Luca Oneto

Model Selection and Error Estimation in a Nutshell

- Monograph -

February 13, 2020

Springer

Do not worry Dad.
You know.
I understood you.
Even your mistakes.
And at the end, I am like you.

Luca Oneto

Foreword

The field of Machine Learning plays an increasingly important role in science and engineering. Over the past two decades, the availability of powerful computing resources has opened the door to synergistic interactions between empirical and theoretical studies of Machine Learning, showing the value of the “learning from example” paradigm in a wide variety of applications. Nevertheless, in many such practical scenarios the performance of many algorithms often depends crucially on manually engineered features or hyperparameter settings, which often make the difference between bad and state-of-the-art performance. Tools from statistical learning theory allow to estimate the statistical performance of learning algorithm and provide a means to better understand the factors that influence algorithm’s behavior, ultimately suggesting ways to improve the algorithms or designing novel ones. This book reviews in an intelligible and synthetic way the problem of tuning and assessing the performance of an algorithm allowing both young researches and experienced data scientists to gain a broad overview of the key problems underlying model selection, state-of-the-art solutions, and key open questions.

Prof. Massimiliano Pontil
London (United Kingdom), February 13, 2020

Massimiliano Pontil is Senior Researcher at Istituto Italiano di Tecnologia, where he leads the CSML research group, and part-time Professor of Computational Statistics and Machine Learning in the Department of Computer

Science at University College London. He received a PhD in Physics from the University of Genoa in 1999 (Advisor Prof. A. Verri). His research interests are in the areas of machine learning, with a focus on statistical learning theory, kernel methods, multitask and transfer learning, online learning, learning over graphs and sparsity regularization. He also has some interests in approximation theory, numerical optimization and statistical estimation, and he has pursued machine learning applications arising in computer vision, bioinformatics and user modeling. Massimiliano has been on the programme committee of the main machine learning conferences, including COLT, ICML and NIPS, he is an Associate Editor of the Machine Learning Journal, an Action Editor for the Journal of Machine Learning Research and he is on the Scientific Advisory Board of the Max Planck Institute for Intelligent Systems Germany. Massimiliano received the Edoardo R. Caianiello award for the Best Italian PhD Thesis on Connectionism in 2002, an EPSRC Advanced Research Fellowship in 2006-2011, and a Best Paper Runner Up from ICML in 2013.

Preface

How can we select the best performing data-driven model? How can we rigorously estimate its generalization error? Statistical Learning Theory (SLT) answers these questions by deriving non-asymptotic bounds on the generalization error of a model or, in other words, by upper bounding the true error of the learned model based just on quantities computed on the available data. However, for a long time, SLT has been considered only an abstract theoretical framework, useful for inspiring new learning approaches, but with limited applicability to practical problems. The purpose of this book is to give an intelligible overview of the problems of Model Selection (MS) and Error Estimation (EE), by focusing on the ideas behind the different SLT-based approaches and simplifying most of the technical aspects with the purpose of making them more accessible and usable in practice. We will start by presenting the seminal works of the 80's until the most recent results, then discuss open problems and finally outline future directions of this field of research.

Prof. Luca Oneto
Genova (Italy), February 13, 2020

Luca Oneto was born in Rapallo, Italy in 1986. He received his BSc and MSc in Electronic Engineering at the University of Genoa, Italy respectively in 2008 and 2010. In 2014 he received his PhD from the same university in the School of Sciences and Technologies for Knowledge and Information Retrieval with the thesis "Learning Based On Empirical Data". In 2017 he obtained the

Italian National Scientific Qualification for the role of Associate Professor in Computer Engineering and in 2018 he obtained the one in Computer Science He worked as Assistant Professor in Computer Engineering at University of Genoa from 2016 to 2019. In 2018 he was co-funder of the spin-off ZenaByte s.r.l. In 2019 he obtained the Italian National Scientific Qualification for the role of Full Professor in Computer Engineering. He is currently Associate Professor in Computer Science at University of Pisa. His first main topic of research is the Statistical Learning Theory with particular focus on the theoretical aspects of the problems of (Semi) Supervised Model Selection and Error Estimation. His second main topic of research is Data Science with particular reference to the solution of real world problems by exploiting and improving the most recent Learning Algorithms and Theoretical Results in the fields of Machine Learning and Data Mining.

Contents

1

Introduction

How can we select the best performing data-driven model and quantify its generalization error? This question has received a solid answer from the field of statistical inference since the last century and before [54, 221].

The now classic approach of parametric statistics [80, 101, 262] identifies a family of models (e.g. linear functions), a noise assumption (e.g. Gaussian) and, given some data, easily provides a criteria for choosing the best model, along with a quantification of the uncertainty or, in modern terms, an estimation of the generalization error in terms of a confidence interval. On the contrary, non-parametric statistics addresses the problem of deriving all the information directly from the data, without any assumption on the model family nor any other information that is external to the data set itself [218, 275]. With the advent of the digital information era, this approach has gained more and more popularity, up to the point of suggesting that effective data-driven models, with the desired accuracy, can be generated by simply collecting more and more data (see the work of Dhar [77] for some insights on this provocative and inexact but, unfortunately, widespread belief).

However, is it really possible to perform statistical inference for building predictive models without any assumption? Unfortunately, the series of no-free-lunch theorems provided a negative answer to this question [276]. They also showed that, in general, is not even possible to solve apparently simpler problems, like differentiating noise from data, no matter how large the data set is [165].

SLT addresses exactly this problem, by trying to find necessary and sufficient conditions for non-parametric inference to build data-driven models from data or, using the language of SLT, learn an optimal model from data [265]. The main SLT results have been obtained by deriving non-asymptotic bounds on the generalization error of a model or, to be more precise, upper and lower bounds on the excess risk between the optimal predictor and the learned model, as a function of the, possibly infinite and unknown, family of models and the number of available samples [265].

An important byproduct of SLT has been the (theoretical) possibility of applying these bounds for solving the problems raised by our first question, about the quality and the performance of the learned model. However, for a long time, SLT has been considered only a theoretical, albeit very sound and deep, statistical framework, without any real applicability to practical problems [263]. Only in the last decade, with important advances in this field, it has been shown that SLT can provide practical answers [10, 25, 85, 110, 147, 195]. We review here the main results of SLT for the purpose to select the best performing data-driven model and to quantify its generalization error. Note that we will not cover all the approaches available in the literature, since it would be almost impossible and also not very useful, but we will cover only the ground-breaking results in the field since all the other approaches can be seen as a small modification or combination of these ground-breaking achievements. Inspired by the idea of Shewchuk [244], our purpose is to provide an intelligible overview of the ideas, the hypotheses, the advantages and the disadvantages of the different approaches developed in the SLT framework. SLT is still an open field of research but can be the starting point for a better understanding of the methodologies able to rigorously assess the performance and reliability of data-driven models.

2

The “Five W” of MS & EE

Before starting with the technical parts we would like to answer five simple questions regarding the problem of MS and EE:

- **What** is MS and EE?
- **Why** should we care?
- **Who** should care?
- **Where** and **When** this problem has been addressed?

For what concerns the first “W”, MS and EE can be defined respectively as the problem of selecting the data-driven model with the highest accuracy and the problem of estimating the error that the selected model will exhibit on previously unseen data by relying only on the available data. Example of MS problems are: choosing between different learning algorithms (e.g. Support Vector Machines [265], Random Forests [51], Neural Networks [39], Gaussian Processes [220], Nearest Neighbor [72], and Decision Trees [216]), setting the hyperparameters of a learning algorithm (e.g. the regularization in Support Vector Machines, the number of layers and neurons per layer in a Neural Network, the depth of a Decision Tree, and k in k-Nearest Neighbor), and choosing the structure of a learning algorithm (e.g. the type of regularizer in Regularize Least Squares [259, 282] and the families of kernels in Multiple Kernel Learning [106]). Once all the choices have been performed during the MS phase and the data-driven model has been built, the EE phase deals with the problem of estimating the error that this model will exibit on previously unseen data based on different tools such as: probability inequalities (e.g. Hoeffding Inequalities [118], Clopper-Pearson Inequalities [67]), concentration inequalities (e.g. Bounded Difference Function [181], Self Bounding Functions [46, 48, 134, 254]), and moment inequalities [47]. Note that, in this work, we will deal only with methods which do not make any assumption on the noise that corrupts the data and that require only quantities that can be computed on the data themselves.

