	0.1214814	-	0.09622706
PPE	0.1800135	-	0.1475197
Max-P	0.1200926	-	0.09204876
Pratt's max-P	0.120101	-	0.0920428
Mid-P	0.1195515	-	0.0914774
Pratt's mid-P	0.1195202	-	0.09145363
ME credible	0.1199726	-	0.0919446
Minimax credible	0.1323257	-	0.1053257
Wald (- CC)	0.1185987	-	0.09046514
Wald (+ CC)	0.1191422	-	0.09100862
Wald (+ BS)	0.1191673	-	0.09103067
Logit-Wald	0.1198076	-	0.09177031
Probit-Wald	0.1195723	-	0.09152385
Arcsine-Wald	0.1191754	-	0.09108126
Score (- CC)	0.1197825	-	0.09174074
Score (+ CC)	0.1203628	-	0.09232926
Normal bootstrap	0.1183	0.0283	0.0906
Studentized bootstrap	0.1185	0.0301	0.0916
Basic bootstrap-t	0.1174	0.0285	0.0902
Percentile bootstrap	0.1195	0.0279	0.0902
ABC bootstrap	0.11955305	0.02962831	0.09149320
Hoeffding's tail	0.1425238	0.06382817	0.1164369
Hoeffding's r.e.	0.1282982	0.03782398	0.0993411
A-V	0.1314303	0.03952137	0.1018402
Bernstein	0.1304628	0.04046507	0.1015556
P-H (best ν)	0.14325330	0.04605585	0.11220717
P-H + Occam	0.17233353	0.06444982	0.13844671

Table 9.4: Test sample based estimators of risk for the fitted models

tighter bound with reduced overcoverage at the cost of no longer guaranteeing coverage. In all cases, the Pratt approximations are accurate to within 10^{-4}. It is further interesting to note that the minimax point estimator lies well outside the 95% max-P interval for error. Another interesting observation is that the Bayesian credible interval based on the ME prior actually yields a *tighter* bound than the max-P interval. This is because exact coverage can be obtained in this case, since the posterior distribution is continuous. If a Bayesian credible interval with the same coverage as the max-P interval were to be constructed, its upper endpoint would be larger than the max-P bound.

Turning to the bounds obtained using asymptotic normality, we see that in both cases

$$\text{Wald (- CC)} < \text{Wald (+ CC)} < \text{Wald (+ BS)} < \text{Mid-P} \ .$$

We expect the Wald interval without continuity correction to exhibit consistent undercoverage, and the continuity correction and Blyth-Still adjustment to modify the intervals to take this into account. The resulting Wald-Blyth-Still bound turns out to be very close to the mid-P bound. The three Wald intervals based on a transformation all yield bounds larger than the Wald-Blyth-Still bound in both cases, as do the score intervals, both with and without CC.

The bootstrap intervals exhibit a similar pattern, with the normal bootstrap bound being lower than the more refined ABC bootstrap estimator in all three cases. Note that the ABC bounds match the mid-P bound quite closely. These bootstrap estimators were obtained using the `boot.ci` and `abc.ci` functions of the `boot` R package (Canty, 2005).[170]

If we consider the most reputable confidence intervals among the ones considered thus far: the max-P, ME credible, Wald-Blyth-Still, score (+CC), and the ABC bootstrap intervals, we see that for both `dt80_symm` and `boost80s` the bounds are all within 0.0015 of each other, while the difference between these bounds and the corresponding test risk is roughly 0.015, only ten times

[170] For estimators employing a variance estimate, we used $\frac{\hat{p}(1-\hat{p})}{m}$.

as much. This serves to illustrate that at the scale under consideration, the selection of a specific confidence interval is not just an academic question, but can have a sizable effect.

However, the differences between these methods are almost negligible when they are compared to the bounds obtained from concentration inequalities. The bounds in the last group are derived from results which hold for all distributions. This generality introduces considerable slackness in the bounds when applied to the binomial distribution, and we see that the resulting bounds are nowhere near competitive. We see that the symmetric tails implied by using Hoeffding's tail inequality yield substantially worse results than directly inverting Hoeffding's r.e. bound. The A-V bound's ability to handle asymmetry means that it is a much tighter approximation to the r.e. bound than the Hoeffding's tail bound. The bound based on Bernstein's inequality is weaker than the Hoeffding r.e. bound (and even the A-V bound), since no additional knowledge about the variance of the test error is employed: the binomial variance is used in all cases. As expected, the P-H ν-deviation is always weaker than Bernstein's inequality, which it employs. The P-H + Occam entry employs the Occam's razor method on P-H ν-deviation over a uniform grid of 100 values of ν, so that the provided value is an upper bound of a 95% confidence interval. In contrast, the P-H (best ν) entry provides the best bound which could have been obtained had ν been selected a priori.

9.4 Training sample estimators

9.4.1 Bootstrap point estimates

In all three fitted models, the training risk is less than the test risk, as is to be expected. We begin this section by presenting point estimates of the true risk employing the bootstrap on the training sample. The estimates we consider are the optimism-adjusted training risk and the .632 bootstrap estimate. These 2 estimators were calculated for dt80_symm and dt80_asymm. The estimates were calculated using the bootpred function in the bootstrap

Model	*Tr. risk*	*Est. opt.*	*Adj. tr. risk*	.632
`dt80_symm`	0.1016028	0.007726161	0.1093290	0.1046921
`dt80_asymm`	0.01586525	0.004962239	0.02082749	0.02129312

Table 9.5: Training sample bootstrap point estimates

R package (Statlib and Tibshirani, 2007), and the results are summarized in Table 9.5.[171]

In the table, the entries in the second last column (the adjusted training risk) are the sum of the corresponding values for training risk and estimated optimism. In the case of `dt80_symm`, both the optimism-adjusted training risk and the .632 estimate are well within the various 95% test sample upper confidence intervals for the risk, while in the case of `dt80_asymm`, both estimates are less than the corresponding test risk. However, we note that the .632 estimate is closer to the test risk than the optimism-adjusted training risk in both cases.

9.4.2 Capacity measures of decision trees

All the training sample interval estimators we will present for decision trees will essentially be based on covering numbers of a class of decision trees. It is not difficult to show that the class of all decision trees has infinite VC dimension, and in fact any sample of distinct points can be shattered by the class of decision trees. Without a restriction on the class, we will not be able to obtain non-trivial training sample estimators.

Two restrictions are of interest to us, both of which reflect the size of the tree in some way. These are the number of splits in the tree, and the depth of the tree.

The basic building block for boosting with stumps and decision trees are single decision-tree splits, known as stumps. In principle, these splits could

[171] The .632+ bootstrap estimator is not commonly available in R, to my knowledge. Since the fitting method used employs pruning to avoid overfitting, we decided it was not necessary to evaluate this estimator, since its main benefit is when overfitting is liable to occur.

be complex functions, but it is most customary for them to create binary splits by mapping an input x to $\{0, 1\}$ by thresholding a single feature of x. This is the basic approach used in the `rpart` package we employed.[172] A decision tree with j binary splits has $j + 1$ terminal nodes, and a total of $2j + 1$ nodes.

The depth of a node in a tree refers to the number of splits lying between the root node and that node. Thus the depth of the root node is always zero. The depth of the tree is the maximal depth of any node in the tree. In particular, the depth of a stump is one.

Let $\mathcal{T}_j$ denote the class of decision trees with at most j splits, $\mathcal{T}'_j$ the class of decision trees of depth at most j, and $\mathcal{S}$ denote the class of all stumps. Generally speaking, we will obtain a bound on $\mathcal{T}'_j$ from $\mathcal{T}_j$, and we will obtain a bound on $\mathcal{T}_j$ from various bounds on $\mathcal{S}$.

We begin by discussing covering numbers for various classes of interest, and then we will discuss Rademacher averages.

Covering numbers for stumps

To obtain bounds on the covering numbers of $\mathcal{S}$, we begin by defining the class somewhat more formally. For an input $x \in \mathrm{I\!R}^N$, a split can be represented as an indicator function of the form $I(x^{(i)} < s)$ or the form $I(x^{(i)} \geq s)$, where $x^{(i)}, i \in 1 : N$ is a coordinate of x and $s \in \mathrm{I\!R}$. If we denote these two functions by $h^-_{i,s}$ and $h^+_{i,s}$ respectively, we can define

$$\mathcal{S} = \bigcup_{i \in 1:N} (\mathcal{S}^-_i \cup \mathcal{S}^+_i) \ ,$$

where $\mathcal{S}^-_i = \{h^-_{i,s} : s \in \mathrm{I\!R}\}$, and $\mathcal{S}^+_i$ is defined likewise.

Let us begin with the simplest bound on the covering numbers of $\mathcal{S}$. Clearly, any element of $\mathcal{S}$ is a halfspace in $\mathrm{I\!R}^N$, so that the VC-dimension of $\mathcal{S}$ is max-

[172] The package also caters for so-called *surrogate splits*, which allows the resulting decision trees to cope with missing data. However, the data set under consideration does not have any missing data, and our results do not apply to trees employing such surrogate splits.

imally $N+1$, which in the case of the spam data set is 58. (See Theorem 5.30 and Example 5.18.) The VCSS lemma (Theorem 5.24) now provides bounds on the shatter coefficient of the class, while (5.54) and (5.55) provide bounds on the packing (and hence covering) numbers of the class, since the VC dimension of a class of $\{0,1\}$ functions is also its pseudodimension.

However, this approach is unnecessarily conservative, since the elements of $\mathcal{S}$ form a very small subset of all hyperplanes: each separating hyperplane corresponding to a halfspace in $\mathcal{S}$ is parallel to $N-1$ of the axes in $\mathrm{I\!R}^N$. Employing part 2 of Theorem 5.34, we can see that for any metric d,

$$\mathcal{N}(\gamma, \mathcal{S}, d) \leq \sum_{i=1}^{N} \left[\mathcal{N}(\gamma, \mathcal{S}_i^-, d) + \mathcal{N}(\gamma, \mathcal{S}_i^+, d)\right] \ .$$

By an argument almost identical to Example 5.11, it follows that the VC dimension of any $\mathcal{S}_i^-$ or $\mathcal{S}_i^+$ is 1. The results in the previous paragraph then allow one to upper bound covering numbers and shatter coefficients of these classes.

We can use this result to bound the VC dimension of the class of splits in the following way: from the VCSS lemma, the shattering coefficient of any $\mathcal{S}_i^-$ or $\mathcal{S}_i^+$ on a sample of size n is at most $n+1$. Thus, we have that

$$\mathcal{N}_{\mathcal{S}}(n) \leq 2N(n+1) \ . \tag{9.1}$$

In the case of $N = 57$, we have that $114(10+1) > 2^{10}$ and $114(11+1) < 2^{11}$, so that it follows that the VC dimension of the class of splits is maximally 10, a substantial improvement over 58.

We now consider improving these bounds based on representation of the data. Until now, we have assumed that the inputs are real numbers. However, the first 54 features are percentages provided to two decimal places, e.g. 4.36%. As such, each of these features can only assume one of 10001 possible values for each observation in the training and test sample. Suppose we know that all future data supplied to a classifier will be in this form. In that case, there are maximally 10001 effective splits in any $\mathcal{S}_i^-$ or $\mathcal{S}_i^+$, $i \in 1:54$, since all splits in such a set for which the value of s are identical to two decimals are indistinguishable. Thus, we can reduce the covering number bound to

$$\mathcal{N}_{\mathcal{S}}(n) \leq 108 \min(n+1, 10001) + 6(n+1) \ , \tag{9.2}$$

where the second term is for the remaining 3 features where we don't have the advantage of representation. This modification does not affect the VC dimension estimate however, since it only yields an improved result for $n > 10000$. When $n \gg 10000$, the benefits reaped here could be substantial, but the remaining three features tend to dwarf the other term, limiting the potential benefits. Unfortunately, we can not apply this approach to all of the features.[173]

Covering numbers for $\mathcal{T}_j$

To obtaining covering numbers for $\mathcal{T}_j$ we must express $\mathcal{T}_j$. Let us for a moment identify any split with the set of points it maps to 1. Our goal is to identify the set of points mapped to 1 by a tree in terms of the corresponding sets for the splits it contains. Clearly, the set is the union of the sets corresponding to the terminal nodes which predict a 1. Each such terminal node is the intersection of a number of sets arising from various splits: either the set corresponding to the split itself, or its negation. Thus, the set mapped to 1 by the tree can be obtained from j elements of $\mathcal{S}$ by using a set theoretic formula (and the set corresponding to any terminal node labelled 1 can be expressed in terms of j' splits, where j' is the depth of the node).

Using this view, we can obtain the VC dimension of $\mathcal{T}_j$ from that of $\mathcal{S}$ by employing Theorem 5.32, yielding

$$\mathrm{VC}(\mathcal{T}_j) \leq 2j\,\mathrm{VC}(\mathcal{S}) \log_2(2j\,\mathrm{VC}(\mathcal{S})) \ .$$

A more direct approach is to employ the covering number bound underlying Theorem 5.32, Theorem 5.33. A bound on the VC-dimension of $\mathcal{T}_j$ can then be obtained by comparing $[\mathcal{N}_{\mathcal{S}}(n)]^j$ to 2^n.

[173] This would result in a finite function class, yielding much better results on large shadow samples.

Splits	VC dim=58	VC dim=10	Cov. numbers
1	795	86	10
2	1823	212	22
3	2938	354	36
4	4110	505	50
5	5324	664	64
6	6572	828	78
7	7848	998	93
8	9148	1171	108
9	10469	1348	124
10	11808	1528	139
11	13165	1711	155
12	14536	1897	171
13	15922	2085	187
14	17320	2276	203
15	18730	2468	219
16	20152	2663	235
17	21584	2859	251
18	23026	3057	268
19	24477	3256	284
20	25937	3457	301

Table 9.6: VC dimension bounds by number of splits in tree

A comparison of these approaches for trees with between one and twenty splits are presented in Table 9.6. It presents a striking illustration of how much can be gained by employing a more refined analysis to obtain a tighter VC dimension bound on a class, and furthermore how much can be gained by utilizing covering number results as far as possible: there is roughly an order of magnitude improvement in the VC dimension bound in each successive column.

As mentioned before, such bounds on VC dimensions can in turn be used to obtain covering numbers or shattering coefficients by employing various results. However, this approach may be wasteful: it might be inefficient to go via the VC-dimension of $\mathcal{T}_j$ to bound its covering numbers if we have access to the covering numbers of $\mathcal{S}$ or its constituent classes. To investigate this[174], we consider bounds on $\ln \mathcal{N}_{\mathcal{T}_5}(7362)$ and $\ln \mathcal{N}_{1,7362}(0.1, \mathcal{T}_5)$ employing

[174] We use 7362 points since this is twice the size of the training sample.

Value	$\mathcal{T}_5 : 5324$	664	64	$\mathcal{S} : 10$	$\mathcal{S}_i^- : 1$
$\ln \mathcal{N}_{\mathcal{T}_5}(7362)$	7045.028	2257.979	364.4275	369.6585	68.2021
$\ln \mathcal{N}_{1,7362}(0.1, \mathcal{T}_5)$	7045.028	2257.979	260.9186	297.3296	60.18074

Table 9.7: Bounds on log-covering numbers for various approaches

the three bounds on the VC dimension above, as well as those obtained when obtaining the shattering coefficients/covering numbers directly from the VC dimension bounds on $\mathcal{S}$ (10) or each $\mathcal{S}_i^-/\mathcal{S}_i^+$ (1). The results are presented in Table 9.7.

In this table, it seems surprising to see the column corresponding to the bound of 64 on the VC dimension of $\mathcal{T}_5$ yielding better results than for the column corresponding to the bound on the VC dimension of $\mathcal{S}$. The final column strongly indicates that we can obtain improved covering numbers by avoiding going via the VC dimensions of the more complex classes, but these other columns seem to gainsay this. A possible explanation of this is the nature of the results used to combine the various classes: the bound used to combine the $\mathcal{S}_i^-$ and $\mathcal{S}_i^+$ to obtain $\mathcal{S}$ was additive in nature: the asymptotic behaviour of the covering numbers was not affected, and continued to grow as a linear function. On the other hand, the bound used to move from $\mathcal{S}$ to $\mathcal{T}_j$ was multiplicative in nature: the degree of polynomial growth of the covering numbers asymptotically grows j-fold. This may help explain why there is little to no improvement, but it does not explain why the bound actually performs worse. This seems to be because the procedure used for calculating the VC dimension bound of 64 took advantage of the knowledge that covering numbers grow in powers of two while the sample size is less than the VC dimension: for a sample of size 64, the covering number bound predicted by the VCSS lemma employing $\text{VC}(\mathcal{S}) = 10$ was greater than 2^{68}, while our knowledge of the growth of covering numbers implies that the actual covering number is at most 2^{68}. This explanation is corroborated by noting that for larger choices of n than 7362, the bounds obtained employing the VC dimension of $\mathcal{S}$ directly begin to outperform those based on the VC dimension of $\mathcal{T}_5$ being 64.

Covering numbers related to $\mathcal{T}_j'$

In this section, we consider alternative covering numbers related to classes of decision trees. We restrict ourselves to the class of decision trees of depth at most j. The approach we use involves representing a decision tree as a thresholded classifier. This allows us to obtain margin bounds in terms of the covering number of the base class, which turns out to be the class of decision trees of depth j with exactly one node labelled 1.

In what follows, we let the decision trees assign labels in $\{-1, 1\}$ instead of $\{0, 1\}$. Any decision tree can then be written as $\operatorname{sgn}\left(\sum_{i=1}^{j'+1} v_i \phi_i(\cdot)\right)$, where j' is the number of splits of the tree, the v_i are *any* real values whose signs correspond to the label of terminal node i, and ϕ_i represents the indicator function indicating membership of terminal node i. This is because any new point will reach exactly one terminal node, say node i, so that the function reduces to $\operatorname{sgn}(v_i)$, the label of terminal node i.

Since the v_i are arbitrary, they can be selected so the sum of their absolute values is 1, for any decision tree. In order to apply margin bounds based on this representation, we need to obtain bounds on the infinity-norm covering numbers of the base class under consideration. Let $\mathcal{L}_j$ denote the class of *leaves* of depth at most j, i.e. indicator functions corresponding to terminal nodes in a decision tree. Then the base hypothesis class can be written as $\operatorname{absconv} \mathcal{L}_j$. By a similar argument, we can see that a classifier generated by boosting decision trees of depth at most j can also be written as a signed thresholded classifier with base hypotheses from the same absolute convex hull.