Regarding the second “W”, MS and EE is a fundamental issue when it comes built data driven models. The first reason is that even advanced practitioners and researchers in the field often fail to perform MS and EE in the correct way. A quite representative evidence of this fact is a recent work published in one of the best journals in the field [95] where the performance of the state-of-the-art learning algorithms have been compared on a large series of benchmark datasets. In this work it is stated that Random Forests is the best performing algorithm but in a recent work [271] other researchers showed that the original work [95] contained a pretty serious flaws in how the MS and EE have been performed: in particular, an hold out dataset was missing leading to biased results. The second reason is rather simple: any data scientist wants to build the best data-driven model in terms of accuracy and have an idea of its actual performances when the model will be exploited in a real world environment. The third reason is that a series of no-free-lunch theorems [97, 276] ensure us that MS and EE will be always necessary since there will never exist a golden learning algorithm able to solve all the data related problems in the optimal way. The last, but not less important, reason is that in literature, due to the lack of knowledge of the problem of MS and EE, most of the research findings are false or report biased results [115, 120].

Regarding the “Who”, the answer is a simple consequence of the “Why” since every data scientist, machine learner, data miner, and every practitioner which uses data should understand these concepts in order to provide reliable analyses, results, insights, and models. Probably the most technical issues of MS and EE are not fundamental for practitioners but having a general idea of the problems and the state-of-the-art tools and techniques able to address them is crucial for everyone in order to obtain rigorous and unbiased results.

Finally for the “Where” and “When” the answer, is simpler. The problem of MS and EE have been addressed in very theoretical Machine Learning and advanced Statistics papers mainly from the 1960 until today. Because of its intrinsic theoretical foundations it is prohibitive for a practitioner to read and comprehend the notation and the ideas proposed in these papers. This book is born from this observation and from a quite inspiring work on the gradient descend algorithm [244]. In fact, most of the time, the ideas behind the methods are quite simple and easy to apply but the technicalities and the knowledge taken for granted make most of the available literature hard to comprehend. In this book we will try to keep the presentation as simple as possible by presenting the problems, the hypotheses, and the ideas behind the methods without going in too many technical details but still presenting the state-of-the-art results in the field. In particular, we will start from the first works of 1960 about probability inequalities [5, 6, 18, 19, 35, 36, 38, 67, 76, 89–91, 118, 137, 170, 175, 195, 255], and proceed with the asymptotic analysis [1, 43, 96, 265] of 1970, and concentration inequalities [44, 46–48, 134, 154, 155, 254, 256, 257] of 1980, then move to the finite sample analysis [3, 10, 13, 22, 23, 25–27, 30, 32, 33, 37, 40, 55, 102–105, 140, 141, 146, 150, 152, 153, 159, 160, 163, 172, 176, 178–180, 185, 190, 192, 194, 196, 198, 201, 207, 217, 227, 229, 231–236, 240–242, 260, 264, 278] of 1990, the milestone results of the 2000 about learnability [49, 93, 131, 174, 186, 195, 212, 238, 238, 261], until the most recent results of 2010 on interactive data analysis [41, 56, 58, 83–85, 94, 114, 115, 124, 145, 189, 191, 219, 250, 273].

3

Preliminaries

In this section we will give an overview of the problem of learning based on empirical data. In particular we will first generally discuss about the inference problems with particular reference to the inductive case and the statistical tools exploited to assess the performance of the induction process. Then we will depict, in details, the Supervised Learning (SL) framework, which represents one of the most successful use of the inductive reasoning. In this section we will also introduce the main subject of this monograph: the MS and EE problems.

3.1 Induction and Statistical Inference

Inference is defined as the act or process of deriving logical conclusions from premises known or assumed to be true [54]. According to the philosopher Charles Sanders Peirce [210] there are three main different approaches to inference (see Figure 3.1): deductive, inductive, and abductive reasoning.
In the deductive reasoning based on a rule it is possible to map a case into a result. An example of deductive reasoning is:

- rule: all the swans in that lake are white
- case: there is a swan coming from that lake
- result: this swan is white

In the inductive reasoning a rule is inferred based on one or more examples of the mapping between case and result. An example of inductive reasoning is:

- case: this swan comes from that lake
- result: this swan is white
- rule: all the swans in that lake are white

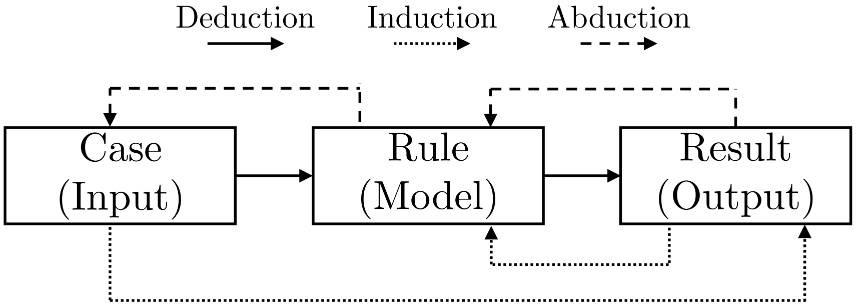

Fig. 3.1. Human inference approaches based on the philosopher Charles Sanders Peirce.

In the abductive reasoning a possible case is inferred based on a result and a rule. An example of abductive reasoning is:

- result: this swan is white
- rule: all the swans in that lake are white
- case: this swan comes from that lake

Deduction [210, 213, 252] is the simplest inference approach since it does not imply any risk. The process is exact in the sense that there is no possibility of making mistakes. Once the rule is assumed to be true and the case is available there is no source of uncertainty in deriving, through deduction, the result. In other words deduction is only able to infer a statement that is already necessarily true from the premises. Mathematics, for example, is one of the most important example of the importance of the deductive reasoning. Inductive reasoning [210, 214], instead, implies a certain level of uncertainty since we are inferring something that is not necessarily true but probable. In particular we only know that it exists at least one case (the one observed) where the inferred rule is correct. The inductive reasoning is a simple inference procedure that allows to increment our level of knowledge, since the induction allows to infer something that is not possible to logically deduce just based on the premises. Since the rule cannot be proved it can be falsified, which means that we can test it with other observations of cases and results [214]. The inductive reasoning is the cornerstone of the scientific method where a phenomenon is observed and a theory is proposed and remains valid until one case falsifies it. Consequently another theory must be developed in order to explain the observations which have falsified the original one. Abduction [210, 214], finally, is the most complex inference approach since abductive reasoning tries to derive a result which, as in the inductive case, cannot be deduced from the premises but, moreover, there is neither one observation which can tell us that the statement is true. By repeating the observations, differently from the inductive case, it is not possible to falsify the statement. In order to falsify the statement, another experiment bust be designed; in other words the phenomenon must be observed from another point of view. Abductive inference is at the basis of the most modern scientific theories where, based on the observed facts and current theories, another theory about unobserved phenomena is built and the experiments for falsifying it are developed.

In this book we will deal only with the problem of induction, since it is the only one that can be exploited for increasing our level of knowledge starting from some observations, and that can be also falsified based on them. In particular we will deal with the problems of inferring the rule, given some examples of cases and results. In the field of learning this problem is called SL, where the rule is an unknown system which maps a point from an input space (the case) to a point in an output space (the result). But this does not conclude the learning process. The quality of the learning procedure must be assessed and tuned by estimating its error based on the desired metric. In order to reach this goal we will make use of statistical inference techniques. Among the others it is possible to identify two main approaches to statistical inference:

- the bayesian one, where the observations are fixed while the unknown parameters that we want the estimate are described probabilistically;
- the frequentist one, where the observations are a repeatable random sample (there is the concept of frequency) while the unknown parameters that we want the estimate remain constant during this repeatable process (there is no way that probabilities can be associated with them).