We thus consider bounding $\mathcal{N}_{\infty,Q}(\gamma, \operatorname{absconv} \mathcal{L}_j)$ for an n-sample Q. Using the relationships between the various norms (see (5.6) and (5.4)), this is bounded by

$$\mathcal{N}_{2,Q}\left(\frac{\gamma}{n}, \operatorname{absconv} \mathcal{L}_j\right) \leq \mathcal{N}_{2,n}\left(\frac{\gamma}{n}, \operatorname{absconv} \mathcal{L}_j\right) \quad .$$

This form allows us to apply Theorem 5.48 with $K = 2$, yielding a bound of[175]

$$\left[2\mathcal{N}_{2,n}\left(\frac{\gamma}{2n},\mathcal{L}_j\right)+1\right]^{\frac{8n^2}{\gamma^2}} .$$

Finally, the 2-norm covering number on the right can in turn be bounded by a 1-norm covering number, by using (5.5), so that we have

$$\mathcal{N}_{\infty,Q}(\gamma,\operatorname{absconv}\mathcal{L}_d) \le \left[2\mathcal{N}_{1,n}\left(\frac{\gamma^2}{4n^2},\mathcal{L}_j\right)+1\right]^{\frac{8n^2}{\gamma^2}} .$$

In order to obtain covering numbers for $\mathcal{L}_j$, we simply note that $\mathcal{L}_j \subseteq \mathcal{T}_j$, so we can use bounds on the covering numbers of $\mathcal{T}_j$.

We expect poor behaviour of these covering numbers due to the confluence of a number of factors. First, the base hypothesis is presented as an absolute convex hull. The convex hull of a class can have very large covering numbers even though the original class has small covering numbers — this serves to negate the tight bounds on decision trees we have constructed in earlier sections. Second, the only result we have available for bounding the covering number of an absolute convex hull employs the 2-norm, while the margin bounds employ an infinity-norm. Third, the only result we have for going from an infinity-norm to a p-norm involves modifying the scale by the sample size. Finally, the covering number bounds in terms of pseudodimension are based on 1-norms, introducing the necessity to square an already small scale. In fact, we shall see that regardless of the initial scale selected for the bound, we shall typically deal with shattering coefficients of the base class.

To illustrate the effects of these issues, consider a sample size of 7362, and $\gamma = 0.1$. In this case, the exponent $\frac{8n^2}{\gamma^2}$ evaluates to a staggering 10786867201, while we obtain a covering number bound of $\mathcal{N}_{1,n}\left(\frac{\gamma^2}{4n^2},\mathcal{L}_1\right) \le$ 839382. This leads to the result

$$\begin{aligned}\ln\mathcal{N}_{\infty,Q}(\gamma,\operatorname{absconv}\mathcal{L}_1) &\le 10786867201\ln 1678765 \\ &= 154614304909 .\end{aligned}$$

We can use this result to apply the bound of (5.46) to the performance of `boost80s` (using $\beta = \frac{1}{2}$ and $\gamma = 0.2$). This yields an upper bound on the

[175] The generally negligible additive constant will often be disregarded in our future calculations.

risk of about 6500. Similarly, all the other margin bounds employing these covering numbers provided similar trivial bounds on the risk, so that further results are not provided.

Rademacher averages for $\mathcal{T}_j$

Obtaining Rademacher averages for classes of decision trees directly is not simple. Some results in this direction employing Gaussian averages are discussed in Bartlett and Mendelson (2002), but converting between Rademacher and Gaussian complexities involves unspecified constants, making the results useless for practical calculations.

Our approach to obtaining Rademacher averages for classes of decision trees is to calculate them from the VC dimension or shatter coefficients using the last two results of Theorem 7.11. Note that the last of these results only provides a probabilistic bound. To get a bound which holds under all conditions, we note that the maximal value of the Rademacher average in our case is 1. Thus, by employing Theorem 7.11 to obtain a bound on $\mathcal{R}_Q(\mathcal{W})$ with probability at least $1-\delta^\star$, we can bound $\mathcal{R}_m(\mathcal{W})$ by

$$\delta^\star(1) + (1-\delta^\star)\sqrt{\frac{2(\ln \mathcal{N}_{\mathcal{W}}(n) - \ln \delta^\star)}{n}} \ . \tag{9.3}$$

This bound can further be optimized over $\delta^\star$.

This approach typically outperforms the alternative method when the bound on the shatter coefficient is not sample dependent. We shall see that this is the case in our examples.

9.4.3 Capacity measures for loss classes

In order to apply bounds in practice, we need to obtain capacity measures on the loss class, not simply the original function class. In this section, we look into obtaining such measures from the corresponding measures of the function classes.

For both the misclassification loss L_m and the asymmetric loss L_a, this problem is easily disposed of for covering numbers by an application of Theorem 5.36. Indeed, both of these losses can be formulated as 1-Lipschitz functions of the distance between the predicted and actual value. As a result the covering numbers of the loss class are no larger than those of the underlying function class for these losses. It should be fairly clear that covering number results for the asymmetric loss will be somewhat looser than those of the misclassification loss.

Slightly more advanced machinery is needed to move from the decision class to the loss class in the case of Rademacher averages. Part 6 of Theorem 7.10 provides a tool for the case of the misclassification loss (with $p = 1$), although the result would not apply to the asymmetric loss. Applying a counterpart of this result for Rademacher averages (not absolute Rademacher averages), one obtains that the Rademacher average of the loss class for any sample of size n does not exceed that of the function class by more than $\frac{1}{\sqrt{n}}$.

However, there does not seem to be a simple way to obtain Rademacher averages for a thresholded class. Thus, in cases where we are considering thresholded classes, we need a way to bypass this problem. The solution is to use an alternative proxy loss function, which acts on the unthresholded values. The resulting bound is similar to the margin bounds we have discussed before, and the derivation is based on Koltchinskii and Panchenko (2002). To understand this, we once again consider the case where the base hypotheses map into $\{-1, 1\}$, and examples are labelled likewise. Then the misclassification loss of a base hypothesis $h' \in \mathcal{H}'$ thresholded at 0 can be expressed in terms of the unthresholded value by $L_m(h'(x), y) = I(yh'(x) \leq 0)$. The essential problem is that this function is not Lipschitz. Suppose $v(\eta) > I(\eta \leq 0)$ is K-Lipschitz. Then the proxy loss defined by $L'(h'(x), y) = v(yh'(x))$ is no smaller than L_m. In fact, it is not difficult to bound the Rademacher averages of the loss class if we can bound the Rademacher averages of the class $\mathcal{H}^\star = \{\phi_{h'}(\cdot) : h' \in \mathcal{H}'\}$, where $\phi_{h'} : \mathcal{Z} \to \mathbb{R}$ is defined by $\phi_{h'}((x, y)) = yh'(x)$.

This is fortunately easy, since by referring to the definition of Rademacher averages, for any sample S,

$$\begin{aligned}\bar{\mathcal{R}}_S\left(\mathcal{H}^\star\right) &= \mathbb{E}_{\zeta_i \sim \text{Unif}\{-1,1\}^m} \sup_{\phi \in \mathcal{H}^\star} \left(\frac{1}{m} \sum_{i=1}^m \zeta_i \phi(x_i, y_i) \right) \\ &= \mathbb{E}_{\zeta_i \sim \text{Unif}\{-1,1\}^m} \sup_{h \in \mathcal{H}'} \left(\frac{1}{m} \sum_{i=1}^m \zeta_i y_i h(x_i) \right) \quad .\end{aligned}$$

If we now make the transformation $\zeta_i' = \zeta_i y_i$, we note that the ζ_i' are also Rademacher r.v.'s, so that the expression equals

$$\mathbb{E}_{\zeta_i' \sim \text{Unif}\{-1,1\}^m} \sup_{h \in \mathcal{H}'} \left(\frac{1}{m} \sum_{i=1}^m \zeta_i' h(x_i) \right) = \bar{\mathcal{R}}_S\left(\mathcal{H}\right) \quad .$$

We now bound the Rademacher averages of the loss class obtained from the proxy loss L'. Let $v(0) = c \geq 1$.[176] We note that $v(\cdot)$ can be written as $(v(\cdot) - c) + c$. Now $v(\cdot) - c$ is also K-Lipschitz, and passes through the origin, so that one can apply part 4 of Theorem 7.10. Adding c can be performed by employing part 5 of the same theorem. Writing $\mathcal{F}'$ for the proxy loss class, for any m-sample S we have

$$\begin{aligned}\bar{\mathcal{R}}_S\left(\mathcal{F}'\right) &= \bar{\mathcal{R}}_S\left(v \circ \mathcal{H}^\star\right) \\ &\leq \bar{\mathcal{R}}_S\left(\left[(v - c) \circ \mathcal{H}^\star\right] + c\right) \\ &\leq \bar{\mathcal{R}}_S\left((v - c) \circ \mathcal{H}^\star\right) + \frac{c}{\sqrt{m}} \\ &\leq K \bar{\mathcal{R}}_S(\mathcal{H}) + \frac{c}{\sqrt{m}} \quad .\end{aligned}$$

Turning to the problem of bounding $\bar{\mathcal{R}}_S(\mathcal{H})$, we note that $\mathcal{H}$ is contained in the absolute convex hull of the leaf functions, absconv $\mathcal{L}_j$. However, since we are not dealing with absolute Rademacher averages, we need to express the class as a convex hull, rather than an absolute convex hull.[177] By definition,

$$\text{absconv}\, \mathcal{L}_j = \text{conv}(\mathcal{L}_j \cup -\mathcal{L}_j) \ ,$$

[176] $v(0)$ must exceed 1, since v upper bounds the misclassification loss.

[177] We could bound the Rademacher average by the absolute Rademacher average, but we do not have an equivalent result to Theorem 7.11 for absolute Rademacher averages.

so that we obtain

$$\begin{aligned}\bar{\mathcal{R}}_S(\mathcal{H}) &\leq \bar{\mathcal{R}}_S(\mathrm{conv}(\mathcal{L}_j \cup -\mathcal{L}_j)) \\ &= \bar{\mathcal{R}}_S(\mathcal{L}_j \cup -\mathcal{L}_j) \\ &\leq \bar{\mathcal{R}}_S(\mathcal{L}_j) + \bar{\mathcal{R}}_S(-\mathcal{L}_j) \\ &\leq 2\bar{\mathcal{R}}_S(\mathcal{L}_j) \ ,\end{aligned}$$

and this final value can be bounded in terms of shatter coefficients using Theorem 7.11.

We now briefly consider the choice of v. We require $v \geq 0$, $v(\eta) > I(\eta \leq 0)$, and that v be K-Lipschitz. In addition, we want v to be as small as possible. For any fixed choice of K, it is not difficult to show that the following v is the optimal choice:

$$v_K(\eta) = \begin{cases} 1 & \eta \leq 0 \\ 1 - K\eta & 0 < \eta < \frac{1}{K} \\ 0 & \eta \geq \frac{1}{K} \end{cases} .$$

For all these choices, we have $c = 1$. In particular v_1 is a squashed version of the so-called hinge loss often used for training support vector machines.

9.5 Bounds on the fitted decision trees

In this section, we shall discuss bounds obtained for the risk of `dt80_symm` and `dt80_asymm`. Note that the bounds discussed above employed a restricted class of decision trees based on the number of splits or the depth of the tree. This approach defines a hierarchy on the class of all decision trees. We employed a “prior” over this hierarchy, based on a minimum expected size: the “prior” was defined by $\alpha(j) = \frac{1}{j(j+1)}$, where in the case of splits, j was the number of splits more than 3, and in the case of depth was the amount the depth exceeded 2 (in both cases, with a minimum value of 1). Note that this prior should be selected *prior* to obtaining the training or test sample. [178]

[178] A slight improvement could be obtained by using the fact that the algorithm never yields trees of depth larger than 31 due to the internal representation of the tree.

We obtained the fitted model `dt80_symm` with the following R command:

```
dt80_symm<-rpart(y~.,spam80,method="class")
```

The following is a summary of the resulting fitted model:

```
> dt80_symm
n= 3681

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 3681 1444 0 (0.60771529 0.39228471)
   2) x.V53< 0.0445 2739  625 0 (0.77181453 0.22818547)
     4) x.V7< 0.065 2494  401 0 (0.83921411 0.16078589)
       8) x.V52< 0.5085 2247  241 0 (0.89274588 0.10725412) *
       9) x.V52>=0.5085 247   87 1 (0.35222672 0.64777328)
        18) x.V57< 35 92   24 0 (0.73913043 0.26086957) *
        19) x.V57>=35 155   19 1 (0.12258065 0.87741935) *
     5) x.V7>=0.065 245   21 1 (0.08571429 0.91428571) *
   3) x.V53>=0.0445 942  123 1 (0.13057325 0.86942675)
     6) x.V25>=0.385 68    7 0 (0.89705882 0.10294118) *
     7) x.V25< 0.385 874   62 1 (0.07093822 0.92906178) *
```

This decision tree has 5 splits and 6 terminal nodes. The splits are on features 7, 25, 52, 53, and 57. The "prior" probability of obtaining 5 splits is $\frac{1}{(5-3)(5-3+1)} = \frac{1}{6}$, so that to obtain strict 95% confidence intervals, we shall typically need to set $\delta = \frac{1-0.95}{6} = \frac{1}{120}$ in our applications of bounds. This allows us to use $\mathcal{T}_5$ as the function class in the bounds.

Note that a larger tree may have yielded a lower training error, but the resulting tree would have been larger, yielding larger capacity measures.

For the asymmetric loss, we employed the command

```
> dt80_asymm<-rpart(y~.,spam80,method="class",parms=list(loss=spam_loss_matrix))
```

obtaining the following result:

```
> dt80_asymm
```

```
n= 3681

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 3681 144.4 0 (0.607715295 0.392284705)
   2) x.V7< 0.01 3041  83.9 0 (0.724103913 0.275896087)
     4) x.V53< 0.0875 2596  45.7 0 (0.823959938 0.176040062)
       8) x.V52< 0.775 2432  33.4 0 (0.862664474 0.137335526) *
       9) x.V52>=0.775 164  12.3 0 (0.250000000 0.750000000)
        18) x.V57< 77 83   4.4 0 (0.469879518 0.530120482)
          36) x.V16< 0.62 67   2.8 0 (0.582089552 0.417910448) *
          37) x.V16>=0.62 16   0.0 1 (0.000000000 1.000000000) *
        19) x.V57>=77 81   2.0 1 (0.024691358 0.975308642) *
     5) x.V53>=0.0875 445  38.2 0 (0.141573034 0.858426966)
      10) x.V52< 0.4095 263  20.2 0 (0.231939163 0.768060837)
        20) x.V24< 0.165 153   9.8 0 (0.359477124 0.640522876)
          40) x.V23< 0.98 121   6.6 0 (0.454545455 0.545454545) *
          41) x.V23>=0.98 32   0.0 1 (0.000000000 1.000000000) *
        21) x.V24>=0.165 110   6.0 1 (0.054545455 0.945454545)
          42) x.V21< 0.875 22   1.7 0 (0.227272727 0.772727273) *
          43) x.V21>=0.875 88   1.0 1 (0.011363636 0.988636364) *
      11) x.V52>=0.4095 182   2.0 1 (0.010989011 0.989010989)
        22) x.V12>=2.365 7   0.5 0 (0.285714286 0.714285714) *
        23) x.V12< 2.365 175   0.0 1 (0.000000000 1.000000000) *
   3) x.V7>=0.01 640  35.0 1 (0.054687500 0.945312500)
     6) x.V52< 0.0765 151  12.4 0 (0.178807947 0.821192053)
      12) x.V13< 0.135 102   7.5 0 (0.264705882 0.735294118)
        24) x.V55< 7.0605 83   5.6 0 (0.325301205 0.674698795) *
        25) x.V55>=7.0605 19   0.0 1 (0.000000000 1.000000000) *
      13) x.V13>=0.135 49   0.0 1 (0.000000000 1.000000000) *
     7) x.V52>=0.0765 489   8.0 1 (0.016359918 0.983640082)
      14) x.V21< 0.085 41   3.6 0 (0.121951220 0.878048780)
        28) x.V3< 0.07 23   1.8 0 (0.217391304 0.782608696) *
        29) x.V3>=0.07 18   0.0 1 (0.000000000 1.000000000) *
      15) x.V21>=0.085 448   3.0 1 (0.006696429 0.993303571) *
```

In this case, we have 15 splits, resulting in a δ modified for “prior” probability of $\frac{0.05}{(15-3)(16-3)} = \frac{1}{3120}$ when we use $\mathcal{T}_{15}$ as the function class. It seems that the asymmetric loss will yield much poorer training sample bounds than the symmetric case: this δ is much lower, and the capacity measures will be substantially higher. Furthermore, the resulting larger value is measured against the scale of a smaller risk value.

The bounds we present are all based on estimates of covering numbers of the function class. In Table 9.7 a number of bounds on covering numbers of various quality are presented. The bounds we present here will be based on the method used in the final column of that table. To illustrate the advantages this column provides, we will first present a single bound using all five techniques employed in the table. We invert the bound on the relative deviation of error of `dt80_symm` obtained by setting the right hand side of (5.19) to $\frac{1}{120}$, and solving for ϵ. The results of the different approaches are summarized in Table 9.8.

In addition, the bounds that will follow will consistently use shatter coefficients, or a limiting value as the scale γ tends to zero. Unless stated otherwise, the covering number results we present are based on obtaining covering numbers from the VC-dimension or pseudodimension of a class, and it seems intuitive that the results we employ will perform best on bounding shatter coefficients. Our intuition was verified experimentally, where bounds employing larger scales consistently performed worse than those with smaller scales. As an illustration, we invert the risk bound obtained on the regular deviation of error using (5.17) at a number of different scales. The results are presented in Table 9.9, and we see that the bounds improve steadily, with the most dramatic improvement at the large scales, as is to be expected. It seems the reason for this is that with a finite VC-dimension, the bounds on the covering numbers we obtain by employing the packing number bounds of (5.54) and (5.55) only improve on the shatter coefficients obtained using the VCSS lemma at rather large scales. This is illustrated by the middle column of the table: for the very large scales, there is a slight reduction in the covering number bound, but on the whole, the cost to the bound of using such a large scale is prohibitive. At smaller scales, the covering number bound does not increase if the scale is reduced, while the accompanying risk bound does, leading to one using the limiting case as the scale tends to zero.