Bayesian inference derives the posterior probability $\mathbb{P}\{H|E\}$, which represents the probability of the event H since we observed the event E (note that since E has been observed, this implies that $\mathbb{P}\{E\} \neq 0$), based on the Bayes' theorem which states that:

$$\mathbb{P}\{H|E\} = \frac{\mathbb{P}\{E|H\}\mathbb{P}\{H\}}{\mathbb{P}\{E\}}. \tag{3.1}$$

where

- $\mathbb{P}\{E|H\}$ is the likelihood, which is the probability of observing E given H;
- $\mathbb{P}\{H\}$ is the prior probability, that is the probability of H before observing E;

- $\mathbb{P}\{E\}$ is the marginal likelihood (or evidence), computable from $\mathbb{P}\{E|H\}$ and $\mathbb{P}\{H\}$.

The proof of the Bayes' theorem is trivial and comes from the definition of conditional probability. In fact by definition we have that

$$\mathbb{P}\{H|E\} = \frac{\mathbb{P}\{H, E\}}{\mathbb{P}\{E\}}, \tag{3.2}$$

$$\mathbb{P}\{E|H\} = \frac{\mathbb{P}\{H, E\}}{\mathbb{P}\{H\}}. \tag{3.3}$$

Note that if $\mathbb{P}\{E\} = 0$ it automatically implies that

$$\mathbb{P}\{H, E\}, \mathbb{P}\{H|E\}, \mathbb{P}\{H|E\} = 0. \tag{3.4}$$

Analogously if $\mathbb{P}\{H\} = 0$ we have that

$$\mathbb{P}\{H, E\}, \mathbb{P}\{H|E\}, \mathbb{P}\{H|E\} = 0. \tag{3.5}$$

Consequently we can state that

$$\mathbb{P}\{H|E\}\mathbb{P}\{E\} = \mathbb{P}\{E|H\}\mathbb{P}\{H\}, \tag{3.6}$$

which implies the Bayes' theorem. The result of a Bayesian approach is the complete probability distribution of the event H given the available observation E. From the distribution it is possible to obtain any information about H. If H is the value of a parameter that characterizes E we can compute the expected value of the parameter, its credible interval, etc.

Frequentist inference has been associated with the frequentist interpretation of probability. In particular the observation of an experiment can be considered as one of an infinite sequence of possible repetitions of the same experiment. Based on this interpretation, the aim of the frequentist approach is to draw conclusions from data which are true with a given (high) probability, among this infinite set of repetitions of the experiment. Another interpretation, which does not rely on the frequency interpretation of probability, states that probability relates to a set of random events which are defined before the observation takes place. In other words an experiment should include, before performing the experiment, decisions about exactly which steps will be taken to reach a conclusion from future observations. In a frequentist approach to inference unknown parameters are considered as fixed and unknown values that can not be treated as random variables in any sense. The result of a frequentist approach is either a true-or-false conclusion from a significance test or a conclusion in the form that a given sample-derived confidence interval covers the true value: both these conclusions has a given probability of being correct.

Finally we would like to underline that Bayesian and Frequentist statistics can be also described in a more informal way. Bayesian statistics tries to describe in a rigorous way the initial prior belief about a problem and the update in the belief as some observation about the problem are made by creating posterior belief. Frequentist statistics, instead, concerns itself with methods that have guarantees about future observations, no matter the personal belief. Both approaches are quite natural: regarding the Frequentist statistics, if you compute something that occurs many times then you can be assured that it will continue to occur. However, also, the idea of belief is very natural: if you have prior knowledge about your data then you should try to incorporate it. Moreover a typical error while applying the Frequentist approach is to run an experiment, look at the output, and change the experiment design. This represents the act of incorporating prior belief without taking it into account. With Bayesian statistics, you can explicitly deal with beliefs.

However in this book we will mainly adopt the Frequentist approach. One can refer to other works for a general review [110] of the Bayesian approach and the difference with the Frequentist one.

3.2 The Supervised Learning Problem

In the SL framework the goal is to approximate a generally unknown system, or rule, $\mathfrak{S} : \mathcal{X} \rightarrow \mathcal{Y}$, which maps a point x from an input space $\mathcal{X}$ into a point y of an output space $\mathcal{Y}$, through another rule $\mathfrak{R} : \mathcal{X} \rightarrow \mathcal{Y}$ learned by observing $\mathfrak{S}$, which again maps a point $x \in \mathcal{X}$ into a point $\widehat{y} \in \mathcal{Y}$ (see Figure 3.2). The space $\mathcal{Z}$ is defined as the cartesian product between the input and the output space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $z \in \mathcal{Z}$ is a point in this space. There are many kinds of unknown rules $\mathfrak{S}$: the deterministic ones, where a point $x \in \mathcal{X}$ is mapped into a single point $y \in \mathcal{Y}$, and the probabilistic ones, where a point $x \in \mathcal{X}$ can be mapped to different points in $y \in \mathcal{Y}$. In this monograph we assume to always deal with probabilistic rules, since they can model also the deterministic and hybrid ones. Consequently we suppose that $\mathfrak{S}$ can be modelled as a probability distribution μ over $\mathcal{Z}$. Note that, in real world applications, the use of probabilistic rules is necessary, for example, because $\mathcal{X}$ can be just a subspace of all the inputs of the system $\mathfrak{S}$ and then,

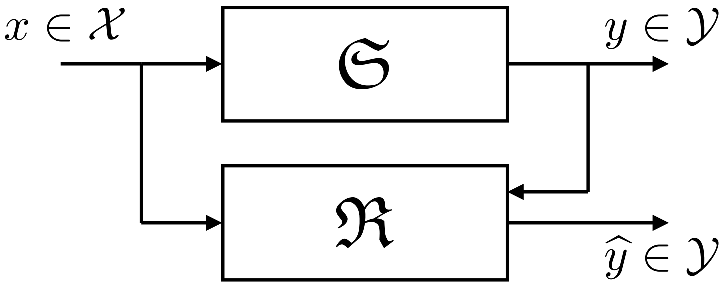

Fig. 3.2. The SL Problem.

even if $\mathfrak{S}$ is a deterministic system, it may degenerate into a probabilistic one. Moreover the inputs and outputs of a real system must be measured, stored and processed, consequently some noise is always introduced and it must be modelled through probabilistic tools.

Based on the properties of the output space it is possible to define different SL problems: the classification problems, where the output space consists in a finite set of possibilities and there is no hierarchy between the different $y \in \mathcal{Y}$, and the regression problems, where $\mathcal{Y}$ is a subset of a possibly infinite set of outputs and there is a hierarchy between the different $y \in \mathcal{Y}$.

From $\mathfrak{S}$ it is possible to retrieve a series of n examples of mapping, which are called labelled samples $\mathcal{D}_n$. Note that, in general, n is fixed, hence we cannot ask for more samples. The goal in the SL framework is to map $\mathcal{D}_n \in \mathcal{Z}^n$ into a rule $\mathfrak{R}$ belonging to a set of possible ones $\mathcal{R}$ during the so called learning phase. The mapping is performed by a learning algorithm: $\mathscr{A} : \mathcal{Z}^n \to \mathcal{R}$. Note that $\mathfrak{R}$ can be:

- a deterministic rule where $\mathfrak{R}$ is a deterministic function $f : \mathcal{X} \to \mathcal{Y}$ and $\mathcal{R}$ is a set of deterministic functions $\mathcal{F}$ (or hypothesis space). In other words, once $f^* \in \mathcal{F}$ is chosen by $\mathscr{A}$ in order to predict the true label $y \in \mathcal{Y}$ associated to a point $x \in \mathcal{X}$, we have to apply f to x and thus we obtain $\widehat{y} = f(x)$ with $\widehat{y} \in \mathcal{Y}$. Note that to each point $x \in \mathcal{X}$ is associated always with the same $\widehat{y} \in \mathcal{Y}$ even if we classify a point $x \in \mathcal{X}$ many times;
- a probabilistic rule (or randomized function) where $\mathfrak{R}$ is a probability distribution Q over a set of deterministic functions $\mathcal{F}$ and $\mathcal{R}$ is a set of possible distributions $\mathcal{Q}$. In other words, once a $\mathsf{Q}^* \in \mathcal{Q}$ is chosen by $\mathscr{A}$ in order to predict the label $y \in \mathcal{Y}$ associated to a point $x \in \mathcal{X}$, we have to sample one function $f \in \mathcal{F}$ according to Q^* and then apply f to x in order to obtain $\widehat{y} = f(x)$ with $\widehat{y} \in \mathcal{Y}$. Note that to each point $x \in \mathcal{X}$ can

be associated to a different $y \in \mathcal{Y}$ if we try to classify a point $x \in \mathcal{X}$ more times.

During the presentation we will deal only with deterministic and probabilistic rules. The same reasoning can be done for $\mathscr{A}$, since it can be:

- a deterministic algorithm, which means that $\mathscr{A}$ returns always the same $\mathfrak{R}$ if $\mathcal{D}_n$ does not change;
- a probabilistic algorithm (or randomized algorithm) where $\mathscr{A}$ may return a different $\mathfrak{R}$ even if we provide the same dataset $\mathcal{D}_n$ to the algorithm $\mathscr{A}$.

The quality of $\mathfrak{R}$ in approximating $\mathfrak{S}$ is measured with reference to a loss function $\ell : \mathcal{R} \times \mathcal{Z} \to \mathbb{R}$. Mostly we will use a $[0, 1]$-bounded loss function since the extension to the $[a, b]$-bounded is trivial, while to analyze the case of unbounded losses, one usually truncates the values at a certain threshold and bounds the probability of exceeding that threshold [111].

Hence it is possible to introduce the concept of generalization (true) error. For a deterministic rule $\mathfrak{R} \in \mathcal{R}$ it is defined as

$$L(\mathfrak{R}) = L(f) = \mathbb{E}_{z \sim \mu}\{\ell(f, z)\}. \tag{3.7}$$

For probabilistic rule, instead, it is defined as

$$L(\mathfrak{R}) = L(\mathsf{Q}) = \mathbb{E}_{z \sim \mu}\mathbb{E}_{f \sim \mathsf{Q}}\{\ell(f, z)\}. \tag{3.8}$$

Note that the generalization error is one of the most informative quantities that we can estimate about $\mathfrak{R}$. It allows us to give a rigorous measure of the error of our approximation by stating that, on average, the error of our approximation with reference to a loss function ℓ is equal to $L(\mathfrak{R})$.

Let us suppose that the probability distribution over $\mathcal{Z}$ is known and that we have access to all the possible rules (probabilistic and deterministic). By all the possible rules we mean all the possible ways of mapping $\mathcal{X}$ to $\mathcal{Y}$. In this case it is possible to build the Bayes' rule, which is the best possible approximation of $\mathfrak{S}$ with reference to a loss function ℓ (see Figure 3.3). The Bayes' rule $\mathscr{B} : \mathcal{X} \to \mathcal{Y}$ is defined as

$$\mathfrak{R}^{\text{Bayes}} : \arg\inf L(\mathfrak{R}). \tag{3.9}$$

Unfortunately, in general it is not possible to have access to all the possible rules but just to a finite set $\mathcal{R}$ of possible ones, therefore, in this case, we can define the best approximation of the Bayes' rule in $\mathcal{R}$, which is $\mathfrak{R}^*$, as

$$\mathfrak{R}^* : \arg \inf_{\mathfrak{R} \in \mathcal{R}} L(\mathfrak{R}). \tag{3.10}$$

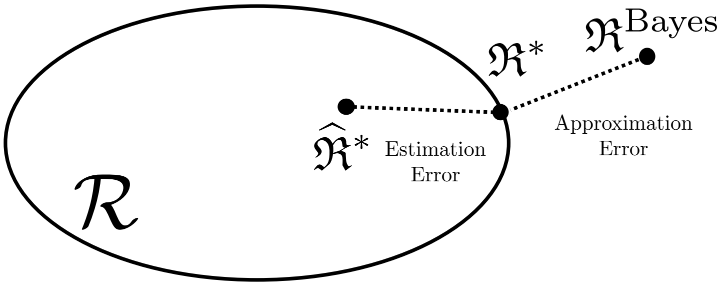

Fig. 3.3. The Bayes' rule, its approximation and the different sources of error.

The error due to the choice of $\mathcal{R}$ is called approximation error since it is the error due to the fact that $\mathfrak{R}^{\text{Bayes}} \notin \mathcal{R}$ (see Figure 3.3). Unfortunately, in general, also the probability distribution over $\mathcal{Z}$ is unknown and consequently $L(\mathfrak{R})$ cannot be computed. The only thing that we can do is to use the algorithm $\mathscr{A}$, which basically maps $\mathcal{D}_n$ into a rule $\mathfrak{R}$ according to an heuristic, that can be more or less strongly theoretically grounded, which tries to find $\mathfrak{R}^* \in \mathcal{R}$ based just on $\mathcal{D}_n$. The result is a rule $\widehat{\mathfrak{R}}^* \in \mathcal{R}$ defined as

$$\widehat{\mathfrak{R}}^* : \mathscr{A}(\mathcal{D}_n). \tag{3.11}$$

The rule $\widehat{\mathfrak{R}}^*$ is affected by the estimation error, with respect to $\mathfrak{R}^*$, due to the fact that the probability distribution over $\mathcal{Z}$ is unknown and just $\mathcal{D}_n$ is available. Consequently, with respect to $\mathfrak{R}^{\text{Bayes}}$, the rule $\widehat{\mathfrak{R}}^*$ is affected by two sources of error: the approximation and the estimation errors (see Figure 3.3).
In this monograph we will deal with the problem of learning from empirical data where the only information available during the learning phase is $\mathcal{D}_n$. Consequently it is possible to identify several problems that arise during any learning process: how $\mathcal{R}$ must be designed, how $\mathscr{A}$ must be designed, how to choose between different set of rules and different algorithms in order to reduce the different sources of error, how the generalization error can be estimated based on $\mathcal{D}_n$.
These issues related to designing $\mathcal{R}$ and $\mathscr{A}$ are out of the scope of this book. For a deep treatment of the problem of algorithms design, implementation and related issues one can refer to many books [39, 63, 113, 237, 274].
The problems that we will face in details in this monograph are, instead the remaining issues related to any learning process: the MS and the EE phases. Any algorithm, more or less explicitly, is characterized by a set of hyperparameters which defines the set of rules from which the algorithm will choose

the final one. Consequently one of the most critical problems in learning is how to choose the best configuration of this hyperparameters h in a set of possible configurations $\mathcal{H}$. The problem could also be generalized to the problem of choosing between different algorithms. Basically one has to choose between different algorithms $\mathscr{A} \in \mathcal{A}$, each one characterized by its configuration of hyperparameters $h \in \mathcal{H}_{\mathscr{A}}$

$$\mathcal{A}_{\mathcal{H}} = \{\mathscr{A}_h : \mathscr{A} \in \mathcal{A}, h \in \mathcal{H}_{\mathscr{A}}\} . \tag{3.12}$$

Note that once the best algorithm $\mathscr{A}_h \in \mathcal{A}_{\mathcal{H}}$ is chosen the final rule can be selected by applying the algorithm to the available data $\mathcal{D}_n$. Consequently the first problem that we will face in this monograph is the one of choosing $\mathscr{A}_h \in \mathcal{A}_{\mathcal{H}}$ based just on the empirical observations ($\mathcal{D}_n$). In the learning process this phase is MS or, more generally, performance tuning phase.