The training sample bounds we obtained for the fitted decision trees are summarized in Table 9.10. The second column indicates which results in

	$\mathcal{T}_5 : 5324$	664	64	$\mathcal{S} : 10$	$\mathcal{S}_i^- : 1$
ϵ	2.76808	1.568556	0.6346007	0.6390637	0.2842910
Error bound	7.864161	2.659692	0.5883787	0.5942359	0.241234

Table 9.8: Effect of different covering numbers on bounds

γ	$\ln \mathcal{N}_{1,7362}(\gamma, \mathcal{T}_5)$	*Risk bound*
2^{0}	48.66101	2.2231695
2^{-1}	53.66580	1.2286518
2^{-2}	58.66637	0.7339022
2^{-3}	58.66637	0.4889653
2^{-4}	63.68060	0.3683756
2^{-5}	68.20211	0.3058756
2^{-6}	68.20211	0.2746256
2^{-7}	68.20211	0.2590006
2^{-8}	68.20211	0.2511881
2^{-9}	68.20211	0.2472818
2^{-10}	68.20211	0.2453287
2^{-11}	68.20211	0.2443521
2^{-12}	68.20211	0.2438639
2^{-13}	68.20211	0.2436197
2^{-14}	68.20211	0.2434976
2^{-15}	68.20211	0.2434366
2^{-16}	68.20211	0.2434061
2^{-17}	68.20211	0.2433908
2^{-18}	68.20211	0.2433832
2^{-19}	68.20211	0.2433794
2^{-20}	68.20211	0.2433775
0	68.20211	0.2433775

Table 9.9: Effect of different scales on covering number bounds

Bound type	*Source*	dt80_symm	dt80_asymm
(Thresholded) reg dev error	(5.17)/(5.39)	0.2433756	0.2573062
(Thresholded) rel dev error	(5.19)/(5.41)	0.241234	0.2641177
Reg dev risk double	(5.27)	0.2553656	0.2645674
Reg dev risk dual	Thm 5.17	0.2155034	0.2092272
B-L ν-dev	(5.31)	0.2443235	0.2669476
P-H ν-dev	(5.33)	0.2203768	0.1621967
Random subsample	(5.37)	0.5008167	0.6864995
Chaining	(5.52)	1.290900	2.026900
Basic Rademacher	(7.7)	0.5565481	-
Refined Rademacher	(7.10)	0.5569935	-

Table 9.10: Training sample bounds for decision trees

the text were used to obtain the results. When a bound has adjustable parameters (such as a choice of β and/or ν), the best result over a grid or line search is reported in this table. Technically, the resulting value is not strictly a 95% upper confidence bound in this case, but it serves to indicate the power of the bound (i.e. the best one could have done had the variable parameters fortuitously been selected optimally in advance). To obtain a 95% upper bound, the weighted union bound can be applied over a grid. This typically has a rather small effect, which we shall consider later in this section.

9.5.1 Discussion of the bounds

In this section we provide commentary, clarifications, and further detail on aspects of the bounds in Table 9.10.

The first two bounds in the table are standard bounds in the form employed. For dt80_symm we applied the simple bounds on error, and for dt80_asymm the risk bounds from the thresholded loss class were applied. This required the covering numbers of the thresholded loss class based on the asymmetric loss, an issue we have not yet discussed. It turns out that the shatter coefficients of the loss class in this case is no larger than twice the shatter coefficient of the function class.

Double sample bound on regular deviation of risk The bound obtained in (5.27) is a generalized and strengthened form of (a one-sided, corrected[179], version of) Alon et al. (1993, Lemma 3.3).

The modifications include being able to choose the values of γ and β, and the replacement of the sample size restriction by employing a (more flexible) choice of α. As a comparison, an implementation of their result for these models yields bounds of 0.9457303 for `dt80_symm` and 1.507155 for `dt80_asymm`.

To make the comparison fair, we note that the result presented above optimized over the choice of β, which the result in Alon et al. (1993) fixed at 0.5, over a grid of 99 values. The values of β employed for the results presented in the table, 0.04 and 0.02 for the symmetric and asymmetric cases respectively, differ substantially from 0.5. For the bound to hold strictly, a weighted union bound must be employed over these grid elements. This increases the reported bound values to 0.2594169 and 0.2670770 for the two models, a rather minor change. For other comparisons we present, we will omit this modification, but it should be borne in mind if a grid is to be employed.

Naïvely employing $\beta = 0.5$ instead of using the grid yields comparable bounds of 0.2596510 and 0.2698580 for the two loss functions. Most of the improvement in the bounds we present thus seems to come from the fact that our generalized bound employs a scale tending to zero, while the scale used in the result from Alon et al. (1993) is related to the value of ϵ.

Dual sample bound on regular deviation of risk The bound based on Theorem 5.17 can be seen as an improved version of the main theorem of Devroye (1982). That result can only handle error, and only considers a shadow sample of size $m^2 - m$. We will compare them shortly, but we first address another issue.

The bounds presented in the table used the representation-independent

[179] The constant 12 quoted in their result should be 8.

bound on the covering numbers presented in (9.1). To show the influence of incorporating this information, we consider the bounds one could obtain by using this information, i.e. by employing (9.2). The bounds using the modified results are 0.2079258 and 0.1961643, which constitutes an improvement of about 0.01. The bounds were all obtained using optimization over the choice of β and the exponent of the sample size used. The optimal choices of β were 0.36 and 0.14 for the original results, and 0.38 and 0.24 for the results presented here. This illustrates a general observation made during the calculation of bounds for this thesis: typically the tighter bounds on covering numbers become, the larger the desirable choice of β becomes.

The most interesting observation here was that the optimal exponent was consistently around 1.4 to 1.5, and the choice of the power has a marked effect on the resulting bound, although performance is reasonably stable for exponents in the range 1.2 to 2.2. An exponent of 1.5 remained near optimal with other methods of bounding the covering numbers (such as the weak VC dimension based approaches) as well.

We now compare our bound with the bound implied by the result in Devroye (1982) in the case of `dt80_symm`. Without using the refined covering numbers, the value obtained from his bound is 0.2252533. This result is not much worse than the bound we present, and after applying the weighted union bound over the grids we employed, it may even be slightly better. These results serve as further confirmation of the idea that a shadow sample's size should be larger than the original sample for good bounds: the bounds obtained using this method are the tightest of the bounds obtained for `dt80_symm`, and the second tightest for `dt80_asymm`.

Bound based on B-L ν-deviation For the bound based on B-L ν-deviation, the values optimizing the bounds in this case were $\nu = 0.003, \beta = 0.1170175$ for `dt80_symm` and $\nu = 0.002, \beta = 0.05562021$ for `dt80_asymm`. (For these results, β was optimized numerically).

The bound is a generalized form of Bartlett and Lugosi (1999, Theorem 1). However, the formulation used there prevents choosing ν as small as we have, and fixes β at 0.5. That bound is minimized for both models at the smallest permissible choice of ν, $\sqrt{\frac{1}{m}} = 0.01648227$. The resulting bounds are 0.2518804 and 0.2787705, which are only slightly inferior to the bounds we obtained.

Bound based on P-H ν-deviation This bound attained its optimal values at $\nu = 0.32, \beta = 0.01$ for `dt80_symm` and at $\nu = 0.32, \beta = 0.07$ for `dt80_asymm`. The resulting bound for `dt80_asymm` was the tightest bound we obtained on the asymmetric loss.

The bound is a generalization of Haussler (1992, Theorem 3), which has one free parameter, ν. The optimal value of this bound is 0.9060489 (at $\nu = 1.07$) for `dt80_symm` and 1.907755 (at $\nu = 1.9$) for `dt80_asymm`. Clearly, our modification of the result yields much better results in this case. The reasons for this improvement are not at all clear, but it may be due to the fact that the original results fix both the value of α and β employed in their symmetrization lemma, subject to a sample size restriction. This restriction is later dropped since their bound is trivial for smaller sample sizes. The form of our result uses a choice of α which is sensitive to the choice of β. On the other hand, the other bounds we have considered do not seem to be as sensitive to the choice of β. This apparent discrepancy seems worth further investigation.

Random subsample bound on regular deviation This bound performs significantly worse than all the other bounds considered so far. It is the first and only bound we consider making use of the random subsample lemma. This performance is as expected — we noted in Section 5.5.7 that the regular double and dual sample covering number bounds tend to be asymptotically tighter than bounds employing the random subsample lemma.

Chaining bound The chaining bound presented here is not technically correct, but only presents the sum of the first hundred terms of the chaining bound as a fairly good approximation. This works because the VC dimension of the class is finite, so that the error in such an approximation can in principle be bounded. For general function classes, the covering numbers approach infinity as the scale approaches zero, so an alternative plan for obtaining a good approximation must be used.

It is also important to note that the size of $\mathcal{B}_j$ in the chaining bound is expressed in terms of covering numbers w.r.t. the 2-norm, not the 1-norm. A bound on the 2-norm covering numbers was obtained from the 1-norm covering numbers by employing (5.5).

Despite the fact that the chaining bound is technically refined, its performance here is poor. The most likely reason for this is that the covering numbers at various scales are bounded by employing the VC dimension. These covering number bounds are rather poor, and we expect that chaining will perform much better if alternative bounds on such covering numbers are available.

Rademacher bounds The Rademacher bounds we present employ the last part of Theorem 7.11 to bound the Rademacher complexity, using (9.3). This approach yields a bound on $\mathcal{R}_{3681}(\mathcal{T}_5)$ of 0.1982399 (when $\delta = 0.00150262$), while employing the third part of Theorem 7.11 yields a bound of 0.2652305.

We note that the resulting bound performs similarly to the bound using the random subsample lemma, i.e. poorer than the dual sample and double sample bounds. This is also as expected. If we bound the Rademacher complexity by employing the shatter coefficients as we do here, we have a bound on maximal deviations of *roughly*[180]

$$2\sqrt{\frac{2\mathcal{N}_{\mathcal{W}}(m) - \log \delta}{m}} \quad .$$

Comparing this to the result obtained when inverting the bound of (5.37), we see that their overall structure is highly similar.

[180] This is obtained by combining (7.7) and the inner portion of the last part of Theorem 7.11, and suppressing a number of other terms.

To obtain results from global Rademacher bounds which improve on the dual sample bounds, it is necessary to employ methods to obtain tighter Rademacher bounds than those obtained from Theorem 7.11.

9.6 Bounds on boosted stumps

In this section, we investigate training sample bounds applied to the model fitted by boosting stumps, `boost80s`. The fitted model was obtained by running

```
> boost80s<-ada(spam80[,2:58],spam80[,1],type="discrete",iter=200,
+ control=rpart.control(maxdepth=1,cp=-1,minsplit=0,xval=0))
```

This yielded the following fitted model:

```
> boost80s
Call:
ada(spam80[, 2:58], y = spam80[, 1], type = "discrete", iter = 200,
    control = rpart.control(maxdepth = 1, cp = -1, minsplit = 0,
        xval = 0))

Loss: exponential Method: discrete   Iteration: 200

Final Confusion Matrix for Data:
          Final Prediction
True value    0    1
         0 2157   80
         1  184 1260

Train Error: 0.072

Out-Of-Bag Error:  0.077  iteration= 190

Additional Estimates of number of iterations:

train.err1 train.kap1
       196        196
```

Informally speaking, the class of unthresholded functions which can be returned by a general boosting algorithm lies in the convex hull of the class of functions implemented by the base classifier. If, as in our case, the number of iterations is limited, the class is more restricted to linear combinations employing a fixed number of terms. However, both of these ways of viewing the unthresholded class yield poor bounds on covering numbers. We have already discussed this issue for the first viewpoint in Section 9.4.2. The major issue with the second viewpoint is that it is not clear how to use the fact that the class is restricted in order to obtain better covering number bounds.

Thus, for the boosting model under consideration, we do not have access to reasonable covering number bounds, precluding the use of most of the bounds applied to the decision trees in the previous section. The bounds we will consider in this section are Rademacher bounds based on a Lipschitz proxy loss (see Section 9.4.3), and double/dual sample PAC-Bayesian bounds, which employ the covering numbers of the class of stumps.

Since we specified that the boosting algorithm will use stumps, the underlying size of decision trees is fixed a priori at one split. Thus no class hierarchy over the decision trees is necessary in this case, so bounds can be applied using $\delta = 0.05$.

Rademacher bounds The actual model generated by the `ada` package yields weights which do not sum to one, and the resulting weighted sum is thresholded at 0.5, since the base stumps actually make predictions in $\{0, 1\}$. As a result, a few linear transformations of the weights and the outputs are necessary in order to calculate unthresholded values for application of the Rademacher bounds (specifically, the unthresholded values are needed to calculate the losses of the proxy loss function on the training points).

From the discussion of Section 9.4.3, it follows that if we employ the proxy loss v_K, we can bound the Rademacher complexity of the proxy loss class under consideration by

$$2K\mathcal{R}_{3681}(\mathcal{S}) + \sqrt{\frac{1}{3681}} \ .$$

(Recall that the boosting was performed using stumps, and $\mathcal{L}_1 = \mathcal{S}$).

We obtain a bound of 0.1152925 for $\mathcal{R}_{3681}(\mathcal{S})$ using the probabilistic bound on Rademacher penalties. Unfortunately, this bound is still too large, yielding a trivial bound in this case. Various bounds can be obtained by using (7.7) with varying choices of v_K, however, in this case, the bounds tend to improve as $K \to 0$. In the limiting case, v_K assigns a loss of 1 to all the points, and the bound reduces to

$$1 + 2\sqrt{\frac{1}{3681}} + \sqrt{\frac{-\ln(0.05)}{7362}} = 1.053137 \ .$$

The same behaviour occurs for the bound obtained from (7.10), with the resulting bound being 1.053409.

Double/dual sample PAC-Bayesian bounds This section will consider double sample and dual sample PAC-Bayesian bounds based on the results in Examples 8.3 and 8.4. Recall that the PAC-Bayesian bounds are based on stochastic classifiers using the Gibbs strategy. For the boosting algorithm, the weight vector obtained training the algorithm can be seen as a (discrete) distribution over the class of stumps. The combined classifier then implements the MAP strategy based on that distribution: it classifies according to the most likely class predicted if a stump is sampled according to the weight vector.

PAC-Bayesian bounds, on the other hand, apply to the Gibbs strategy: each new point is classified by sampling a stump according to the weight vector, and making a prediction employing the stump. The PAC-Bayesian bounds we shall consider bound the risk of this classifier.

This classifier's performance is substantially worse than the other classifiers we have considered, perhaps mainly due to the fact that the base classifiers are so simple — each stump's individual performance is quite poor. The training error of this Gibbs classifier is 0.3220136.

In order to implement these bounds, we used the earlier results for bounding the covering numbers of the class of stumps, which corresponds to the class $\mathcal{H}'$ in the bounds referenced. The double sample bound in this case evaluated to 0.419319, which is the closest a bound has come to the training risk amongst the bounds considered.

However, we were able to obtain even tighter bounds by employing the dual sample bound, which gives us the flexibility to consider larger shadow sample sizes. The optimal settings for this bound were a shadow sample size of $u = 3681^{1.5}$ and $\beta = 0.55$. These settings yielded a bound of 0.3770343. (Using the refined covering number bound from (9.2) yielded a reduction to 0.3736865 at $u = 3681^{1.5}, \beta = 0.65$).

As an indication of the tightness of this bound, the test error of the Gibbs classifier is 0.3240981, slightly higher than the training error. Applying the Hoeffding r.e. inequality to obtain a 95% test sample confidence interval on this classifier, one obtains an upper bound of 0.3644480.

Without any doubt, these bounds are the most successful we have considered. Unfortunately, they come at the cost of having to employ the Gibbs classifier. For certain problems, one can approximate a deterministic classifier arbitrarily closely by a related Gibbs classifier, allowing one to obtain bounds using this powerful strategy. For more details on this aspect, the interested reader is referred to Langford and Caruana (2002).

9.7 General discussion

Improving covering numbers The training sample bounds applied to the decision trees clearly showed the improvements that can be obtained with covering number-based bounds if some effort is spent to obtain good covering number bounds, rather than naïvely applying the simplest tools available. We showed that if the class size can be restricted by employing the representation of the data, the bounds could also be improved. As we mentioned before, any computer implementation of an algorithm automatically restricts the function class to a countable set due to its representation of numbers using a finite set of bits. Employing this restriction may further lead to tighter bounds.

A related idea is that for good training bounds, the representation of the data should be as coarse as possible, and the range as restricted as possible. For example, if we had prior information that the first 48 features of any point would never have a value of more than 5%, and the measurements would be multiples of 0.05% (i.e. feature values would be rounded off/truncated as necessary), instead of 20002 potential splits on a feature, there are only 202. Similar limitations could be imposed on the other features.

A further avenue for potentially reducing covering numbers which we have not investigated is relationships between features. In the e-mail classification problem under consideration, there are some inherent relationships between the features: the sum of the first 48 features can not exceed 100, and neither can the sum of features 49 to 54. Furthermore, features 55 and 56 can not exceed feature 57, and feature 57 is a multiple of feature 55. Exploiting these relationships to tighten the bounds may be possible, and seems worth investigation.

Bounds not calculated A number of bounds described in this thesis were not implemented on this data set, or the algorithms considered. We shall briefly consider the algorithms we did not present here, and why they were not presented.

- A simple Occam's razor bound could not be applied because the class of decision trees (of a given size) is uncountable.
- The margin bounds were investigated, but the covering numbers exploded on the convex hull of the class of stumps, as described in Section 9.4.2.

- Sample compression bounds were not applicable since we did not have a useful compression scheme formulation of decision trees.
- Luckiness and algorithmic luckiness bounds were not applicable, since no useful luckiness function was found.[181]
- To apply local Rademacher bounds, one needs bounds on the subclasses with control on the variance. It is unclear how such bounds can be obtained for this problem.
- The general PAC-Bayesian bounds and the Occam's Hammer bound both require "prior" distributions on the class of classifiers. I was unable to find a sensible way to express distributions over the class of all decision trees, or even of all stumps.
- Shell decomposition bounds are defined on a countable hypothesis class. They can be extended to continuous spaces by employing PAC-Bayesian arguments, but then a "prior" over the class needs to be defined.

For a different problem and loss function, some of these bounds would have been more applicable, while traditional covering number approaches may not have been practical, or would have yielded terrible results. Generally speaking, the structure of the input space together with the algorithm employed determine which bounds are suitable candidates for implementation.

Other bounds on decision trees The bounds provided in this thesis are by no means exhaustive. Furthermore, various algorithms spawn custom-made bounds which apply knowledge of the structure of the algorithm under consideration.

The microchoice bound (Langford and Blum, 1999),which arose from the concept of self-bounding algorithms introduced in Freund (1998), is generally

[181] Luckiness functions such as the one in Example 6.7 are not worthwhile unless the bounds obtained on the luckiness function are sample dependent.

a very competitive bound for decision trees with binary features. However, when the features are continuous, this bound is no longer applicable.[182]

Two other approaches to obtaining bounds for decision trees are presented in Golea et al. (1998) and Mansour and McAllester (2000). The first article employs the view of a decision tree as the convex hull of the leaf functions, but uses a more advanced concept of the *effective number of leaves* of a tree. It uses these ideas to present a margin bound on decision trees. This margin bound is, however, based on the covering numbers of the base class, rather than its convex hull. These results, which are based on the margin bounds presented in Schapire et al. (1997), are unfortunately based on unspecified constants. Mansour and McAllester (2000) present what the authors call a *compositional* bound — the problem of optimizing the bound over the class of decision trees can be decomposed into independently optimizing the left and right subtrees and a little extra work. They also note that the approach used in Golea et al. (1998) only performs well when "almost all training data reaches a single leaf". On the other hand, a drawback of their method for our sample is that they assume the class of splits is countable.

Another very recent result which is highly relevant here is Shah (2007), which presents sample compression bounds for decision trees. His approach uses a more general sample compression approach than that outlined in this thesis. In the alternative scheme he employs, the compression function maps the sample to a subsample as well as an extra message with further information for the reconstruction function. This modified approach, which seems to have developed from the PAC-MDL bound proposed in Blum and Langford (2003), makes the sample compression framework much more flexible. The general approach, developed in the author's doctoral thesis (Shah, 2006), allows one to employ a data-dependent "prior" over the compression set and the class of possible messages, with each individual hypothesis being bounded by an arbitrary bound (as with Occam's hammer, one can employ bounds on the binomial tail deviation).