MS is tightly related with the second issue that we will face in this monograph, which is how to estimate the generalization error of the learned rule $\mathfrak{R}$. This phase is called EE or, more generally, performance assessment phase. The link between the MS and EE phases is that the purpose of any learning procedure is to find the best possible approximation of the Bayes' rule given $\mathcal{D}_n$, and by best possible approximation we mean the rule with the smallest generalization error. In order to estimate the generalization error of a $\mathfrak{R}$ the state-of-the-art approach is to use probabilities and concentration inequalities [35, 36, 38, 44, 46–48, 67, 118, 134, 154, 155, 170, 175, 254–257] which allow to prove that

$$\mathbb{P}\{L(\mathfrak{R}) \geqslant \varDelta\} \leqslant \delta(\varDelta), \tag{3.13}$$

where the confidence δ is a function of the accuracy $\varDelta$ and vice-versa. In other words we are able to guarantee, with high probability, that the generalization error will be larger than a particular quantity or alternatively that the probability of the generalization error to be large is small.

The MS phase is tightly related with the EE one since, if we find a way to accurately estimate the generalization error of a rule, we can easily obtain a criterion to select among different rules (i.e. different algorithms or different configurations of the hyperparameters of the algorithm) by choosing the one which minimizes the estimated generalization error. This approach, which is the golden standard, obviously has some drawbacks (see Figure 3.4): the sharp is the estimation of the generalization error the higher will be the probability to select the best model (perform a good MS phase) while the loose is the estimation the lower will be the probability to perform a good MS phase.

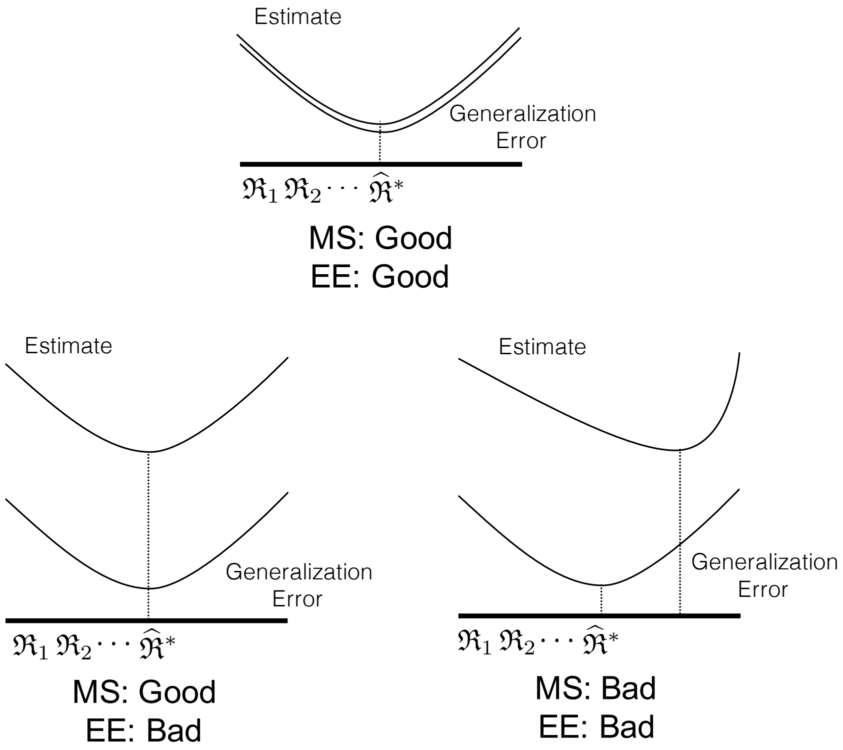
Estimate
Generalization Error
$\mathfrak{R}_1 \mathfrak{R}_2 \cdots \widehat{\mathfrak{R}}^*$
MS: Good
EE: Good
Estimate
Generalization Error
$\mathfrak{R}_1 \mathfrak{R}_2 \cdots \widehat{\mathfrak{R}}^*$
MS: Good
EE: Bad
Estimate
Generalization Error
$\mathfrak{R}_1 \mathfrak{R}_2 \cdots \widehat{\mathfrak{R}}^*$
MS: Bad
EE: Bad

Fig. 3.4. Advantages and disadvantages of using EE for MS purposes.

Note that, in this book, we will just present methods for MS and EE that rely only on $\mathcal{D}_n$ in order to be applied. Each method presented will not require any additional oracle or a-priori knowledge in order to be adopted in practice.

3.3 What Methods Will Be Presented and How?

In this book we will present six approaches to MS and EE:

- Resampling methods (Hold Out, Cross Validation and Bootstrap);
- Complexity-based methods (Union and Shell bounds, V. N. Vapnik and A. Chernovenkis Theory, Rademacher Complexity Theory);
- Compression bound;
- Algorithmic Stability Theory;
- PAC-Bayes Theory;
- Differential Privacy Theory.

At the best knowledge of the authors these six approaches are the state-of-the art ones and any other approach is a modification or combination of the latter. In order to not overcomplicate the presentation and the notation (and avoid the agonizing pain) we will use a different notation, as simple as possible, for each family of methods. This also has the benefit for the reader to focus on just one particular approach.

For each of the methods we will present the general idea, the hypothesis, the field of application, the technicalities of the method itself, its advantages and disadvantages. We will not report too many technical details which can be retrieved in the original papers.

4

Resampling Methods

Resampling methods [10, 19, 119, 137], also called Out-of-Sample methods, are favoured by practitioners because they work well in many situations and allow the application of simple statistical techniques for estimating the quantities of interest. Some examples of resampling methods are the Hold Out (HO), the well-known k-Fold Cross Validation [19, 137] (KCV), the Leave-One-Out [100] (LOO), and the Bootstrap [5, 91] (BTS).

The idea behind these methods is quite simple: if a rule performs well on data that have not been used for selecting the rule itself than probably the rule will generalize, namely it will have small generalization error.

The underlying hypothesis that needs to be clearly expressed for supporting the idea behind the resampling methods is that the data must come from a phenomenon that does not change in time; basically we are assuming that the available data and the future sampled data must be independent and identically distributed (i.i.d.).

Note that these methods can be applied to both deterministic and probabilistic rules and algorithms.

Resampling techniques rely on a similar approach: the original dataset is resampled, with or without replacement, to build two independent datasets called, respectively, the learning and validation (or estimation) sets. The first one is used for learning different rules (through different learning algorithms and different configurations of the hyperparameters for each algorithm), while the second one is exploited for estimating the generalization error of each rule in order to choose the best one (MS phase). Note that the error on the learning set is obviously optimistically biased but also the one on the validation set is optimistically biased since we reuse the data many times in order to select the best rule (the overvalidation phenomenon). Therefore, in order to estimate the generalization error of the final rule, it is necessary to possess a third set, called the test set, by nesting two of the resampling procedures mentioned above. Since the data in the test set are i.i.d. with respect to the learning and validation sets the error of the selected rule on the test set represents an unbiased estimator of the generalization error which can be easily used to estimate it (EE phase).

Note that the resampling procedure itself may introduce artifacts in the estimation process (e.g. unlucky splittings) and must be carefully designed. In particular it is not clear how to split the data when a small sample is available [10] and what the size of learning, validation, and test sets should be with respect to the available observations [6, 20].

More formally let us consider the SL framework where $\mathcal{X}$ and $\mathcal{Y}$ are, respectively, the input and the output spaces. We consider a set of i.i.d. labeled samples $\mathcal{D}_n : \{z_1, \cdots, z_n\}$ of size n where $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $z \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. The samples in $\mathcal{D}_n$ are sampled from an unknown probability distribution μ over $\mathcal{Z}$. A learning algorithm $\mathscr{A}_h$, characterized by its configuration of the hyperparameters h, maps $\mathcal{D}_n$ into a rule $\mathfrak{R} : \mathscr{A}_h(\mathcal{D}_n)$. The rule $\mathfrak{R}$ maps an element of the input space $\mathcal{X}$ into an element of the output space $\mathcal{Y}$. In particular, $\mathscr{A}_h$ allows designing a rule $\mathfrak{R} \in \mathcal{R}$ and the set of rules $\mathcal{R}$. $\mathcal{R}$ can be, in general, unknown [10, 25, 195]. The accuracy of a rule $\mathfrak{R}$ in representing the hidden relationship μ is measured with reference to a loss function $\ell : \mathcal{R} \times \mathcal{Z} \to [0, 1]$. The quantity which we are interested in is the generalization error [265], namely the error that a model will perform on new data generated by μ and previously unseen[1]

$$L(\mathfrak{R}) = \mathbb{E}_z \ell(\mathfrak{R}, z). \tag{4.1}$$

Unfortunately, since μ is unknown, $L(\mathfrak{R})$ cannot be computed and, consequently, must be estimated. Then we have to resort to its empirical estimator, in this case the empirical error [265]

$$\widehat{L}(\mathfrak{R}, \mathcal{D}_n) = \frac{1}{n} \sum_{\boldsymbol{z} \in \mathcal{D}_n} \ell(\mathfrak{R}, z), \tag{4.2}$$

together with its variance [175]

$$\widehat{V}(\mathfrak{R}, \mathcal{D}_n) = \frac{1}{n(n-1)} \sum_{z' \in \mathcal{D}_n} \sum_{z'' \in \mathcal{D}_n} [\ell(\mathfrak{R}, z') - \ell(\mathfrak{R}, z'')]^2. \tag{4.3}$$

[1] For improving the readability of the paper we abbreviate $\mathbb{E}_{z \sim \mu}$ with $\mathbb{E}_z$

As we described before these techniques rely on a similar idea: the original dataset $\mathcal{D}_n$ is resampled once or many (n_r) times, with or without replacement, to build three independent datasets called learning, validation, and test sets, respectively $\mathcal{L}_l^r$, $\mathcal{V}_v^r$, and $\mathcal{T}_t^r$, with $r \in \{1, \cdots, n_r\}$. Note that $\mathcal{L}_l^r \cap \mathcal{V}_v^r = \oslash$, $\mathcal{L}_l^r \cap \mathcal{T}_t^r = \oslash$, and $\mathcal{V}_v^r \cap \mathcal{T}_t^r = \oslash$.

Then, in order to select the best algorithm $\mathscr{A}_h^*$ in a set of possible ones $\mathcal{A}$, together with the best configuration of its hyperparameters chosen in a set of possible ones for each algorithm $\mathcal{H}_{\mathscr{A}}$

$$\mathcal{A}_{\mathcal{H}} = \{\mathscr{A}_h : \mathscr{A} \in \mathcal{A}, h \in \mathcal{H}_{\mathscr{A}}\}, \tag{4.4}$$

or, in other words, to perform the MS phase, we have to apply the following procedure

$$\mathscr{A}_h^* : \min_{\mathscr{A}_h \in \mathcal{A}_{\mathcal{H}}} \frac{1}{n_r} \sum_{r=1}^{n_r} \widehat{L}(\mathscr{A}_h(\mathcal{L}_l^r), \mathcal{V}_v^r). \tag{4.5}$$

Since the data in $\mathcal{L}_l^r$ are i.i.d. from the one in $\mathcal{V}_v^r$, the idea is that $\mathscr{A}_h^*$ should be the algorithm, together with its hyperparameters configuration, which allows to achieve a small error on a dataset that is independent from the training set.

But why are we selecting this criterion for choosing $\mathscr{A}_h^*$? The reason is simple and yet quite theoretical.

Let us suppose that $|\mathcal{A}_{\mathcal{H}}| = 1$. In this case since the data in $\mathcal{V}_v^r$ are i.i.d. with respect to the ones in $\mathcal{L}_l^r$, also the errors that $\mathscr{A}_h(\mathcal{L}_l^r)$ commits on each $z \in \mathcal{V}_v^r$ are i.i.d. Then, by exploiting for example the Hoeffding Inequality [118], we can state that

$$\mathbb{P}_{\mathcal{V}_v^r} \left\{ L(\mathscr{A}_h(\mathcal{L}_l^r)) \geqslant \widehat{L}(\mathscr{A}_h(\mathcal{L}_l^r), \mathcal{V}_v^r) + \varDelta \right\} \leqslant e^{-2vt^2}, \tag{4.6}$$

$$\mathbb{P}_{\mathcal{V}_v^r} \left\{ \left| L(\mathscr{A}_h(\mathcal{L}_l^r)) - \widehat{L}(\mathscr{A}_h(\mathcal{L}_l^r), \mathcal{V}_v^r) \right| \geqslant \varDelta \right\} \leqslant 2e^{-2vt^2}, \tag{4.7}$$

or, alternatively, that with probability $(1 - \delta)$

$$L(\mathscr{A}_h(\mathcal{L}_l^r)) \leqslant \widehat{L}(\mathscr{A}_h(\mathcal{L}_l^r), \mathcal{V}_v^r) + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2v}}, \tag{4.8}$$

$$\left| L(\mathscr{A}_h(\mathcal{L}_l^r)) - \widehat{L}(\mathscr{A}_h(\mathcal{L}_l^r), \mathcal{V}_v^r) \right| \leqslant \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2v}}. \tag{4.9}$$

These exponential bounds state that, with high probability, the distance between the generalization error and the empirical error goes to zero as

$O\left(\sqrt{1/v}\right)$. The larger is $\mathcal{V}_v^r$ the more accurate will be the estimation but the smaller will be $\mathcal{L}_l^r$ and then the less data we have for building the rule. Basically the term $\sqrt{\ln{(1/\delta)}/2v}$ measures the uncertainty due to the fact that instead of observing the whole population we have observed just $\mathcal{V}_v^r$ and then the empirical error has been computed over this set.

If $|\mathcal{A}_\mathcal{H}| = n_c > 1$ we need to have n_c different validation sets $\mathcal{V}_v^r$, one for each element in $\mathcal{A}_\mathcal{H}$ in order to estimate the generalization error of each rule. This is obviously not possible since we should partition $\mathcal{D}_n$ in too many sets of too small cardinality.

At the same time, if $n_c > 1$ and we use always the same validation set $\mathcal{V}_v^r$, in order to estimate the generalization error of each rule we cannot use the Hoeffding Inequality since $\mathcal{V}_v^r$ is used n_c times. The solution is to employ the Hoeffding Inequality [118] in combination with the Bonferroni Correction [45] and obtain that

$$\mathbb{P}_{\mathcal{V}_v^r}\left\{L(\mathscr{A}_h(\mathcal{L}_l^r)) \geqslant \widehat{L}(\mathscr{A}_h(\mathcal{L}_l^r), \mathcal{V}_v^r) + \varDelta\right\} \leqslant n_c e^{-2vt^2}, \quad \forall \mathscr{A}_h \in \mathcal{A}_\mathcal{H}, \qquad (4.10)$$

$$\mathbb{P}_{\mathcal{V}_v^r}\left\{\left|L(\mathscr{A}_h(\mathcal{L}_l^r)) - \widehat{L}(\mathscr{A}_h(\mathcal{L}_l^r), \mathcal{V}_v^r)\right| \geqslant \varDelta\right\} \leqslant 2n_c e^{-2vt^2}, \quad \forall \mathscr{A}_h \in \mathcal{A}_\mathcal{H}, \quad (4.11)$$

or, alternatively, that with probability $(1-\delta)$

$$L(\mathscr{A}_h(\mathcal{L}_l^r)) \leqslant \widehat{L}(\mathscr{A}_h(\mathcal{L}_l^r), \mathcal{V}_v^r) + \sqrt{\frac{\ln{(n_c)}}{2v}} + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2v}}, \quad \forall \mathscr{A}_h \in \mathcal{A}_\mathcal{H} \qquad (4.12)$$

$$\left|L(\mathscr{A}_h(\mathcal{L}_l^r)) - \widehat{L}(\mathscr{A}_h(\mathcal{L}_l^r), \mathcal{V}_v^r)\right| \leqslant \sqrt{\frac{\ln{(n_c)}}{2v}} + \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2v}}, \quad \forall \mathscr{A}_h \in \mathcal{A}_\mathcal{H}. \quad (4.13)$$

In this case the bound is composed by three terms: the first one is the empirical error of the rule on the validation set, the second one $\sqrt{\ln{(n_c)}/2v}$ depends on the number of times we exploited the validation set, and the third one $\sqrt{\ln{(1/\delta)}/2v}$ has the same meaning described above.