[182] Technically, the bound can handle splits on features with multiple values, but not cases where the size of the class of splits is infinite.

Shah (2007) suggests appropriate compression and reconstruction functions in this framework, as well as advice on the selection of appropriate "priors". This bound handles the uncountability of the class of splits by effectively discretizing the class up to an accuracy acceptable to the practitioner. However, the higher the desired accuracy, the looser the bounds will become. The article provides empirical results of his bound on a number of benchmark data sets. In all cases the bounds were non-trivial, despite the trees constructed being fairly large (11-61 nodes). Unfortunately, I became aware of these developments too late to include a fair reflection of these results in my thesis.

A fair comparison The reader may have noticed that the comparisons of training sample bounds and test sample bounds presented in this chapter are not entirely fair. Since training sample bounds dispense with the need for a test sample, a fair comparison should compare the training sample bounds applied to models fitted on the full data set to test sample bounds on models fitted using the reduced 80% training sample.

We expect better performance from such training sample bounds than we obtained from the training sample bounds applied to the 80% training sample for two reasons. First, we expect the training risk of the model fit on the full data set to be smaller than that fitted on the 80% training sample. Second, we would expect the width of the interval to be smaller.

For the spam data, we employed the same algorithms on the full training sample as well, obtaining the models `dt100_symm`, `dt100_asymm`, and `boost100s`. However, the improvement in training risk was very small — less than 5% of the training risk on the 80% sample in all three cases. In addition, the resulting decision tree models both had an extra split, so that the SRM "prior" is more punitive, partially countering the potential gains of a larger sample size. In general, the training sample bounds obtained were similar to those obtained on the 80% training sample, so we elected not to present them, since it would be unclear how the differences between the bounds should be attributed to differences in decision rule risk, training sample size, and differences in model complexity.

The most competitive bound we considered was the dual sample PAC Bayesian bound applied to the boosting algorithm. As an illustration, we note that the training error of the Gibbs classifier corresponding to the fitted model `boost100s` has a training error of 0.3379392 (larger than that of `boost80s`). The corresponding bound was 0.3829682. The term added to the training risk for this bound decreased from 0.0550207 for `boost80s` to 0.045029, predominantly due to the increase in the sample size. If the training risk had improved notably due to the increase in size of the training sample from `boost80s` to `boost100s`, one can imagine that this bound may well have outperformed a test sample bound obtained on `boost80s`. This type of situation is most likely to occur when data are rather scarce, and withholding new data is likely to hamper the quality of the resulting decision rule considerably. However, the problem in this situation is that bounds for the Gibbs classifier are being compared, rather than the actual fitted model.

9.8 Conclusion

The bounds presented in this chapter clearly illustrate that parameters such as the β from the symmetrization lemma, and the shadow sample size u in the dual sample lemma can have a notable influence on training sample bounds. Furthermore, we note that by using a more flexible choice of scale in double sample risk bounds, we were able to get bounds which were not much poorer than double sample error bounds. This indicates that the appropriate choice of scale should be sensitive to the structure of the loss class, and that the traditional default of using a fraction of ϵ is not necessarily appropriate.

These observations also suggest that some improvement can be attained by generalizing the available double sample bounds for error to dual sample bounds, and deriving a symmetrization lemma for error in terms of appropriate choices of α and β. In addition, it might be worth investigating extending bounds on the P-H deviation to more general shadow sample sizes, and obtaining a tighter symmetrization lemma for error, since the P-H

ν-deviation bound performed quite well without these refinements.

Although the training sample bounds were, on the whole, poor in comparison to the test sample bounds, many were by no means trivial. Furthermore, the results were obtained using covering numbers which were in no way data-dependent: suprema of covering numbers were used throughout. The effects of optimizing covering numbers directly, rather than simply employing bounds on the VC dimension, was also clearly illustrated.

A selection of functions employed for the calculation of many of the estimators presented in this chapter appear in Appendix C.[183] One interesting feature of the code is the calculation of the α employed in the bound for a given β: the code selects the value of α in response to the size of β and the shadow sample u, by choosing the best of the candidate α functions considered in the thesis.

[183] Disclaimer: these functions were written by me, for me! They would need thorough sanitation before being made generally available. If you wish to use them, it is strongly recommended that you verify them against the theory first, since they were often implemented for very specific circumstances.

Chapter 10

Conclusion and Future Research Directions

This chapter considers the progress made in the various objectives for the thesis. We review the contributions and progress made, and consider the way forward for further progress in the field.

10.1 Review of objectives

Chapter 2 presented a generalization of David Haussler's decision-theoretic model of learning which incorporated a new component which we called a strategy. This component allowed us to deal with cases where the decision rule is stochastic. Notions of risk of an hypothesis and of a decision rule were introduced. In Chapter 3 we introduced the idea of viewing interval estimators from the perspective of bounding measures of deviation. The rest of the results in the thesis were discussed using these perspectives.

The thesis introduced a number of approaches to risk estimation, with a focus on training sample interval estimation. In order to keep the material accessible, plenty of background information was provided. The results presented are a reasonably comprehensive introduction to the major approaches to training sample interval estimation of risk.

The third objective concerned presenting generalized results allowing more flexibility in the selection of various parameter values. This approach was successfully employed to generalize a number of classical and data-dependent covering number-based and PAC-Bayesian bounds. These more general forms allowed for flexible selection of the scale parameter γ, the parameter β used in the relevant symmetrization lemma (as well as the choice of the corresponding α), and the shadow sample size u. These generalizations and other contributions of the thesis are discussed in more detail in Section 10.2

These extended results were compared to the classical results in Chapter 9, where we noted that the choice of scale in covering number bounds can have a sizeable effect. Furthermore, we saw that a double sample approach to symmetrization was not near optimal for our problem. Instead a shadow sample of size $u \approx m^c$ performed well, with $1.2 \leq c \leq 2.2$. For other work investigating alternative shadow sample sizes, the reader is referred to Catoni (2004a). Our investigation of the choice of β in symmetrization lemmas indicated that the impact of this choice is typically rather small, and that the tighter a bound is, the larger the appropriate choice of β seemed to be.

Furthermore, we considered the benefits which could be reaped by focusing more closely on direct bounding of covering numbers instead of using bounds on dimension measures of large classes to obtain them. We illustrated that this approach can yield dramatic improvements.

Generally speaking, the training sample interval estimators we obtained for the data set and algorithms we considered were simply not competitive with those obtained by employing a test sample. On the other hand, many of the results were certainly not trivial, and we illustrated that the PAC-Bayesian methodology could yield competitive bounds for decision rules employing the Gibbs strategy.

In general, none of the training sample bounds we presented are uniformly better than the other methods, and the appropriate choice among these techniques depends on the specific loss class being used for the problem. In certain cases, good bounds on covering numbers of the loss class are available, but not for Rademacher averages, while in other cases the situation is the opposite. When a Gibbs classifier is being employed, the PAC-Bayesian methodology will typically yield better results than covering number- and Rademacher average-based approaches. Thus, the appropriate choice of training sample bound for any given problem will depend on a fine analysis of the loss class, which is currently beyond the expertise of the typical practitioner.[184]

Note that we did not present numerical results for all the approaches presented in the thesis in Chapter 9. As a result, it is not clear how competitive the other approaches could be made in practice. In principle, many of these bounds could be substantially tighter than the ones we evaluated, but it is often very difficult to get useful bounds on the values which enhance the results.

Furthermore, the results presented in this thesis can not hope to be exhaustive. Our aim was to provide representative results of good bounds for the approaches considered. For a number of approaches, the results we present are not the state of the art, and in many cases improvements can be obtained under certain conditions. In particular, tighter bounds are often available for certain classes of algorithms. Notable in this regard are bounds based on algorithm stability, local Rademacher bounds for ERM, bounds for convex loss functions, and bounds employing conditions on the noise level in a model[185]. Once again, however, such bounds are only useful if they can be successfully evaluated. We will return to the difficulty of evaluating training sample bounds, and possible ways of alleviating the situation, in Section 10.4.1.

[184] We discuss this further in Section 10.4.1.

[185] For a discussion of these approaches, the reader is referred to Bousquet and Elisseeff (2002) and Koltchinskii (2006) for the first two approaches, and Bousquet et al. (2005) for overviews of developments for the other two approaches.

10.2 Contributions

In this section, we will consider the contributions made in this thesis, focussing on our third objective, viz. presenting more general forms of training sample bounds. Probably the largest contribution consisted of modifying various bounds employing covering numbers to allow arbitrary choice of β, u, and γ by the practitioner, rather than being restricted to the values of these parameters chosen for convenience in earlier derivations.[186]

β is a parameter employed in various symmetrization lemmas. A corresponding choice of α is typically obtained by using a concentration inequality. We point out that one can actually select α on the basis of more than one concentration inequality. Furthermore, traditionally, once the value of β was fixed (typically at $\beta = \frac{1}{2}$), the choice of α was fixed by employing a restriction on the sample size m for which the results presented held. Our approach removes the sample size restriction by directly using the appropriate choice of α. This allows improvements in the result when the shadow sample size exceeds the minimum sample size specified by the traditional results. The final bounds presented retain β as a parameter, allowing the practitioner to select β as he wishes, or to optimize over β with an appropriate union bound argument. The potential value of alternative choices of β besides $\frac{1}{2}$ is illustrated by the alternative choice used in Shawe-Taylor et al. (1993).

The size of the shadow sample u in nearly all traditional covering number results is set to equal the sample size m. When this traditional wisdom has been questioned, improved results have typically been obtained, as in Catoni (2004a), Devroye (1982), Shawe-Taylor et al. (1993). Our approach provides bounds where the shadow sample size can be specified by the practitioner.

Finally, it is customary for covering number based bounds on error to be stated in terms of shatter coefficients, while bounds on risk were stated using covering numbers. The scale γ of these covering numbers were typically chosen as a fraction of ϵ, the bound on the measure of deviation under

[186] Of course, the typical practitioner is highly unlikely to know the appropriate choices to make, so the development of heuristics for selecting these parameters is important future work — see Section 10.4.

consideration. While this was convenient, it is not necessarily a good choice. We present results for bounding error in terms of covering numbers, and bounds for risk where the scale can be selected by the practitioner. The bounds on error reduce to the traditional shatter coefficient bounds as the scale becomes arbitrarily small. Our experiments further indicate that for risk based on asymmetric loss the benefits to be reaped by alternative choices of scale are substantial.

Flexibility in the choice of u and β is typically attained by the use of the more flexible forms of symmetrization lemmas which we present in Theorems 5.4, 5.5, 5.6, 5.7, 5.15, 6.1 and 6.2.

The freedom to choose $m \neq u$ also relies on an appropriate dual sample lemma for the measure of deviation under consideration. We only provided such results explicitly for regular deviation and for the realizable case — see Theorems 5.10 and 5.14. However, similar arguments were employed to obtain the data-dependent bound of Theorem 6.3 in Chapter 6.

The other modification to many of the bounds, allowing the practitioner the freedom to choose the scale γ of the covering numbers, could generally be achieved by careful bookkeeping.

Combining these approaches led to many of the bounds in the thesis based on covering number arguments generalizing previous results. These include the various bounds in the latter part of Chapter 5 (except for the chaining bounds), Chapter 6, and Section 8.1.3.

The dual sample bound of Theorem 5.10 was obtained by generalizing an argument presented for error in Devroye (1982). However, the result we obtained is applicable to risk as well. This allowed us to obtain a new simple covering number bound on regular deviation of risk with faster decay than previous bounds of the same type. Bounds with the same decay rate have been obtained with more sophisticated tools such as chaining, however. For further information, see the discussion after Theorem 5.17, until the end of the subsection. Our formulation of the dual sample bound also permitted us to extend the PAC-Bayesian bounds presented in Section 8.1.3 to general loss functions, although the original results by Catoni were restricted to error.

We also present a generalized form of the shell decomposition bound in Section 8.2, which holds for general loss functions. In addition, we allow the practitioner to select the number of bins used for the bound, and introduce two “priors”.

Section 5.6 presents a somewhat more general view of margin bounds than is traditional. We expect the greatest contribution from this perspective to be the more general margin concept for zero-one loss functions presented in Section 5.6.3 which views the margin as the distance to a boundary between various classifications.

We have already discussed the modifications we made to Haussler’s framework in terms of introduction of the strategy concept. In addition, we attempted to deconstruct known results into basic theorems.[187] The idea of viewing interval estimators as inverting bounds on various measures of deviation was valuable in this sense.

Finally, a number of variations on existing results were presented in what we felt were more convenient forms. For example, the bound in (7.12) can be extracted from the proof of Theorem 2.1 in Bartlett et al. (2004). We feel this form is more generally applicable, since one has not yet committed to various relaxations they employ in the rest of the proof. Similarly, some results in Section 8.1 parallel results in Catoni (2004b), except that our arguments used results underlying Hoeffding’s inequality, rather than Bernstein’s inequality.

10.3 Utility of training sample bounds

The original training sample bounds were based on convergence theorems. However, the conditions under which uniform laws of large numbers and

[187] Most of our generalizations flowed from generalizing these basic theorems, and then applying similar reasoning as before.

uniform central limit theorems hold are now fairly well understood (see for example Dudley, 1999). Furthermore, upper and lower bounds on the asymptotic rate of convergence are available in many cases (e.g. Anthony and Bartlett, 1999). However, while the asymptotic rate is often known exactly, or up to a factor of $\ln m$, the actual gap between the upper and lower bounds tend to be extremely large due to the constants employed.

Training sample bounds arise in response to the question of what else can be done with these convergence results. When applied as is, these bounds were not suitable for practical use as interval estimators. However, researchers proposed other uses for the resultant bounds. By considering the quantities influencing the bound, new insight into the factors influencing the risk of a decision rule could be obtained. For example, the covering numbers employed in bounds provide theoretical support for the ideas of capacity control and regularization.

Furthermore, it was suggested that the bounds could be used as heuristic tools to obtain the appropriate level of capacity control. Thus, it was proposed that training sample bounds could be used for model selection.[188]

Another application of training sample bounds was the design of new algorithms. The most well-known example of this is undoubtedly the support vector machine. This algorithm was motivated by the heuristic of finding a separating hyperplane which minimized a training sample bound.

It should be clear from these applications that progress in training sample bounds is a worthy pursuit, regardless of the fact that the bounds we have considered do not seem to be competitive with those obtained using a test sample: clearly, tighter bounds will be more predictive of generalization, making the bounds more useful for identifying factors influencing risk, model selection, and designing new algorithms.

It seems that for training sample bounds to become consistently competitive with bounds based on a test sample, at least one of two conditions needs

[188] A recent book on the use of concentration inequalities and the resulting training sample bounds for model selection is Massart (2006).

to be met: either a new theoretical breakthrough in bound design will be needed, or effective methods for obtaining (good bounds on) the quantities in the most refined bounds must be found. In addition, progress in making the bounds available to practitioners will be needed (see Section 10.4.1).

We conclude this section by noting a few restrictions on the applicability of the results in this thesis.

The most notable restrictions in the thesis are that of bounded loss functions and independent, identically distributed data points. Some work has been done on training sample bounds for unbounded loss functions, but the results are by necessity not distribution-free (e.g. Vapnik, 1998). Some of the theorems we employ can be extended to martingales, and for some other results identical distributions are not required (e.g. Catoni, 2004a). In addition, many of the results we present can be generalized to various mixing processes. These generalizations are discussed in Vidyasagar (2002).

Another restriction which one must always be aware of when considering training sample bounds are the various independence requirements. Notably, “priors” employed for a specific bound typically need to be specified before seeing any of the training data.

10.4 Further research

The work done in this thesis suggests a number of avenues for future work.

Our work has introduced new parameters in bounds which can be specified by practitioners, or optimized using a weighted union bound. It is natural to investigate the effect of these parameters, and to try to obtain guidelines or heuristics for the choice of these parameters or appropriate “priors”. Furthermore, the interactions between the parameters may also be important. Consider for example the choice of the shadow sample size u. As u increases, the variance of $r_P(w)$ decreases, making the regular deviation $r_S(w)-r_P(w)$ an increasingly accurate reflection of $r_S(w)-r_D(w)$. On the other hand, the covering numbers under consideration increase as the sample size increases.

The appropriate choice of u seems to represent a trade-off between these factors, but the optimal trade-off for the resultant bound is likely to depend on the value of β and the corresponding α in the symmetrization lemma employed.

Another challenge is to attempt to obtain dual sample bounds for other measures of deviation besides regular deviation. As a related issue, we note that besides the classical covering number bounds of Chapter 5, all the training sample bounds we considered involved bounds on regular deviation. We suspect obtaining analogs of these other approaches in terms of other measures of deviation would be valuable. The only work I am currently aware of in this vein is the work on data-dependent bounds on relative deviation in Andonova Jaeger (2005).

Another potentially useful activity is investigating the values of unspecified constants in classical results. This includes refining constants where poor constants are presented as well as obtaining bounds on unspecified constants. Both lower and upper bounds on such constants would be useful. For example, knowledge of (bounds on) the values of the absolute constants in Theorem 7.5 may have allowed us to obtain tighter bounds on the decision trees investigated in Chapter 9, perhaps using the results presented in Bartlett and Mendelson (2002).

This thesis has not explicitly discussed the concept of bracketing covering numbers. This concept, which was already employed for proving a uniform version of the strong law of large numbers in Pollard (1984) (see Theorem II.2.2), and appears in a number of aspects of empirical process theory (e.g. Dudley, 1999, van der Vaart and Wellner, 1996), is closely related to the general covering numbers we do consider. Specifically, the bracketing covering number of a class $\mathcal{V}$ can be shown to be closely related to the external covering number using the metric $d_{1,D}$. Langford (2002, Chapter 9) provides an extension of PAC-Bayesian bounds for finite hypothesis classes to infinite classes using this covering number. These results suggest that future work on obtaining bounds using these covering numbers may yield tight training sample bounds. In addition further research into bounding such covering numbers may be useful.

Another field for future research is investigating effective ways to obtain good “priors” or reduce the effective size of function classes by interaction with practitioners. For example, if the practitioner could specify the likely range and maximum meaningful resolution for each feature of the data set, covering numbers could in practice be dramatically reduced.

Another approach worth investigating is the concept of training sample bootstrap interval estimators. It is not clear to me how such an estimator might be constructed, but if one could, it may well be an extremely useful and powerful tool for risk estimation.

10.4.1 The way forward

In this section, we briefly discuss making training sample bounds more available to practitioners, and suggest implementing a website for enhancing the effectivity of researchers in the field while lowering the barrier to entry for newcomers to the field.

Perhaps the most useful application of training sample bounds at the moment is model selection. However, their applicability is severely limited by a number of factors.

The first is the question of whether model selection using training sample bounds outperforms the much more well-known methods such as the hold-out method, cross-validation and the various approaches mentioned in Section 5.1.2. Note that performance here should also take into account whether it is necessary to fit multiple models.