At this point we can provide a reason behind the approach of Eq. (4.5). With the approach proposed in Eq. (4.5) we are choosing the algorithms together with the configuration of their hyperparameters which minimized the estimated generalization error of the rules selected by those algorithms (the approach of Figure 3.4) averaged over the n_r repetitions of the splitting procedure.

The EE phase came straightforward from the description of the MS phase, where we can state that the following bounds hold with probability $(1-\delta)$

$$L(\mathscr{A}_h^*(\mathcal{D}_n)) \leqslant \widehat{L}(\mathscr{A}_h^*(\mathcal{D}_n)) + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2t}}, \tag{4.14}$$

$$\left|L(\mathscr{A}_h^*(\mathcal{D}_n)) - \widehat{L}(\mathscr{A}_h^*(\mathcal{D}_n))\right| \leqslant \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2t}}, \tag{4.15}$$

and where $\widehat{L}(\mathscr{A}_h^*(\mathcal{D}_n)) = \frac{1}{n_r}\sum_{r=1}^{n_r} \widehat{L}(\mathscr{A}_h^*(\mathcal{L}_l^r \cup \mathcal{V}_v^r), \mathcal{T}_t^r)$ for brevity.
In fact, in this case, $\mathcal{T}_t^r$ is i.i.d. from $\mathcal{L}_l^r \cup \mathcal{V}_v^r$, it is used just once and the Hoeffding Inequality [118] can be exploited.
Note that after $\mathscr{A}_h^*$ is found, one can extract the best rule by training the algorithm with the whole dataset [16] $\mathscr{A}_h^*(\mathcal{D}_n)$. Note that this approach, although sound and adopted as common practice, is not theoretically grounded. The rigorous approach would be to randomly select one of the rules $\mathscr{A}_h^*(\mathcal{L}_l^r \cup \mathcal{V}_v^r)$ with $r \in \{1, \cdots, n_r\}$ each time a new sample has to be labeled, but this procedure is usually not taken into account for practical reasons [10, 16, 19].
Note also that if $n_r = 1$, if l, v, and t are aprioristically set such that $n = l + v + t$ and if the resample procedure is performed without replacement we get the hold out method [10]. For implementing the complete k-fold cross validation, instead, we have to set $n_r \leqslant \binom{n}{k}\binom{n-n/k}{k}$, $l = (k-2)n/k$, $v = n/k$, and $t = n/k$ and the resampling must be done without replacement [10, 19, 137]. Finally, for implementing the bootstrap, $l = n$ and $\mathcal{L}_l^r$ must be sampled with replacement from $\mathcal{D}_n$, while $\mathcal{V}_v^r$ and $\mathcal{T}_t^r$ are sampled without replacement from the sample of $\mathcal{D}_n$ that has not been sampled in $\mathcal{L}_l^r$ [10, 91]. Note that for the bootstrap procedure $n_r \leqslant \binom{2n-1}{n}$.
Finally note that the bounds of Eqns. (4.14) and (4.15) can be sharpened both in the rate of convergence and both in the constants involved in the bounds [7].
For example we can use a Chernoff-type bound [65] which is sharper when the empirical error is small and it can exhibit a fast convergence rate $O\left(1/t\right)$

$$\left|L(\mathscr{A}_h^*(\mathcal{D}_n)) - \widehat{L}(\mathscr{A}_h^*(\mathcal{D}_n))\right| \leqslant \sqrt{\widehat{L}(\mathscr{A}_h^*(\mathcal{D}_n))\frac{3\ln\left(\frac{2}{\delta}\right)}{t}} + \frac{3\ln\left(\frac{2}{\delta}\right)}{t}, \tag{4.16}$$

where the bound holds with probability $(1-\delta)$.
Another option is to use a Bennet-type bound [35, 175] which is sharper when the variance of the empirical error is small and, as the Chernoff-type bound, it can exhibit a fast convergence rate $O\left(1/t\right)$

$$\left|L(\mathscr{A}_h^*(\mathcal{D}_n)) - \widehat{L}(\mathscr{A}_h^*(\mathcal{D}_n))\right| \leqslant \sqrt{\widehat{V}(\mathscr{A}_h^*(\mathcal{D}_n))\frac{2\ln\left(\frac{3}{\delta}\right)}{t}} + \frac{7\ln\left(\frac{3}{\delta}\right)}{3(t-1)}, \tag{4.17}$$

where the bound holds with probability $(1-\delta)$ and where, for brevity, we set $\widehat{V}(\mathscr{A}_h^*(\mathcal{D}_n)) = \frac{1}{n_r}\sum_{r=1}^{n_r}\widehat{V}(\mathscr{A}_h^*(\mathcal{L}_l^r \cup \mathcal{V}_v^r), \mathcal{T}_t^r)$.
The state-of-art option is to use the Clopper-Pearson bound [67]. The latter, in its original form, can be applied just to the cases when $\ell \in \{0,1\}$ (e.g. the classification problems where $\ell = 0$ if the predicted class is the same of the actual class and $\ell = 1$ otherwise). Based on the result of Clopper-Pearson [67] we can state that

$$L(\mathscr{A}_h^*(\mathcal{D}_n)) \in \begin{bmatrix} \mathtt{Q}\left[\frac{\delta}{2}; t\widehat{L}(\mathscr{A}_h^*(\mathcal{D}_n)), t - t\widehat{L}(\mathscr{A}_h^*(\mathcal{D}_n)) + 1\right], \\ \mathtt{Q}\left[1-\frac{\delta}{2}; t\widehat{L}(\mathscr{A}_h^*(\mathcal{D}_n)) + 1, t - t\widehat{L}(\mathscr{A}_h^*(\mathcal{D}_n))\right] \end{bmatrix}, \tag{4.18}$$

with probability $(1-\delta)$ and where $\mathtt{Q}[p; v, w]$ is the p-th quantile of the Beta distribution with shape parameters v and w. Recently the Clopper-Pearson bound has been extended [59, 193] in order to be applied to the case of $[0,1]$-bounded losses. Let u be a random variable uniformly distributed over $[0,1]$ and let $\{u_1, \cdots, u_{n_t}\}$ be n_t variables sampled i.i.d. from u. Then we can state that

$$L(\mathscr{A}_h^*(\mathcal{D}_n)) \in \begin{bmatrix} \mathtt{Q}\left[\frac{\delta}{2}; t\widehat{L}^u(\mathscr{A}_h^*(\mathcal{D}_n)), t - t\widehat{L}^u(\mathscr{A}_h^*(\mathcal{D}_n)) + 1\right], \\ \mathtt{Q}\left[1-\frac{\delta}{2}; t\widehat{L}^u(\mathscr{A}_h^*(\mathcal{D}_n)) + 1, t - t\widehat{L}^u(\mathscr{A}_h^*(\mathcal{D}_n))\right] \end{bmatrix}, \tag{4.19}$$

with probability $(1-\delta)$ and where

$$\widehat{L}^u(\mathscr{A}_h^*(\mathcal{D}_n)) = \frac{1}{n_r}\sum_{r=1}^{n_r}\frac{1}{t}\sum_{\boldsymbol{z}\in\mathcal{T}_t^r}[\ell(\mathscr{A}_h^*(\mathcal{L}_l^r \cup \mathcal{V}_v^r), z) \geqslant u_i] \tag{4.20}$$

(each u_i is associated with a different $z \in \mathcal{T}_t^r$ and the Iverson bracket notation [121] is exploited).
The pseudocode of the resampling-based MS and EE strategy is summarized and simplified in Algorithm 1.