Even if it can be shown that training sample bounds are suitable for model selection, a much greater barrier to their application is the large amount of technical knowledge necessary to apply them effectively. In this respect, availability of code, and even integration of model selection employing training sample bounds into packages implementing various algorithms would be invaluable. Furthermore, this approach may allow one to calculate bounds which one could not calculate any other way. As an example, microchoice bounds (Langford and Blum, 1999) need to evaluate the number of choices an algorithm makes during its execution in order to calculate the bound. Currently, this bound is only available for custom decision tree implementations, to my knowledge. Incorporating this bound into a standard decision tree package for a system like `R` by collaborating with the package authors will make the bound much more useful. Such an approach will also make it more likely that bounds will be correctly implemented in packages. For a recent discussion on the potential benefits of wider availability of source code to a research community, from the perspective of the machine learning community, please see Sonnenburg et al. (2007).

A final recommendation for encouraging progress in the field of training sample risk bounds is the establishment of an online repository for information, similar to that available for researchers in kernel methods at `http://www.kernel-machines.org/`.

In the late 1990s and early 2000s, the European Strategic Program on Research in Information Technology (ESPRIT) funded a working group focusing on neural networks and computational learning theory, known as NeuroCOLT, and later NeuroCOLT2. One of the focus areas of this group effectively involved deriving training sample bounds, and most of the development in the field of training sample bounds at the time was due to members of this group. The group's website at `http://www.neurocolt.com`, which provides a repository of the group's technical reports, is still a valuable source of publications in the field.

As such, our suggestion is not entirely unprecedented in the field. However, we suggest a broader, more inclusive, repository with more community involvement. We briefly mention a few potential benefits of such a site.

- Links to publications relevant to the field can be collected centrally. Furthermore, detailed proofs which could not be published due to article length constraints, source code, and errata or modifications[189]

[189] In this thesis, a variety of flaws and inaccuracies were discovered in various sources,

could be made available with the article.

- Issues which are unclear in a publication can be resolved once, and collected with other information on the publication. Later readers of the publication can then find this clarification with little effort.

- Publications in the field, or highly relevant to the field, tend to be scattered throughout a variety of journals. Announcing relevant new publications on the site will help researchers stay up-to-date on recent developments more easily. Similarly, the release of software implementing various bounds could be tracked.

- A record of performance of various training sample bounds on a variety of benchmark data sets and algorithms can be maintained, allowing researchers and practitioners to more easily assess the state of the art.[190]

and we have pointed out a number of them that were relevant to our arguments. In many cases, when these inaccuracies are found, it is difficult to make other researchers aware of them. Furthermore, if they do not affect the asymptotic results, many researchers will not consider the inaccuracies important.

[190] One complication with this approach is the matter of verifiability of “priors”. However, the problem does not seem insurmountable.

Appendix A

List of Symbols

Symbol	*Usage*
$\mathcal{A}$	action class for the learning problem, with typical element r
$\mathscr{A}$	r.v. corresponding to α in hierarchical Bayes estimation
A	generic set; diagonal operator
$A(S, w)$	high confidence region for $r_D(w)$
A_j	subset of $\mathcal{W}$ corresponding to a link pair — typical element of $\mathcal{D}_j$
Aluck	algorithmic luckiness function
a	acceleration constant in BC_a bootstrap interval; natural number
a_j	diagonal elements of diagonal operator A
absconv(A)	absolute convex hull of A
α	parameter of symmetrization lemma; "prior"; parameter of Beta distribution
$\alpha^\star$	"prior" used in conjunction with symmetrization lemma
$\alpha(u, \beta)$	function yielding appropriate parameter of symmetrization lemma
$\mathcal{B}_j$	j-th link set for chaining
$\mathscr{B}$	r.v. corresponding to β in hierarchical Bayes estimation
B	number of bootstrap samples; ball in a space
$B_\varepsilon(R)$	ε-blowup of the set R
$B(\alpha, \beta)$	Beta function
Bin(k, p)	binomial distribution with parameters k and p
b	bootstrap sample index; base of a logarithm; control on the norm in concentration inequalities

Symbol	*Usage*
β	parameter of symmetrization lemma; parameter of Beta distribution
$\mathcal{C}$	concept class, with typical element c
$\mathscr{C}$	generic class of sets, with typical element c
$C(k,\epsilon)$	probability bound for a concentration inequality
C_P	Mallows' C_P statistic
c	generic real value; concept — typical element of $\mathcal{C}$; set — typical element of $\mathscr{C}$
$\mathrm{conv}(A)$	convex hull of A
χ^2	chi-square distribution
$\mathcal{D}_j$	partition of $\mathcal{W}$ based on the link set $\mathcal{B}_j$, with typical element A_j
$\mathscr{D}$	diameter of a set
D	distribution generating input-output pairs
D'	a modified version of D with independent inputs and outputs
D_p	integral used for deriving relative deviation bounds
d	generic metric, pseudometric or prametric
d^H	Haussler extension of a metric
$d_{p,Q}$	metric on the Lebesgue space $L^p(Q)$
$\bar{d}$	point-set extension of d
$\varDelta$	level function for Occam's hammer
δ	confidence level
$\delta_{\mathcal{X}}$	class of Kronecker delta functions
δ_x	Kronecker delta function
$\digamma$	multiplicative constant
$\mathcal{E}$	generic space, with typical element η
$\mathscr{E}$	generic event or predicate
E	generic r.v., typically taking values in $\mathcal{E}$; generic stochastic process
(E_1, E_2)	constants for a Euclidean class
Ent	entropy
env	envelope
e_P	error with respect to distribution P
erf	Gauss error function
ϵ	probabilistic bound on a measure of deviation

Symbol	*Usage*
ϵ_0	estimate of optimism in the bootstrap world
ε	noise term in regression setting; parameter for ε-insensitive loss; parameter for ε-blowup of a set
η	generic element of $\mathcal{E}$; generic real value or vector
$\mathcal{F}$	loss class, with typical element f
$\mathcal{F}_i$	subclass of $\mathcal{F}$, used for SRM
$\mathscr{F}$	Fourier transform
Fi	fold i for cross-validation
F_Q	c.d.f. of a distribution Q
f	typical element of $\mathcal{F}$
fat	fat-shattering dimension
fatV	level fat-shattering dimension
$\mathcal{G}_{\mathcal{H}}(\mathcal{Q})$	Gibbs class associated with $\mathcal{H}$ and $\mathcal{Q}$
$\mathcal{G}_{\mathcal{H}}$	Gibbs class associated with $\mathcal{H}$ and $\mathcal{Q})\mathcal{H}$
$\mathscr{G}$	Gaussian penalty, average or complexity
G	grid of points; r.v. with asymptotic normal distribution
g	strategy
$\mathit{\Gamma}$	Gamma function
γ	scale for covering numbers
γ^-, γ^+	functions or constants for trimming
$\mathcal{H}$	hypothesis class
$\mathcal{H}'$	class of base hypotheses
$\mathscr{H}$	Hankel transform
H_0, H_a	null and alternative hypotheses
$H(Q)$	$H(Q) = \log_b \lvert Q_{\mathcal{W}} \rvert$
h	hypothesis — element of $\mathcal{H}$; function in functional Bennett's inequality
h'	base hypothesis — element of $\mathcal{H}'$
h_Q	element of $\mathcal{G}_{\mathcal{H}}(\mathcal{Q})$ associated with Q
I	indicator function
i	generic natural number
id	identity strategy

Symbol	*Usage*
$\mathcal{J}$	generic set
$\mathscr{J}$	Bessel function of the first kind
J	radius of balls for margin distribution bounds
j	generic natural number; element of set $\mathcal{J}$
$\mathcal{K}$	generic set
$\mathscr{K}$	kernel function
$\mathscr{K}_\nu$	ν-periodic extension of $\mathscr{K}$
K	generic constant
KL	Kullback-Leibler divergence
k	$\vert T\vert$ — size of test sample; element of a set $\mathcal{K}$
κ	multiplicative constant
$\mathcal{L}$	class of functionals
$\mathcal{L}_j$	class of decision tree leaves of depth at most j
$\mathscr{L}$	generic lower endpoint of an interval
L	loss function
L'	proxy loss function
L''	(L, L')-intermediary
$L^p(Q)$	Lebesgue space with p-norm w.r.t. the distribution Q
L_a	asymmetric loss function for classification
L_m	misclassification loss function
L_γ	margin loss function
L_ε	ε-insensitive loss function
LBT	lower endpoint of inverted binomial tail deviation
Leg	Legendre-Fenchel transform
Lik	likelihood function
LKL	lower endpoint of inverted KL deviation
Luck	luckiness function
l	$l = m + k$ — combined size of the training and test sample; element of range(L)
ℓ^p	space of sequences with the p-norm
ℓ^p_n	space of n-sequences with the p-norm
$\mathit{\Lambda}$	likelihood ratio

Symbol	*Usage*
λ	parameter for an m.g.f.
λ_i	eigenvalues of an operator
$\mathcal{M}$	packing number
M	$S \oplus P$, a dual sample; dimension of a product space
$M(E)$	median of the r.v. E
m	$\lvert S \rvert$ — size of training sample
$m_0(\epsilon, \delta)$	true sample complexity of an algorithm
$\mathrm{mode}(E)$	mode of the r.v. E
μ	mean of a r.v.
$\mathcal{N}$	covering number; shatter coefficient
$\bar{\mathcal{N}}$	external covering number
$\mathcal{N}_{p,Q}$	covering number w.r.t. $d_{p,Q}$
$\mathcal{N}_{p,\mathcal{S}}$	supremum of $\mathcal{N}_{p,Q}$ for $Q \in \mathcal{S}$
$\mathcal{N}_{p,n}$	supremum of $\mathcal{N}_{p,Q}$ for n-samples Q
$\mathcal{N}_{\mathcal{W}}$	shatter coefficient of $\mathcal{W}$
$\mathcal{N}_{\mathscr{C}}(R)$	analog of $\lvert Q_{\mathcal{W}} \rvert$ for the class of sets $\mathscr{C}$
$\mathcal{N}_{\mathscr{C}}(n)$	shatter coefficient of $\mathscr{C}$
$\mathscr{N}$	entropy number
N	dimension of a space
$N(\mu, \sigma^2)$	normal distribution
n	generic natural number; size of a generic sample Q
∇	gradient operator
ν	parameter for B-L and P-H ν-deviation; Gibbs distribution; width of periodic extension of $\mathscr{K}$
O	order notation
O_p	stochastic order notation
op	optimism of an estimator
ω	smallness function for a luckiness function; modified scaling factor for the .632+ estimator
$\mathcal{P}(\tau, \tau')$	class of couplings between τ and τ'
$\mathscr{P}_j$	j-th projection function for chaining
P	shadow sample of size u from D; generic sample from D

Symbol	*Usage*
$P_{\tau(Q)}$	the last u components of a permutation τ of an $(m+u)$-sample Q
p	$p = e_D(w)$ for some decision rule; constant for choice of norm
$\hat{p}$	$\hat{p} = r/k$, an estimate of p
$p^\star$	point estimate of p
p_i	elements of $P_{\tau(Q)}$
pdim	pseudodimension
$\mathit{\Phi}$	c.d.f. of $N(0,1)$ distribution; feature map of a kernel
ϕ	generic function; typical element of $\mathcal{V}$
φ	p.d.f. of $N(0,1)$ distribution
$\mathit{\Pi}$	function used for derivation of chaining bound
$\pi_{(\gamma^-,\gamma^+)}$	trimming function
$\mathit{\Psi}$	function in Bennett's inequality
ψ	measure of deviation; function in functional Bennet's inequality
ψ_i	eigenfunctions of an operator
$\mathcal{Q}$	class of distributions
$\mathcal{Q}_{\mathcal{E}}$	class of all distributions over $\mathcal{E}$
Q	$S \oplus P$, a dual sample; a generic n-sample; generic distribution, typical element of $\mathcal{Q}$
Q_w	the decision rule w restricted to the sample Q
$Q_{\mathcal{W}}$	the class of decision rules $\mathcal{W}$ restricted to the sample Q
q_i	element of the sample Q
$\mathcal{R}$	Rademacher penalty, average or complexity
R	generic set, with typical element r
$\hat{R}$	a measure of overfitting used with the .632+ estimator
r	number of test errors of a decision rule; action — typical element of $\mathcal{A}$; generic element of R
r_P	risk with respect to distribution P
ρ	margin
ϱ	generic value for a dimension
$\mathcal{S}$	class of distributions for the learning problem, typically $\mathcal{Q}_{\mathcal{Z}}$; class of stumps
$\mathcal{S}_i^+, \mathcal{S}_i^-$	classes of unidirectional splits on feature i

Symbol	*Usage*
$\mathscr{S}$	positive definite symmetric matrix
S	training sample (and the associated empirical distribution)
S_X	the set of input components of the training sample
S_Y	the set of output components of the training sample
$S^{\star b}$	bootstrap sample
S_n	symmetric group on $[1:n]$
$S^{\star}_{2m}$	swapping subgroup of S_{2m}
$S_{\setminus i}$	jackknife sample
$S_{\tau(Q)}$	the first m components of a permutation τ of an $(m+u)$-sample Q
s	threshold for indicator functions and thresholded classifiers; additive constants for linear expansions
s_i	elements of $S_{\tau(Q)}$
sgn	sign function
star	star hull
supp	support of a function or distribution
$\mathit{\Sigma}$	σ-algebra, typically of $\mathcal{E}$
σ	standard deviation of a r.v.; bandwidth of Gaussian kernel
ς	bound on standard deviation of a r.v.
$\mathcal{T}$	class of binary decision trees
$\mathcal{T}_j$	class of binary decision trees with at most j splits
$\mathcal{T}'_j$	class of binary decision trees with depth at most j
T	test sample (and the associated empirical distribution); operator
t	generic parameter of a distribution
τ	permutation — typical element of S_{m+u} or $S^{\star}_{2m}$; measure, typically over $\mathcal{E}$
$\mathit{\Theta}$	algorithm
ϑ	function of r.v.'s
$\mathscr{U}$	generic upper endpoint of an interval
U	r.v. with distribution $\mathrm{Unif}[0,1]$
UBT	upper endpoint of inverted binomial tail deviation
UKL	upper endpoint of inverted KL deviation
Unif A	uniform distribution over A

Symbol	*Usage*
u	shadow sample size; realized value of r.v. U
$u(m)$	shadow sample size for ω-smallness of a luckiness function
Υ	function used for derivation of chaining bound
$\mathcal{V}$	generic class of functions, with typical element ϕ
$\mathscr{V}$	generic vector with components v_i
V	a generic random variable, often a function of a number of E_i
VC	VC dimension
v	generic function, set, vector, or number
v_ϕ	fixed point of the sub-root function ϕ
$\mathcal{W}$	decision class, with typical element w
$\mathcal{W}^\star$	a cover of $\mathcal{W}$, with typical element $w^\star$
$\mathcal{W}_t$	thresholded decision class
$\mathcal{W}(Q, w)$	subclass of "luckier" decision rules than w on Q
W	a generic random variable, often a function of a number of V_i
w	decision rule — typical element of $\mathcal{W}$
$w^\star$	decision rule — typical element of $\mathcal{W}^\star$
w_S	decision rule dependent on the sample S
w_s	decision rule thresholded at s
$\mathcal{X}$	input space, with typical element x
X	r.v. representing an input
x	input — typical element of $\mathcal{X}$
x_i	training input — see z_i
$x_i^\star$	test input — see $z_i^\star$
ξ	embedding function for margin distribution bounds
$\mathcal{Y}$	output space, with typical element y
$\mathscr{Y}$	$\mathscr{Y} = \sup_{w \in \mathcal{W}'}[r_D(w) - r_S(w)]$
Y	r.v. representing an output
y	output — typical element of $\mathcal{Y}$
y_i	training output — see z_i
$y_i^\star$	test output — see $z_i^\star$
$\mathcal{Z}$	$\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, with typical element z
Z	r.v. in $\mathcal{Z}$, typically with distribution D;

Symbol	*Usage*
	r.v. with asymptotic $N(0,1)$ distribution
z	$z = (x, y)$, an input-output pair — typical element of $\mathcal{Z}$
z_i	$z_i = (x_i, y_i)$, element of the training sample S
$z_i^\star$	$z_i^\star = (x_i^\star, y_i^\star)$, element of the test sample T
z_δ	critical value of the normal distribution
ζ	Rademacher variable

Appendix B

List of Abbreviations

This appendix lists the abbreviations used in this thesis, their page of first occurrence, and what they stand for.

Abbreviation	*Page*	*In full*
AIC	127	Akaike information criterion
AV	87	Angluin-Valiant
BIC	127	Bayes information criterion
B-L	41	Bartlett-Lugosi
BS	61	Blyth-Still
BSC	63	Blyth-Still-Casella
BSW	61	Blyth-Still-Wald
BU	31	Best unbiased
CC	59	Continuity correction
c.d.f.	43	cumulative distribution function
CV	19	Cross-validation
DD-SRM	274	Data-dependent structural risk minimization
ERM	270	Empirical risk minimization
GC	237	Glivenko-Cantelli
i.i.d.	14	independent, identically distributed
KL	35	Kullback-Leibler
HPD	49	Highest posterior density

Abbreviation	*Page*	*In full*
LOO-CV	134	Leave-one-out cross-validation
LR	52	Likelihood ratio
MAP	16	Maximum a posteriori
MDL	128	Minimum description length
ME	34	Maximum entropy
m.g.f.	80	moment generating function
ML	31	Maximum likelihood
MM	31	Method of moments
MRE	32	Minimum risk equivariant
MSE	31	Mean squared error
PAC	13	Probably approximately correct
PPE	39	Posterior probability estimator
P-H	42	Pollard-Haussler
r.v.	15	random variable
RSE	127	Residual squared error
SRM	268	Structural risk minimization
SV	22	Support vector
UMVU	31	Uniform minimum-variance unbiased
VC	231	Vapnik-Chervonenkis
VCSS	232	Vapnik-Chervonenkis-Sauer-Shelah

Appendix C

Code of R functions

This appendix includes various functions employed in calculating most of the bounds and capacity measures presented in Chapter 9. The functions are presented in alphabetical order of their names.

Disclaimer: these functions were written by me, for me! They would need thorough sanitation before being made generally available. If you wish to use them, it is strongly recommended that you verify them against the theory first, since they were often implemented for very specific circumstances.

Function `alpha_bph` evaluates the value of α_{BPH} for a given choice of β.

```
function(second_sample_size,beta,nu) {

        sqrt(-8*log(1-beta)/(9*second_sample_size*nu))

}
```

Function `alpha_c` evaluates the value of α_C for a given choice of β.

```
function(second_sample_size,beta) {

        sqrt(1/(4*second_sample_size*(1-beta)))

}
```

Function `alpha_cc` evaluates the value of α_{CC} for a given choice of β.

```
function(second_sample_size,beta) {

        sqrt(beta/(4*second_sample_size*(1-beta)))

}
```

Function alpha_h evaluates the value of α_H for a given choice of β.

```
function(second_sample_size,beta) {

        sqrt(-log(1-beta)/(2*second_sample_size))

}
```

Function av_upper calculates an upper bound based on the AV bounds.

```
function(test_error, test_size, lower_conf, upper_conf) {

   av_upper_bound<-test_error-(log(upper_conf)/test_size)
       *(1+sqrt(1-(2*test_error*test_size/log(upper_conf))))

   av_upper_bound

}
```

Function bernstein_binom_var_upper yields an upper bound based on Bernstein's inequality, using the binomial variance estimate.

```
function(test_error, test_size, lower_conf, upper_conf) {

   optim_function<-function(r)
   {
        val<- r - test_error - (-log(upper_conf)+
          sqrt(-18*test_size*r*(1-r)*log(upper_conf)))/(3*test_size)
    }

   if (optim_function(1)>0) {
       res<-uniroot(optim_function,c(0,1))
    } else {
        res<-list(root=1)
    }
   res$root

}
```

Function bernstein_phdev_risk_var_upper calculates an interval based on a bound on the P-H ν-deviation found using Bernstein's inequality.

```
function(test_error, test_size,
lower_conf, upper_conf,nu) {

   interim_val<-(-3*(1+sqrt((18*test_size*nu)
     /(-log(upper_conf)))))/(1-(18*test_size*nu)/(-log(upper_conf)))

   bound<- ((1+interim_val)*test_error+interim_val*nu)/(1-interim_val)
   bound

}
```

Function best_alpha evaluates the functions alpha_h, alpha_c and alpha_cc, and returns the minimum of the three.

```
function(second_sample_size,beta) {
pmin(alpha_h(second_sample_size,beta),
        alpha_c(second_sample_size,beta),
        alpha_cc(second_sample_size,beta))
}
```

Function bl_dev_error_bound_epsilon returns a bound on the B-L ν-deviation using a double sample result. The resulting bound is typically inverted using the function invert_bl_deviation. It employs the function best_alpha.

```
function(sample_size,dbl_sample_log_shat_coef,conf,nu) {

           find_best_beta_for_nu<-function(nu_val) {
           opt<-function(beta) { best_alpha(sample_size,beta)-nu_val}
                   uniroot(opt,c(0,1))$root
           }
           best_beta<-sapply(nu,find_best_beta_for_nu)
   res<-sqrt((4*(dbl_sample_log_shat_coef-log(best_beta*conf)))/
           sample_size)
}
```

Function corrected_dt_log_cov_number returns a bound on the natural logarithm of the covering number of a class of decision trees, based on the number of splits and the number of (continuous) features. It employs the function corrected_log_cov_num_from_vc_dim.

```
function(sample_size,num_splits,num_features,scale) {

        num_splits*(log(2*num_features)+
            corrected_log_cov_num_from_vc_dim(sample_size,
            1,scale/num_splits,error=TRUE))

}
```

Function corrected_log_cov_num_from_vc_dim returns a bound on the natural logarithm of the covering numbers of a class based on the VC dimension of the class. It employs the function log_shat_coef

```
function(sample_size,vc_dim,scale,error=FALSE) {

    estimate1<-log_shat_coef(sample_size,vc_dim)

    estimate2<-log(2)+vc_dim*log((2*exp(1)/scale)*
                    log(2*exp(1)/scale))

    if (error) {
            scale<-floor(scale*sample_size)/sample_size
            estimate3<-log(vc_dim+1)+1+vc_dim*
                        log(2*exp(1)/scale)
    } else {
            estimate3<-estimate2
    }

    combined_estim<-pmin(estimate1,estimate2,estimate3)
    combined_estim[scale > 1] <- 0
    combined_estim
}
```

Function credible_beta returns a Bayesian credible interval for a binomial proportion based on a Beta prior.

```
function(errors,size,beta_shape_1,beta_shape_2,lower_conf,upper_conf)
{
    credible_beta_lower<-qbeta(lower_conf,errors+beta_shape_1,
        (size-errors)+beta_shape_2)
    credible_beta_upper<-qbeta(1-upper_conf,errors+beta_shape_1,
        (size-errors)+beta_shape_2)
    c(credible_beta_lower,credible_beta_upper)
}
```

Function `dbl_sample_pac_bayes_error_bound` returns an error bound based on a PAC-Bayesian argument over a double sample.

```
function(train_error,sample_size,log_cov_num,conf) {

    train_error+1/sample_size+sqrt(2*(log(2)+
          log_cov_num-log(conf))/sample_size)

}
```

Function `dt80_symm_abc_eval` calculates the sample risk on a bootstrap sample for use with the `abc.ci` function of the `boot` package.

```
function(orig_data,bootstrap_weights){
    bootstrap_weights %*% (orig_data$y!=predict(dt80_symm,
         orig_data,type="class"))
}
```

Function `dt80_symm_boot_eval2` calculates the sample mean and sample variance of loss on a bootstrap sample, for use with the `boot` function in the package `boot`.

```
function(orig_data,bootstrap_sample){
boot_sample<-orig_data[bootstrap_sample,]
boot_error<-length(boot_sample$y[boot_sample$y!=
    predict(dt80_symm,boot_sample,type="class")])/920
boot_var<-boot_error*(1-boot_error)/length(boot_sample)
c(boot_error,boot_var)
}
```

Function `dual_sample_pac_bayes_risk_bound` returns a risk bound based on a PAC-Bayesian argument over a dual sample.

```
function(train_risk,first_size,second_size,
    log_cov_num,conf,beta,scale) {

    m<-first_size
    u<-second_size

    train_risk+best_alpha(u,beta)+(2*m+u)*(m+u)*scale/(m*u)+
        (m+u)*sqrt((log_cov_num-log(conf)-log(beta))/(2*m))/u

}
```

Function further_corrected_dt_chaining_bound approximates a chaining bound for decision trees based on the number of splits in the tree and the number of (continuous) features in the data. It employs the functions corrected_dt_log_cov_number and best_alpha.

```
function(train_risk,sample_size,num_splits,num_features,
    conf,beta,num_terms=100) {

        modified_delta<-conf*beta/2
        modified_epsilon<- function(delta_val) {
           3*sqrt(2/sample_size)*sum(2^(-(1:num_terms))*sqrt(
                corrected_dt_log_cov_number(sample_size,num_splits,
                num_features,(2^(-(1:num_terms))/2)^2)-
                log(delta_val)+log(1:num_terms)+log(1+(1:num_terms))))
        }
        epsilon_vec<-sapply(modified_delta,modified_epsilon)
        bound<-train_risk+2*epsilon_vec+best_alpha(sample_size,beta)
        bound
}
```

Function hoeffding_tail yields a confidence interval based on Hoeffding's tail inequality.

```
function(test_error, test_size,
lower_conf, upper_conf) {

   lower_bound_val<-sqrt(-log(lower_conf)/(2*test_size))
   upper_bound_val<-sqrt(-log(upper_conf)/(2*test_size))
   hoeffding_tail_lower<-invert_simple_deviation("lower",
        lower_bound_val,test_error)
   hoeffding_tail_upper<-invert_simple_deviation("upper",
        upper_bound_val,test_error)
   c(hoeffding_tail_lower,hoeffding_tail_upper)
}
```

Function hoeffding_re_upper yields a confidence interval based on Hoeffding's r.e. inequality. This employs the function re_invert_kl_deviation.

```
function(test_error, test_size,
lower_conf, upper_conf, tol=0.0001) {

        upper_optim_function<-function(r) {
            re_invert_kl_deviation("upper",
                -log(upper_conf)/test_size,r,test_size,tol)-test_error
        }

        if (upper_optim_function(1-tol) > 0) {
            if (upper_optim_function(tol) < 0) {
                res<-uniroot(upper_optim_function,c(tol,1-tol))
            } else {
                res<-list(root=0)
            }
        } else {
            res<-list(root=1)
        }
        res$root

}
```

Function integer_prior returns the "prior" probability over a sequence of natural numbers.

```
function(value,min_val=1) {
1/((value-min_val+1)*(value-min_val+2))
}
```

Function invert_binomial_tail_deviation computes max-P or mid-P interval endpoints based on a bound on the binomial tail deviation.

```
function(type,bound,estimate,sample_size,mid=FALSE){

    if (type=="upper") {
        optim_function<-function(x)
        pbinom(estimate*sample_size,sample_size,x)-
```

```
            0.5*mid*dbinom(estimate*sample_size,sample_size,x)-bound
        if (sign(optim_function(0)) !=
            sign(optim_function(1))) {
            res<-uniroot(optim_function,c(0,1))
        } else {
            res<-list(root=1)
        }
    } else if (type=="lower") {
        optim_function<-function(x)
        pbinom(estimate*sample_size-1,sample_size,x)+
            0.5*mid*dbinom(estimate*sample_size,sample_size,x)-(1-bound)
        if (sign(optim_function(0)) !=
            sign(optim_function(1))) {
            res<-uniroot(optim_function,c(0,1))
        } else {
            res<-list(root=0)
        }
    }
    res$root

}
```

Function invert_bl_deviation inverts a bound on the B-L ν-deviation. It employs the function invert_relative_deviation.

```
function(type,bound,estimate,nu){

        invert_relative_deviation("upper",bound+nu,estimate)

}
```

Function invert_rao_deviation computes interval endpoints based on a bound on the Rao deviation.

```
function(type,bound,estimate,sample_size) {
   if (type=="upper") {
        denominator<- sample_size+bound^2
        sqrt_factor<-sqrt(sample_size*estimate*(1-estimate)+(bound^2)/4)
        numerator<-sample_size*estimate+(bound^2)/2+bound*sqrt_factor
        res<-numerator/denominator
    } else if (type=="lower") {
        res<-invert_rao_deviation("upper",-bound,estimate,sample_size)
    }
    res
}
```

Function invert_relative_deviation inverts a bound on the relative deviation of risk.

```
function(type,bound,estimate){

    if (type=="upper") {
        res<-(2*estimate+bound^2+bound*sqrt(bound^2+4*estimate))/2
    } else if (type=="lower") {
        res<-invert_relative_deviation("upper",-bound,estimate)
    }
    res
}
```

Function invert.transform inverts the data transformations performed by the function transform.

```
function(type,value,inv_custom) {

    if (type=="logit") {
        new_val<- 1-1/(1+exp(value))
    } else if (type=="probit") {
        new_val<-pnorm(value)
    } else if (type=="cloglog") {
        new_val<- exp(-exp(-value))
    } else if (type=="arcsine") {
        new_val<- (sin(value))^2
    } else if (type=="custom") {
        new_val<- inv_custom(value)
    } else {
        if (type != "identity") {
            print("Unknown transform specified, identity
            returned")
        }
        new_val<-value
    }
    new_val
}
```

Function kl_bernoulli calculates the KL deviation between two Bernoulli distributions.

```
function(p1,p2) {

    res<- p1*log(p1/p2)+(1-p1)*log((1-p1)/(1-p2))
}
```

Function log_shat_coef returns a bound on the natural logarithm of the shatter coefficient of a class, based on the VC dimension of the class.

```
function(sample_size,vc_dim) {

        estimate1<-0
        for (i in 0:vc_dim) {
                estimate1<-estimate1+choose(sample_size,i)
        }

        estimate1<-log(estimate1)
        estimate2<-estimate1
        estimate2[estimate2 ==  Inf]<-log(2)+
            vc_dim*log(sample_size[estimate2 ==
            Inf])-lfactorial(vc_dim)

        if (any(estimate2 != estimate1)) {
          print("Using non-combinatorial approximation for some sample sizes")
        }

        estimate2
}
```

Function ph_dev_risk_bound calculates a bound based on a double sample bound on P-H ν-deviation. It employs the function alpha_bph.

```
function(train_risk,sample_size,dbl_sample_log_cov_num,conf,
    scale,beta,nu) {

        epsilon<-scale+alpha_bph(sample_size,beta,nu)+
            sqrt((dbl_sample_log_cov_num+log(2)-
            log(conf*beta))/(2*nu*sample_size))
        ((1+epsilon)*train_risk+epsilon*nu)/(1-epsilon)

}
```

Function pratt_approx implements the core expression in Pratt's max-P interval approximation.

```
function(errors,total,conf) {

    crit_val<-qnorm(1-conf)
    1/(1+(((errors+1)/(total-errors))^2)*(((81*(errors+1)*
        (total-errors)-9*total-8-3*crit_val*sqrt(9*(errors+1)*
        (total-errors)*(9*total+5-crit_val^2)+total+1))/
        (81*(errors+1)^2-9*(errors+1)*(2+crit_val^2)+1))^3))

}
```

Function `pratt_max_upper` return's Pratt's approximation to the upper bound of the max-P interval. It employs the function `pratt_approx`.

```
function(errors,total,conf) {
    pratt_approx(errors,total, conf)
}
```

Function `pratt_mid_upper` returns Pratt's approximation to the mid-P interval. It employs `pratt_max_upper`.

```
function(errors,total,conf) {
    mean(c(pratt_max_upper(errors,total, conf),
        pratt_max_upper(errors-1,total, conf)))
}
```

Function `predict_boost_with_dist` calculates the sample risk of the Gibbs classifier where the distribution is specified by the weights of a boosted model.

```
function(boost_object,to_predict,loss_func) {


    loss_on_indiv_predict<-function(num) {
        prediction<-predict(boost_object$model$trees[num][[1]],
            to_predict,type="class")
        mean(loss_func(as.numeric(prediction)-1,to_predict$y))
    }

        sum_of_weights<-sum(boost_object$model$alpha)
        sum(boost80s$model$alpha*sapply(1:boost_object$iter,
            loss_on_indiv_predict)/sum_of_weights)
}
```

Function rademacher_avg returns a bound on the Rademacher average of a sample based on the natural logarithm of the empirical shattering coefficient.

```
function(emp_log_shat_coef,sample_size) {

        2*sqrt(emp_log_shat_coef/sample_size)+exp(-emp_log_shat_coef)

}
```

Function rademacher_pen returns a probabilistic bound on the Rademacher penalty for a sample based on the natural logarithm of the supremum of the empirical shattering coefficient.

```
function(sup_log_shat_coef,sample_size,conf) {

        sqrt(2*(sup_log_shat_coef-log(conf))/sample_size)

}
```

Function random_subsample_bound returns a bound calculated based on a regular deviation bound obtained using the random subsample lemma. It employs the function best_alpha.

```
function(train_risk,sample_size,sample_log_cov_num,conf, scale,beta) {

        train_risk+best_alpha(sample_size,beta)+
            2*(scale+sqrt(2*(sample_log_cov_num+log(2)-
            log(conf*beta))/sample_size))

}
```

Function refined_dt_log_shat_coef returns a bound on the natural logarithm of the shattering coefficient of a decision tree implemented on the spam data, using additional information on representation of the first 54 features.

```
function(sample_size,num_splits) {
        split_log_shat_coef<-log(108*min(sample_size+1,10001)+
            6*(sample_size+1))
        num_splits*split_log_shat_coef

}
```

Function `refined_symm_ineq_rademacher_bound` returns a risk bound obtained by combining the Rademacher symmetrization inequality with the functional Bernstein's inequality.

```
function(train_risk,subclass_rademacher_complexity,
    subclass_variance_bound,sample_size,conf) {
        train_risk+2*subclass_rademacher_complexity+
            (sqrt(-18*log(conf)*(sample_size*subclass_variance_bound+
            4*subclass_rademacher_complexity))-log(conf))/(3*sample_size)
}
```

Function `reg_dev_error_bound_scale` returns a bound based on inverting a double sample bound on regular deviation of error.

```
function(train_error,sample_size,dbl_sample_log_cov_num,conf, scale) {
train_error+2*scale+(1+sqrt(((sample_size+1)*(dbl_sample_log_cov_num+
    log(2)-log(conf)))+1))/sample_size
}
```

Function `reg_dev_risk_bound_dual_sample_scale` returns a bound based on a dual sample bound on regular deviation. It employs the function `best_alpha`.

```
function(train_error,sample_size,second_sample_size,
    dual_sample_log_cov_num,conf, scale,beta) {

        m<-sample_size
        u<-second_sample_size

        train_error+(2*m+u)*(m+u)*scale/(m*u)+best_alpha(u,beta)+
            ((m+u)/u)*sqrt((dual_sample_log_cov_num-log(conf*beta))/(2*m))

}
```

Function `reg_dev_risk_bound_scale` evaluates the double sample regular deviation bound on risk. It employs the function `best_alpha`.

```
function(train_error,sample_size,dbl_sample_log_cov_num,conf,
    scale,beta) {

        train_error+2*scale+best_alpha(sample_size,beta)+
            sqrt(((sample_size+1)*(dbl_sample_log_cov_num+
            log(2*sample_size)-log(conf*beta))+1)/sample_size

}
```

Function `re_invert_kl_deviation` inverts a bound on KL deviation. It employs the function `kl_bernoulli`.

```
function(type,bound,estimate,sample_size,tol){

    if (type=="upper") {
        optim_function<-function(x) {
        res<-kl_bernoulli(estimate*x,estimate)-bound
        }
            if (optim_function(tol) < 0) {
                res<-list(root=0)
            } else {
                res<-uniroot(optim_function,c(tol,1))
            }
    } else if (type=="lower") {
        optim_function<-function(x)
        kl_bernoulli(estimate*x,estimate)-bound
        if (sign(optim_function(0)) !=
            sign(optim_function(1))) {
            res<-uniroot(optim_function,c(0,1))
        } else {
            res<-list(root=0)
        }
    }
    estimate*res$root

}
```

Function `rel_dev_error_bound_epsilon` calculates an upper bound on the relative deviation of error based on a double sample result. The deviation is typically inverted with the function `invert_relative_deviation`.

```
function(sample_size,dbl_sample_log_shat_coef,conf) {
    res<-sqrt((4*(log(4)+dbl_sample_log_shat_coef-log(conf)))/
            sample_size)
    if (any(sample_size*res^2 <= 1)) {
            print("Some values too small to apply bound")
    }
    res
}
```

Function `score` computes the score interval for a proportion. This employs the function `invert_rao_deviation`

```
function(test_error, test_size,
lower_conf,upper_conf,cc=TRUE,bs=FALSE) {

    score_lower_crit<-qnorm(1-lower_conf)
    score_upper_crit<-qnorm(1-upper_conf)
    score_interval<-c(invert_rao_deviation("lower",score_lower_crit,
        test_error-cc/(2*test_size),test_size),invert_rao_deviation(
        "upper",score_upper_crit,test_error+cc/(2*test_size),test_size))

}
```

Function `sq_deriv.transform` calculates a squared derivative of a value transformed using `transform`, for creating a Wald interval on the transformed values.

```
function(type,value,deriv_custom) {

    if (type=="logit") {
        new_val<- 1/(value*(1-value))
    } else if (type=="probit") {
        new_val<- 1/dnorm(qnorm(value))
    } else if (type=="cloglog") {
        new_val<- -1/(value*log(value))
    } else if (type=="arcsine") {
        new_val<- 1/(2*sqrt(value*(1-value)))
    } else if (type=="custom") {
        new_val<- deriv_custom(value)
    } else {
        if (type != "identity") {
            print("Unknown transform specified, identity
            returned")
        }
        new_val<-1
    }
    sq_val<-new_val^2
}
```

Function symm_ineq_rademacher_bound returns a risk bound obtained by combining the Rademacher symmetrization inequality with the bounded difference inequality.

```
function(train_risk,rademacher_complexity,sample_size,conf) {
train_risk+2*rademacher_complexity+sqrt(-log(conf)/(2*sample_size))
}
```

Function t_lip_error calculates the sample risk on a boosted model employing the optimal t-Lipschitz proxy loss on the misclassification loss.

```
function(t,sum_of_weights,predictions,actual) {

    final<-predictions*actual/sum_of_weights
    sample_size<-length(final)
        transform<-function(val,t_val) {
                pmin(1,pmax(0,1-t_val*val))
        }
    err_rate<-function(t_val) { sum(transform(final,t_val)) }
    res<-sapply(t,err_rate)
    res/sample_size
}
```

Function transform performs some standard data transformations for constructing confidence intervals.

```
function(type,value,custom) {

    if (type=="logit") {
        new_val<-log(value/(1-value))
    } else if (type=="probit") {
        new_val<-qnorm(value)
    } else if (type=="cloglog") {
        new_val<- -log(-log(value))
    } else if (type=="arcsine") {
        new_val<- asin(sqrt(value))
    } else if (type=="custom") {
        new_val<- custom(value)
    } else {
        if (type != "identity") {
            print("Unknown transform specified, identity applied")
        }
        new_val<-value
    }
    new_val
}
```

Function wald generates intervals based on the Wald test.

```
function(test_error, test_size,
lower_conf,upper_conf,cc=TRUE, bs=FALSE) {

    var_hat<- test_error*(1-test_error)/test_size
    wald_lower_crit<-qnorm(1-lower_conf)
    wald_upper_crit<-qnorm(1-upper_conf)
    if (bs) {
        bs_adjust<-(sqrt(test_size/(test_size-wald_upper_crit^2-
            2*wald_upper_crit/sqrt(test_size)-1/test_size)
))
    } else {
        bs_adjust<-1
    }
    wald_lower<-test_error-(bs_adjust*wald_lower_crit*sqrt(var_hat)+
        cc/(2*test_size))
    wald_upper<-test_error+(bs_adjust*wald_upper_crit*sqrt(var_hat)+
        cc/(2*test_size))
    c(invert_wald_deviation("lower",wald_lower_crit,test_error,var_hat),
invert_wald_deviation("upper",wald_upper_crit,test_error,var_hat))

    c(wald_lower, wald_upper)
}
```

Function wald.transform obtains a Wald interval on transformed data. It makes use of the functions wald, transform, sq_deriv.transform and invert.transform.

```
function(type, test_error, test_size,
lower_conf,upper_conf,cc=FALSE, bs=FALSE,
custom,inv_custom,deriv_custom) {
    var_hat<-test_error*(1-test_error)/test_size
    wald_lower_crit<-qnorm(1-lower_conf)
    wald_upper_crit<-qnorm(1-upper_conf)
    transformed_test_error<-transform(type,test_error,custom)
    variance_transformed_estimate<-var_hat*
        sq_deriv.transform(type,test_error,deriv_custom)

    invert.transform(type,c(invert_wald_deviation("lower",
        wald_lower_crit,transformed_test_error,
        variance_transformed_estimate),invert_wald_deviation(
        "upper",wald_upper_crit,transformed_test_error,
        variance_transformed_estimate)),inv_custom)
}
```

Bibliography

Note: We have indicated references whose contents we have not seen first-hand with a dagger (†). In these cases, the contents of the source has been inferred by another source, also cited.

Agresti, A. and Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, **52** (2), 119–126.

Agresti, A. and Min, Y. (2005). Frequentist performance of Bayesian confidence intervals for comparing proportions in 2 x 2 contingency tables. *Biometrics*, **61** (2), 515–523.

Akaike, H. (1974). †A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19** (6), 716–723.

Alexander, K. S. (1984). Probability inequalities for empirical processes and a law of the iterated logarithm. *Annals of Probability*, **12** (4), 1041–1067. Correction published in Annals of Probability, Vol. 15, No. 1. (Jan., 1987), pp. 428–430.

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, **16** (1), 125–127.

Alon, N., Ben-David, S., Cesa-Bianchi, N. and Haussler, D. (1993). Scale-sensitive dimensions, uniform convergence, and learnability. *Proceedings of the Conference on Foundations of Computer Science (FOCS)*. Also in Journal of the ACM, volume 44, number 4, 1997, 615–631.

ANDONOVA JAEGER, S. (2005). Generalization bounds and complexities based on sparsity and clustering for convex combinations of functions from random classes. *Journal of Machine Learning Research*, **6**, 307–340.

ANGLUIN, D. (1992). Computational learning theory: Survey and selected bibliography. *Proceedings of the 24th Annual ACM Symposium on Theory of Computing*, pp. 351–369. ACM Press.

ANGLUIN, D. and VALIANT, L. (1979). †Fast probabilistic algorithms for Hamiltonian circuits and matchings. *Journal of Computer and System Sciences*, **19**, 155–193.

ANTHONY, M. (1994). Probabilistic analysis of learning in artificial neural networks: The PAC model and its variants. Technical Report NC-TR-94-003, NeuroCOLT.

ANTHONY, M. (1997). Probabilistic analysis of learning in artificial neural networks: The PAC model and its variants. *Neural Computing Surveys*, **1**, 1–47.

ANTHONY, M. and BARTLETT, P. L. (1994). Function learning from interpolation. Technical Report NC-TR-94-013, NeuroCOLT.

ANTHONY, M. and BARTLETT, P. L. (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge University Press.

ANTHONY, M. and SHAWE-TAYLOR, J. (1993). A result of Vapnik with applications. *Discrete Applied Mathematics*, **47**, 207–217.

ASUNCION, A. and NEWMAN, D. (2007). UCI machine learning repository. `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

AUDIBERT, J. and BOUSQUET, O. (2007). Combining PAC-Bayesian and generic chaining bounds. *Journal for Machine Learning Research*.

AZUMA, K. (1967). †Weighted sums of certain dependent random variables. *Tôhoku Mathematics Journal*, **19** (2), 357–367.

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

Bartlett, P. L. (2003). Prediction algorithms: Complexity, concentration and convexity. *Proceedings of the 13th IFAC Symposium on System Identification*, pp. 1507–1517.

Bartlett, P. L., Boucheron, S. and Lugosi, G. (2000). Model selection and error estimation. *Proceedings of the 13th Annual Conference on Computational Learning Theory*, pp. 286–297. Morgan Kaufmann, San Francisco.

Bartlett, P. L., Boucheron, S. and Lugosi, G. (2002). Model selection and error estimation. *Machine Learning*, **48**, 85–113.

Bartlett, P. L., Bousquet, O. and Mendelson, S. (2004). Local Rademacher complexities. To appear in the Annals of Statistics.

Bartlett, P. L., Jordan, M. I. and McAuliffe, J. D. (2003a). Convexity, classification and risk bounds. Tech. Rep. 638, Department of Statistics, University of California, Berkeley, Berkeley, California. Accepted subject to revisions by Journal of the American Statistical Association.

Bartlett, P. L., Jordan, M. I. and McAuliffe, J. D. (2003b). Large margin classifiers: Convex loss, low noise and convergence rates. *Advances in Neural Information Processing Systems 16 (NIPS-2003).*

Bartlett, P. L., Long, P. M. and Williamson, R. C. (1996). Fat shattering and the learnability of real-valued functions. *Journal of Computer and System Sciences*, **52** (3), 434–452.

Bartlett, P. L. and Lugosi, G. (1999). An inequality for uniform deviations of sample averages from their means. *Statistics and Probability Letters*, **44**, 55–62.

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

Bennett, G. (1962). Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, **57**, 33–45.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analyis.* Springer, New York.

Bernstein, S. N. (1924). †Sur une modification de línéqualité de Tchebichef. *Annals of the Science Institute of Sav. Ukraine, Sect. Math. I.* Russian with French summary.

Bernstein, S. N. (1927). †*Theory of Probability.* Unknown publisher, Moscow.

Bishop, Y. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* The MIT Press, Cambridge, MA.

Blanchard, G. and Fleuret, F. (2007). Occam's hammer. *Computational Learning Theory 2007.*

Blum, A. and Langford, J. (2003). PAC-MDL bounds. *Computational Learning Theory 2003.*

Blumer, A., Ehrenfeucht, A., Haussler, D. and Warmuth, M. K. (1986). Classifying learnable geometric concepts with the Vapnik-Chervonenkis dimension. *Proceedings of the 18th Annual ACM Symposium on Theory of Computing*, pp. 273–282. ACM Press.

Blumer, A., Ehrenfeucht, A., Haussler, D. and Warmuth, M. K. (1989). Learnability and the Vapnik-Chervonenkis dimension. *Journal of the Association for Computing Machinery*, **36** (4), 929–965.

Blyth, C. R. and Still, H. A. (1983). Binomial confidence intervals. *Journal of the American Statistical Association*, **78** (381), 108–116.

Boser, B. E., Guyon, I. M. and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. D. Haussler (editor) *Proceedings* *of the Fifth Annual Workshop on Computational Learning Theory*, pp. 144–152. ACM Press, Pittsburg, PA.

Botha, L. M. (1992). *Aanpassende Beraming van Multinomiaalwaarskynlikhede met Toepassing in die Ontleding van Gebeurlikheidstabelle.* Master's thesis, Potchefstroom Universiteit vir Christen Hoër Onderwys. English title: Adaptive Estimation of Multinomial Probabilities with Application in the Analysis of Contingency Tables.

Boucheron, S., Lugosi, G. and Massart, P. (1999). A sharp concentration inequality with applications. Technical Report NC-TR-99-057, NeuroCOLT.

Bousquet, O. (2002a). A Bennett concentration inequality and its application to suprema of empirical processes. *Les Comptes Rendus de l'Académie des Sciences*, **334**, 495–500.

Bousquet, O. (2002b). *Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms.* Ph.D. thesis, Ecole Polytechnique.

Bousquet, O., Boucheron, S. and Lugosi, G. (2005). Theory of classification: A survey of some recent advances. *ESAIM: Probability and Statistics*, **9**, 323–375.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, **2**, 499–526.

Bousquet, O. and Herrmann, D. J. L. (2003). On the complexity of learning the kernel matrix. *Advances in Neural Information Processing Systems 15 (NIPS2003)*, pp. 415–422. The MIT Press, Cambridge, MA.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, **26** (2), 123–140.

Brent, R. P. (1973). †*Algorithms for Minimization without Derivatives.* Prentice-Hall, Englewood Cliffs, NJ. ISBN 0-13-022335-2.

Brown, L. D., Cai, T. T. and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, **16** (2), 101–133.

Brown, L. D., Cai, T. T. and DasGupta, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions. *The Annals of Statistics*, **30** (1), 160–201.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Knowledge Discovery and Data Mining*, **2** (2), 121–167.

Cannon, A., Ettinger, J. M., Hush, D. R. and Scovel, C. (2002). Machine learning with data-dependent hypothesis classes. *Journal of Machine Learning Research*, **2**, 335–358. Also Los Alamos National Laboratory Technical Report LA-UR-01-2583.

Canty, A. (2005). *boot: Bootstrap R (S-Plus) Functions.* S original by Angelo Canty <cantya@mcmaster.ca>. R port by Brian Ripley <ripley@stats.ox.ac.uk>. R package version 1.2-22.

Carpenter, J. and Bithell, J. (2000). Bootstrap confidence intervals: When, which, what? a practical guide for medical statisticians. *Statistics in Medicine*, **19**, 1141–1164.

Casella, G. (1986). Refining binomial confidence intervals. *The Canadian Journal of Statistics*, **14** (2), 113–129.

Catoni, O. (2003). Localized empirical complexity bounds and randomized estimators. Preprint.

Catoni, O. (2004a). Improved Vapnik Cervonenkis bounds. Preprint.

Catoni, O. (2004b). A PAC-Bayesian approach to adaptive classification. Technical Report 840, Laboratoire de Probabilités en Modèles Aléatoires, Université Paris 6 (Site Chevaleret). Revised version of October 9, 2004. Submitted to Annals of Statistics.

CESA-BIANCHI, N. and HAUSSLER, D. (1998). A graph-theoretic generalization of the Sauer-Shelah lemma. *Discrete Applied Mathematics*, **86**, 27–35.

CHERNOFF, H. (1952). †A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, **23**, 493–507.

CLOPPER, C. and PEARSON, E. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, **26**, 404–413.

CORTES, C. and VAPNIK, V. (1995). Support-Vector networks. *Machine Learning*, **20** (3), 273–297.

CRISTIANINI, N. and SHAWE-TAYLOR, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods.* Cambridge University Press, Cambridge, UK. ISBN 0-521-78019-5.

CROW, E. L. (1956). Confidence intervals for a proportion. *Biometrika*, **43**, 423–435.

CULP, M., JOHNSON, K., and MICHAILIDIS, G. (2006). *ada: Performs boosting algorithms for a binary response.* R package version 3.1-38.

DAVIES, E. and SIMON, B. (1984). Ultracontractivity and the heat kernel for Schrödinger operators and Dirichlet Laplacians. *Journal of Functional Analysis*, **59**, 335–395.

DAVISON, A. C., HINKLEY, D. V. and SCHECHTMAN, E. (1986). Efficient bootstrap simulation. *Biometrika*, **73** (3), 555–566.

DE FINETTI, B. (1931). †Funzione caratteristica di un fenomeno aleatorio. *Atti della R. Academia Nazionale dei Lincei, Serie 6. Memorie, Classe di Scienze Fisiche, Mathematice e Naturale*, **4**, 251–299.

DEVROYE, L. (1982). Bounds for the uniform deviation of empirical measures. *Journal of Multivariate Analysis*, **12**, 72–79.

Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Applications of Mathematics. Springer, New York. ISBN 0-387-94618-7.

Devroye, L. and Lugosi, G. (2001). *Combinatorial Methods in Density Estimation*. Springer Series in Statistics. Springer, New York. ISBN 0-387-95117-2.

Dougherty, E. R. (2001). Small sample issues for microarray-based classification. *Comparative and Functional Genomics*, **2**, 28–34.

Dudewicz, E. J. and Mishra, S. N. (1988). *Modern Mathematical Statistics*. Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, New York. ISBN 0-471-60716-9.

Dudley, R. M. (1978). Central limit theorems for empirical measures. *Annals of Probability*, **6** (6), 899–929.

Dudley, R. M. (1979). †Balls in $\mathbb{R}^k$ do not cut all subsets of $k + 2$ points. *Advances in Mathematics*, **31** (3), 306–308.

Dudley, R. M. (1987). Universal Donsker classes and metric entropy. *Annals of Probability*, **15** (4), 1306–1326.

Dudley, R. M. (1999). *Uniform Central Limit Theorems*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge. ISBN 0-521-46102-2.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, **7** (1), 1–26.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, **78** (382), 316–331.

Efron, B. (1986). How biased is the apparent error rate of a prediction rule. *Journal of the American Statistical Association*, **81** (394), 461–470.

EFRON, B. (1992). Jackknife-after-bootstrap standard errors and influence functions. *Journal of the Royal Statistical Society, Series B*, **54** (1), 83–127.

EFRON, B. and TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap.* Monographs on Statistics and Applied Probability. Chapman & Hall.

EFRON, B. and TIBSHIRANI, R. J. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, **92** (438), 548–560.

FLOYD, S. and WARMUTH, M. K. (1995). Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine Learning*, **21** (3), 269–304.

FREUND, Y. (1998). Self bounding learning algorithms. *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT-98).*

FREUND, Y. and SCHAPIRE, R. E. (1999). Large margin classification using the perceptron algorithm. *Machine Learning*, **37** (3), 277–296.

GAT, Y. (1999). A bound concerning the generalization ability of a certain class of learning algorithms. Tech. Rep. Technical Report No. 548, Department of Statistics, University of California, Berkeley.

GAT, Y. (2000a). Generalization bounds for incremental search classification algorithms. Tech. Rep. Technical Report No. 575, Department of Statistics, University of California, Berkeley.

GAT, Y. (2000b). *Overfit Bounds for Classification Algorithms.* Ph.D. thesis, Department of Statistics, University of California, Berkeley.

GEMAN, S., BIENENSTOCK, E. and DOURSAT, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, **4** (1), 1–58. ISSN 0899-7667.

GINÉ, E. and ZINN, J. (1984). Some limit theorems for empirical processes. *Annals of Probability*, **12** (4), 929–989.

GLICK, N. (1978). †Additive estimators for probabilities of correct classification. *Pattern Recognition*, **10**, 211–222.

GOLDMAN, S. A. (1999). Computational learning theory. *Algorithms and Theory of Computation Handbook.* CRC Press.

GOLEA, M., BARTLETT, P. L., LEE, W. S. and MASON, L. (1998). Generalization in decision trees and DNF: Does size matter? *Advances in Neural Information Processing Systems 10, 1997 (NIPS97)*, pp. 259–265.

GORDON, R. D. (1941). Values of Mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *Annals of Mathematical Statistics*, **12** (3), 364–366.

GRAEPEL, T., HERBRICH, R. and SHAWE-TAYLOR, J. (2005). PAC-Bayesian compression bounds on the prediction error of learning algorithms for classification. *Machine Learning*, **59** (1), 55–76.

GRENANDER, U. (1981). †*Abstract Inference.* Wiley, New York.

GROSS, L. (1975). Logarithmic Sobolev inequalities. *American Journal of Mathematics*, **97**, 1061–1083.

GUERMEUR, Y. (2004). Large margin multi-category discriminant models and scale-sensitive ψ-dimensions. Rapport de recherche 5314, Institut National de Recherche en Informatique et en Automatique. ISRN INRIA/RR–5314–FR+ENG.

GUO, Y., BARTLETT, P. L., SHAWE-TAYLOR, J. and WILLIAMSON, R. C. (2002). Covering numbers for support vector machines. *IEEE Transactions on Information Theory*, **48** (1), 239–250.

GURVITS, L. (1997). †A note on a scale-sensitive dimension of linear bounded functionals in Banach spaces. *Proceedings of Algorithm Learning Theory, ALT-97.*

GUYON, I., BOSER, B. E. and VAPNIK, V. N. (1993). Automatic capacity tuning of very large VC-dimension classifiers. *Advances in Neural*

Information Processing Systems 5 (NIPS'92), pp. 147–155. Morgan Kaufmann, San Mateo, CA.

Hall, P. (1982). Improving the normal approximation when constructing one-sided confidence intervals for binomial or Poisson parameters. *Biometrika*, **69** (3), 647–652.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals (with discussion and rejoinder). *The Annals of Statistics*, **16** (3), 927–953.

Han, T. (1978). †Non negative entropy measures of multivariate symmetric correlations. *Information and Control*, **36**, 133–156.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer Series in Statistics. Springer.

Haussler, D. (1986). Epsilon-nets and simplex range queries. *Proceedings of the Second Annual Symposium on Computational Geometry*, pp. 61–71. ACM Press.

Haussler, D. (1991). Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. Tech. Rep. UCSC-CRL-91-41, University of California, Santa Cruz. Revised March 1992. Later published in Journal of Combinatorial Theory Series A. 1995 Feb;69(2):217–232.

Haussler, D. (1992). Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, **100**, 78–150.

Haussler, D. (1996). Part 1: Overview of the probably approximately correct (PAC) learning framework, and part 2: Decision theoretic generalizations of the PAC model for neural net applications. P. Smolensky, M. C. Mozer and D. E. Rumelhart (editors) *Mathematical Perspectives on Neural Networks.* Lawrence Erlbaum Publishers, Mahwah, NJ. Contains reprinted material from 'Decision Theoretic

Generalizations of the PAC Model for Neural Nets and Other Learning Applications', Information and Computation, Vol. 100, September, 1992, pp. 78-150.

HAUSSLER, D., KEARNS, M., SEUNG, H. S. and TISHBY, N. (1996). Rigorous learning curve bounds from statistical mechanics. *Machine Learning*, **25**, 195–236.

HAUSSLER, D. and LONG, P. M. (1995). A generalization of Sauer's lemma. *Journal of Combinatorial Theory, Series A*, **71** (2), 219–240.

HERBRICH, R. (2002). *Learning Kernel Classifiers: Theory and Algorithms*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA.

HERBRICH, R. and GRAEPEL, T. (2001). A PAC-Bayesian margin bound for linear classifiers: Why SVMs work. *Advances in Neural Information Processing Systems*, vol. 13, pp. 224–230.

HERBRICH, R., GRAEPEL, T. and CAMPBELL, C. (1999). Bayesian learning in reproducing kernel Hilbert spaces. Tech. Rep. TR 99-11, Department of Computer Science, Technical University of Berlin, Franklin Street 28/29, 10587 Berlin. Http://stat.cs.tu-berlin.de/publications.

HERBRICH, R. and WILLIAMSON, R. C. (2002). Algorithmic luckiness. *Journal of Machine Learning Research*, **3**, 175–212. Errata released in 2004.

HERBRICH, R. and WILLIAMSON, R. C. (2004). Errata: "algorithmic luckiness". Available from `http://www.jmlr.org/papers/volume3/herbrich02a/errata.pdf`.

HEWITT, E. and SAVAGE, L. (1955). †Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society*, **80**, 470–501.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, **58**, 13–30.

HUSH, D. R. and SCOVEL, C. (1999). †On a result of Koltchinskii.

HUSH, D. R. and SCOVEL, C. (2001). On the VC dimension of bounded margin classifiers. *Machine Learning*, **45**, 33–44.

HUSH, D. R. and SCOVEL, C. (2004). Fat-shattering of affine functions. *Combinatorics, Probability and Computing*, **13** (3), 353–360.

JAYNES, E. T. (1968). Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, **SSC4**, 227–241.

KÄÄRIÄINEN, M. (2004). Relating the Rademacher and VC bounds. Tech. Rep. C-2004-57, Department of Computer Science, University of Helsinki.

KEARNS, M. J. and SCHAPIRE, R. E. (1994). Efficient distribution-free learning of probabilistic concepts. S. J. Hanson, G. A. Drastal and R. L. Rivest (editors) *Computational Learning Theory and Natural Learning Systems, Volume I: Constraints and Prospect*. Bradford/MIT Press. Also in 31st Annual IEEE Symposium on Foundations of Computer Science, pp. 382–391, 1990.

KLASNER, N. and SIMON, H. U. (1995). From noise-free to noise-tolerant and from on-line to batch learning. *Proceedings of the Eighth Annual Conference on Computational Learning Theory*, pp. 250–264.

KOLTCHINSKII, V. (2001). Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, **47** (5), 1902–1914.

KOLTCHINSKII, V. (2006). 2004 IMS medallion lecture: Local Rademacher complexities and oracle inequalities in risk minimization. *Annals of Statistics*, **34** (6), 2593–2656.

KOLTCHINSKII, V. and PANCHENKO, D. (2000). Rademacher processes and bounding the risk of function learning. E. Gine, D. Mason and J. Wellner (editors) *High Dimensional Probability II*, pp. 443–459. Birkhaüser, Boston.

KOLTCHINSKII, V. and PANCHENKO, D. (2002). Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, **30** (1).

KOLTCHINSKII, V. I. (1982). †On the central limit theorem for empirical measures. *Theory of Probability and Mathematical Statistics (Kiev)*, **24**, 71–82.

KROON, R. S. (2003). *Support Vector Machines, Generalization Bounds and Transduction.* Master's thesis, Department of Computer Science, University of Stellenbosch.

LANGFORD, J. (2002). *Quantitatively Tight Sample Complexity Bounds.* Ph.D. thesis, Department of Computer Science, Carnegie Mellon University.

LANGFORD, J. (2003). Tutorial on practical prediction theory for classification. *International Conference on Machine Learning 2003.* Tutorial presented at conference. Early version of a paper published in JMLR, March 2005.

LANGFORD, J. and BLUM, A. (1999). Microchoice bounds and self bounding learning algorithms. *Computational Learning Theory 1999 (COLT-99)*, pp. 209–214. Another version appeared later in Machine Learning (2002).

LANGFORD, J. and CARUANA, R. (2002). (not) bounding the true error. Paper accompanying the talk "Stochastic Neural Networks" at the 2002 "Bounds less than 0.5" conference.

LANGFORD, J. and MCALLESTER, D. A. (2000). Computable shell decomposition bounds. *Computational Learning Theory 2000 (COLT-2000).* Another version appeared later in JMLR.

LANGFORD, J. and MCALLESTER, D. A. (2004). Computable shell decomposition bounds. *Journal of Machine Learning Research*, **5**, 529–547. An earlier version appeared in the proceedings of COLT-2000.

LANGFORD, J. and SEEGER, M. (2001). Bounds for averaging classifiers. Tech. rep., Carnegie Mellon University.

LANGFORD, J. and SHAWE-TAYLOR, J. (2002). PAC-Bayes and margins. *Neural Information Processing Systems 2002 (NIPS-2002).*

LAURITZEN, S. (2007). Exchangeability and de Finetti's theorem. Graduate lecture series notes. Available from `http://www.stats.ox.ac.uk/~steffen/teaching/grad/definetti.pdf`.

LEDOUX, M. (1996). On talagrand's deviation inequalities for product measures. *ESAIM: Probability and Statistics*, **1**, 63–87.

LEDOUX, M. (1997). Concentration of measure and logarithmic Sobolev inequalities. Lecture notes from Berlin.

LEDOUX, M. (2001). *The Concentration of Measure Phenomenon.* No. 89 in Mathematical Surveys and Monographs. American Mathematical Society.

LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces.* No. 23 in Ergebnisse der Mathematik und Ihrer Grenzgebiete: A Series of Modern Surveys in Mathematics. Springer.

LEHMANN, E. and SCHEFFÉ, H. (1950). †Completeness, similar regions and unbiased estimation. *Sankhya*, **10**, 305–340. Continued in 15:219–236 (1955), and corrections in 17:250 (1956).

LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation.* Springer, New York.

LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses.* Springer Texts in Statistics. Springer, New York.

Littlestone, N. and Warmuth, M. (1986). Relating data compression and learnability. Unpublished manuscript, University of California, Santa Cruz.

Lugosi, G. (2004). Concentration-of-measure inequalities. Material from the 2003 Summer School on Machine Learning at Australian National University, and from the Workshop on Combinatorics, Probability and Algorithms at the University of Montreal's Centre of Mathematical Research.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Mallows, C. (1973). Some comments on c_p. *Technometrics*, **15**, 661–675.

Mansour, Y. and McAllester, D. A. (2000). Generalization bounds for decision trees. *Computational Learning Theory 2000.*

Marton, K. (1986). A simple proof of the blowing-up lemma. *IEEE Transactions on Information Theory*, **32** (3), 445–446.

Massart, P. (1998). Optimal constants for hoeffding type inequalities. Tech. Rep. 98.86, Université Paris-Sud.

Massart, P. (2000). About the constants in talagrand's concentration inequalities for empirical processes. *Annals of Probability*, **28** (2), 863–884.

Massart, P. (2006). *Concentration Inequalities and Model Selection.* Lecture Notes in Mathematics. Springer.

McAllester, D. A. (1998). Some PAC-Bayesian theorems. *Computational Learning Theory 1998.* Also appeared as Machine Learning 37(3): 355-363 (1999).

McAllester, D. A. (1999). PAC-Bayesian model averaging. *Proceedings of the 12th Annual Conference on Computational Learning Theory (COLT-99).* This paper is subsumed by "PAC-Bayesian Stochastic

Model Selection", which appeared in Machine Learning 51(1): 5-21 (2003).

McAllester, D. A. (2001). PAC-Bayesian stochastic model selection. *Machine Learning*, **51** (1), 5–21. Revised version of "PAC-Bayesian Model Averaging", which appeared in COLT-99.

McAllester, D. A. (2003). Simplified PAC-Bayesian margin bounds. *Computational Learning Theory 2003 (COLT-2003).*

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

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

Mendelson, S. and Philips, P. C. (2003). Random subclass bounds. B. Schölkopf and M. Warmuth (editors) *Proceedings of the 16th annual conference on Learning Theory COLT03*, Lecture Notes in Computer Sciences 2777, pp. 329–343. Springer.

Mendelson, S. and Philips, P. C. (2004). On the importance of small coordinate projections. *Journal of Machine Learning Research*, **5**, 219–238.

Nolan, D. and Pollard, D. (1987). U-processes: Rates of convergence. *Annals of Statistics*, **15** (2), 780–799.

Okamoto, M. (1958). †Some inequalities relating to the partial sum of binomial probabilities. *Annals of the Institute of Statistical Mathematics*, **10**, 29–35.

Philips, P. C. (2005). *Data-dependent Analysis of Learning Algorithms.* Ph.D. thesis, Australian National University, Canberra, Australia.

Pollard, D. (1982). †A central limit theorem for empirical processes. *Journal of the Australian Mathematical Society (Series A)*, **33**, 235–248.

Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer, New York.

Pollard, D. (1986). †Rates of uniform almost-sure convergence for empirical processes indexed by unbounded classes of functions. Manuscript.

Pollard, D. (1990). *Empirical Processes: Theory and Applications.* Institute of Mathematical Statistics, Hayward, CA.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org`.

Rio, E. (2000). †Inégalités de concentration pour les processus empiriques de classes de parties. *Probability Theory and Related Fields*, **119**, 163–175.

Rio, E. (2002). †Une inégalité de Bennett pour les maxima de processus empiriques. *Ann. I. H. Poincaré*, **38** (6), 1053–1057.

Rissanen, J. (1978). †Modeling by shortest data description. *Automatica*, **14**, 465–471.

Sauer, N. (1972). On the density of families of sets. *Journal of Combinatorial Theory Series A*, **13**, 145–147.

Schapire, R. E. (1999). A brief introduction to boosting. *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI'99).*

Schapire, R. E., Freund, Y., Bartlett, P. and Lee, W. S. (1997). Boosting the margin: A new explanation for the effectiveness of voting methods. *Proceedings of the 14th International Conference on Machine Learning*, pp. 322–330. Morgan Kaufmann. Also appeared in The Annals of Statistics, 26(5):1651-1686, 1998.

Schölkopf, B., Shawe-Taylor, J., Smola, A. J. and Williamson, R. C. (1999a). Generalization bounds via eigenvalues of the Gram matrix. Technical Report NC2-TR-1999-035, NeuroCOLT2.

Schölkopf, B., Shawe-Taylor, J., Smola, A. J. and Williamson, R. C. (1999b). Kernel-dependent support vector error bounds. D. Willshaw and A. Murray (editors) *Proceedings of the International Conference on Artificial Neural Networks 1999*, vol. 1, pp. 103–108. IEEE Conference Publications.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels — Support Vector Machines, Regularization, Optimization and Beyond*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA. ISBN 0-262-19475-9.

Schwartz, G. (1979). †Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Seeger, M. (2001). The proof of McAllester's PAC-Bayesian theorem.

Seeger, M. (2002). PAC-Bayesian generalization error bounds for Gaussian process classification. *Journal of Machine Learning Research*, **3**, 233–269.

Shah, M. (2006). *Sample Compression, Margins and Generalization: Extensions to the Set Covering Machine*. Ph.D. thesis, University of Ottawa, Canada.

Shah, M. (2007). Sample compression bounds for decision trees. *Proceedings of the Twenty Fourth International Conference on Machine Learning (ICML-2007)*, pp. 799–806. ACM Press, New York, NY, USA.

Shawe-Taylor, J., Anthony, M. and Biggs, N. L. (1993). Bouunding sample size with the Vapnik-Chervonenkis dimension. *Discrete Applied Mathematics*, **42** (1), 65–73.

Shawe-Taylor, J., Bartlett, P. L., Williamson, R. C. and Anthony, M. (1996). A framework for structural risk minimization. Technical Report NC-TR-96-032, NeuroCOLT.

Shawe-Taylor, J., Bartlett, P. L., Williamson, R. C. and Anthony, M. (1998). Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, **44** (5), 1926–1940. Also NeuroCOLT Technical Report NC-TR-96-053, October 1996.

Shawe-Taylor, J. and Cristianini, N. (1998a). Margin distribution bounds on generalization. Technical Report NC-TR-98-020, NeuroCOLT.

Shawe-Taylor, J. and Cristianini, N. (1998b). Robust bounds on generalization from the margin distribution. NeuroCOLT Technical Report NC-TR-98-029, ESPRIT NeuroCOLT2 Working Group.

Shawe-Taylor, J. and Cristianini, N. (1999). Further results on the margin distribution. *Proceedings of the Twelfth Annual Conference on Computational Learning Theory (COLT99).*

Shawe-Taylor, J. and Cristianini, N. (2000). On the generalisation of soft margin algorithms. Technical Report NC-TR-00-082, NeuroCOLT.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis.* Cambridge University Press.

Shawe-Taylor, J. and Williamson, R. C. (1997). A PAC analysis of a Bayesian estimator. *Proceedings of the Tenth Annual Conference on Computational Learning Theory.* ACM Press.

Shelah, S. (1972). †A combinatorial problem: Stability and order of models and theory of infinitory languages. *Pacific Journal of Mathematics*, **41**, 247–261.

Shen, X. and Wang, L. (1997). On methods of sieves and penalization. *Annals of Statistics*, **25** (6), 2555–2591.

Simon, H. U. (1997). Bounds on the number of examples needed for learning functions. *SIAM Journal on Computing*, **26** (3), 751–763. Previously appeared in Proceedings of Computational Learning Theory: Eurocolt ’93, p. 83–94.

Sonnenburg, S., Braun, M. L., Ong, C. S., Bengio, S., Bottou, L., Holmes, G., LeCun, Y., Müller, K., Pereira, F., Rasmussen, C. E., Rätsch, G., Schölkopf, B., Smola, A., Vincent, P., Weston, J. and Williamson, R. C. (2007). The need for open source software in machine learning. *Journal of Machine Learning Research*, **8**, 2443–2466.

Statlib and Tibshirani, R. (2007). *bootstrap: Functions for the Book "An Introduction to the Bootstra"*. S original. R port by Friedrich Leisch. R package version 1.0-21.

Steele, J. M. (1975). †*Combinatorial Entropy and Uniform Limit Laws*. Ph.D. thesis, Stanford University, Stanford, CA.

Sterne, T. E. (1954). Some remarks on confidence or fiducial limits. *Biometrika*, **41**, 275–278.

Stone, M. (1974). Cross-validatory choice and assessment of statistical prediction (with discussion). *Journal of the Royal Statistical Society, Series B*, **36** (2), 111–147.

Talagrand, M. (1988). An isoperimetric theorem on the cube and the Kintchine-Kahane inequalities. *Proceedings of the American Mathematical Society*, **104** (3), 905–909.

Talagrand, M. (1989). Isoperimetry and integrability of the sum of independent Banach-space valued random variables. *The Annals of Probability*, **17** (4), 1546–1570.

Talagrand, M. (1994). Sharper bounds for Gaussian and empirical processes. *The Annals of Probability*, **22**, 20–76.

Talagrand, M. (1995). Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'I.H.E.S.*, **81**, 73–205.

Talagrand, M. (1996a). Majorizing measures: The generic chaining. *The Annals of Probability*, **24** (3), 1049–1103.

Talagrand, M. (1996b). New concentration inequalities for product spaces. *Inventiones Mathematicae*, **126**, 505–563.

Talagrand, M. (1996c). A new look at independence. *Annals of Probability*, **24**, 1–34.

Talagrand, M. (1996d). Transportation cost for Gaussian and other product measures. *Geometric and Functional Analysis*, **6**, 587–600.

Talagrand, M. (2005). *The Generic Chaining.* Springer Monographs in Mathematics. Springer.

Therneau, T. M. and Atkinson, B. (2007). *rpart: Recursive Partitioning.* R port by Brian Ripley <ripley@stats.ox.ac.uk>. R package version 3.1-38. S-PLUS 6.x original at `http://mayoresearch.mayo.edu/mayo/research/biostat/splusfunctions.cf%m`.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the Association for Computing Machinery*, **27** (11), 1134–1142.

van de Geer, S. (2000). *Empirical Processes in M-Estimation.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer Series in Statistics. Springer.

Vapnik, V. N. (1982). *Estimation of Dependences Based on Empirical Data.* Springer Series in Statistics. Springer-Verlag. ISBN 0-387-90733-5. Translated by Samuel Kotz.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory.* Springer-Verlag.

Vapnik, V. N. (1998). *Statistical Learning Theory.* Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications and Control. Wiley-Interscience, New York. ISBN 0-471-03003-1.

Vapnik, V. N. (2007). Personal Communication.

Vapnik, V. N. and Chervonenkis, A. Y. (1971). †On the uniform convergence of relative frequencies to their probabilities. *Theory of Probab. Appl.*, **16** (2), 264–280.

Vidyasagar, M. (2002). *Learning and Generalization: with Applications to Neural Networks*. Communications and Control Engineering. Springer, second ed.

Vollset, S. E. (1993). Confidence intervals for a binomial proportion. *Statistics in Medicine*, **12**, 809–824.

von Luxburg, U. and Bousquet, O. (2004). Distance-based classification with Lipschitz functions. *Journal of Machine Learning Research*, **5**, 669–695.

Wallace, C. S. and Boulton, D. M. (1968). †An information measure for classification. *The Computer Journal*, **11**, 185–195.

Wenocur, R. S. and Dudley, R. M. (1981). †Some special Vapnik-Chervonenkis classes. *Discrete Mathematics*, **33**, 313–318.

Williamson, R. C., Shawe-Taylor, J., Schölkopf, B. and Smola, A. J. (1999). Sample based generalization bounds. Technical Report NC-TR-99-055, NeuroCOLT.

Williamson, R. C., Smola, A. J. and Schölkopf, B. (1998a). Entropy numbers, operators and support vector kernels. Technical Report NC-TR-98-023, NeuroCOLT.

Williamson, R. C., Smola, A. J. and Schölkopf, B. (1998b). Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. Technical Report NC-TR-98-019, NeuroCOLT/Royal Holloway college, University of London.

Williamson, R. C., Smola, A. J. and Schölkopf, B. (2000). Entropy numbers of linear function classes. *Proceedings of the 13th Annual Conference on Computational Learning Theory*, pp. 309–319. Morgan Kaufmann, San Francisco.

Wilson, E. B. (1927). †Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, **22**, 209–212.

Yurinskii, V. (1974). †Exponential bounds for large deviations. *Theory of Probability and Applications*, **19**, 154–155.