Algorithm 1: Resampling Methods: MS and EE Strategy.

```
Input: $\mathcal{A}_{\mathcal{H}}$, $\mathcal{D}_n$, Method (HO, LOO, KCV, BOO), $n_r$, $l$, $v$, $t$, and $\delta$
Output: Optimal Model $\mathscr{A}_h^*(\mathcal{D}_n)$ and its estimated generalization error
        $L(\mathscr{A}_h^*(\mathcal{D}_n))$
$L^*_{\text{MS}} = +\infty$;
for $\mathscr{A}_h \in \mathcal{A}_{\mathcal{H}}$ do
    $L_{\text{MS}} = 0$, $L_{\text{EE}} = 0$, $V_{\text{EE}} = 0$ $L^u_{\text{EE}} = 0$ ;
    for $r \leftarrow 1$ to $n_r$ do
        Sample $\{u_1, \cdots u_t\}$ from $u$;
        Split $\mathcal{D}_n$ in $\mathcal{L}^r_l$, $\mathcal{V}^r_v$, and $\mathcal{T}^r_t$ according to the selected Method ;
        $L_{\text{MS}} = L_{\text{MS}} + {}^1\!/_{n_r}\widehat{L}(\mathscr{A}_h(\mathcal{L}^r_l), \mathcal{V}^r_v)$ ;
        $L_{\text{EE}} = L_{\text{EE}} + {}^1\!/_{n_r}\widehat{L}(\mathscr{A}_h(\mathcal{L}^r_l \cup \mathcal{V}^r_v), \mathcal{T}^r_t)$ ;
        $V_{\text{EE}} = V_{\text{EE}} + {}^1\!/_{n_r}\widehat{V}(\mathscr{A}_h(\mathcal{L}^r_l \cup \mathcal{V}^r_v), \mathcal{T}^r_t)$ ;
        $L^u_{\text{EE}} = L^u_{\text{EE}} + \frac{1}{n_r t}\sum_{\boldsymbol{z} \in \mathcal{T}^r_t}[\ell(\mathscr{A}_h^*(\mathcal{L}^r_l \cup \mathcal{V}^r_v), z) \geqslant u_i]$ ;
    if $L^*_{MS} > L_{MS}$ then
        $L^*_{\text{MS}} = L_{\text{MS}}$;
        $\mathscr{A}_h^*(\mathcal{D}_n) = \mathscr{A}_h(\mathcal{D}_n)$;
        Estimate $L(\mathscr{A}_h^*(\mathcal{D}_n))$ with one of the bounds of Eqns. (4.14),
          (4.16), (4.17), and (4.19) based on $\delta$, $\widehat{L}(\mathscr{A}_h^*(\mathcal{D}_n)) = L_{\text{EE}}$,
          $\widehat{V}(\mathscr{A}_h^*(\mathcal{D}_n)) = V_{\text{EE}}$, and $\widehat{L}^u(\mathscr{A}_h^*(\mathcal{D}_n)) = L^u_{\text{EE}}$;
```

5

Complexity-Based Methods

The idea behind the complexity-based methods is that if an algorithm chooses from a small set of rules it will probably generalize. Basically, if we have a small set of rules and one of them has small empirical error, the risk of overfitting the data is small since the probability that this event has happened by chance is small. Vice versa if we have a large set of rules and one of them has small empirical error the risk that this event has happened for chance is high.

Complexity-based methods are probably the most investigated methods because of their relation with many state-of-the-art learning algorithms. In fact, most learning algorithms define a set of rules and in this set they select the one with the minimum empirical error. This procedure is called Empirical Risk Minimization [265] (ERM). ERM usually leads to overfit the training set and for this reason the set of rules must be resized in order to be neither too rich to overfit the available data, leading to large generalization error, nor too simple. This allows to have small empirical error, leading to a small generalization error. The process or resizing the set of rules in order to achieve small generalization error is called Structural Risk Minimization [265] (SRM) and it is obviously connected to the MS phase. Another key concept in complexity-based methods is the localization principle, which answers a simple question: if a set of rules contains few useful rules (rules with small empirical error) and many useful rules (rules with high empirical error which will never be selected by any learning algorithm), is it really large? The answer is that just the rules with small empirical error should be taken into account when measuring the size of a set of rules, while the other rules should be disregarded (since they will never be selected by the learning algorithm).

Complexity-based methods apply to deterministic rules and algorithms. The set of rules from which the algorithm chooses must be known, before observing the data, and the available samples must be i.i.d. samples. Obviously complexity-based methods cannot be applied to many learning algorithms for which the set of rules in unknown or data dependent.

In this section we will show the three main results in the context of the complexity-based methods:

1. the first approaches deal with the problem of finite sized sets of rules: the Union Bound method [45, 265], which takes into account the whole set of rules, and the Shell Bound method [148, 149], which takes into account just the rules with small empirical error;
2. the second approaches are based on the seminal work of V. N. Vapnik and A. Chernovenkis and deal with infinite sized sets of rules for the particular case of binary classification: the VC-Theory [265], which takes into account the whole set of rules, and the Local VC-Theory [192] which takes into account just the rules with small empirical error. Extensions to the general SL framework have been proposed during the year [28, 240, 265, 280], but overcomplicated and made obsolete by the Rademacher Complexity Theory;
3. the last approach is the Rademacher Complexity Theory which deals with infinite sized sets of rules and the general SL framework: the Global Rademacher Complexity Theory [30, 139, 194, 196], which takes into account the whole set of rules, and the Local Rademacher Complexity Theory [26, 27, 140, 164, 198] which takes into account just the rules with small empirical error.

For all the complexity-based approaches, we will use a common notation that we presented here. We recall then the standard SL framework [30, 265], where the goal is to approximate a relationship between inputs from a set $\mathcal{X}$ and outputs from a set $\mathcal{Y}$. The relationship between inputs and outputs is encoded by a fixed, but unknown, probability distribution μ over $\mathcal{X} \times \mathcal{Y}$. The element $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is defined as a labelled sample: the training phase consists in exploiting a set $\mathcal{D}_n : \{(x_1, x_1), \cdots, (x_n, y_n)\}$ of labelled samples in a learning algorithm $\mathscr{A}_{\mathcal{F}}$, which returns a function $f : \mathcal{X} \to \mathcal{Y}$ chosen in a fixed set $\mathcal{F}$ of possible functions. The learning algorithm maps $\mathcal{D}_n$ to $f \in \mathcal{F}$ and the accuracy in representing the hidden relationship μ is measured with reference to a loss function $\ell : \mathcal{F} \times \mathcal{X} \times \mathcal{Y} \to [0, 1]$. For any $f \in \mathcal{F}$, we define the generalization error $L(f)$ as the expectation of $\ell(f(x), y)$ with respect to μ

$$L(f) = \mathbb{E}_{x,y}\ell(f(x), y), \tag{5.1}$$

where we assume that each labelled sample is i.i.d. and generated according to μ. Since μ is unknown, we can only compute its empirical estimate, the empirical error

$$\widehat{L}(f) = \frac{1}{n}\sum_{i=1}^{n}\ell(f(x_i), y_i). \tag{5.2}$$

together with the empirical variance

$$\widehat{V}(f) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=2}^{n}[\ell(f(x_i), y_i) - \ell(f(x_j), y_j)]^2. \tag{5.3}$$

All the quantities strictly related with each particular theory will be presented in each subsection.
All the fully empirical bounds on the generalization ability of a function $f \in \mathcal{F}$ that will be presented in each following subsection of the Complexity based methods can be used for MS and EE purposes as described in the preliminaries [10, 25, 240]. In fact the bounds will always have the following form

$$\mathbb{P}_{\mathcal{D}_n}\{L(f) \leqslant \mathit{\Delta}(f, \mathcal{D}_n, \mathcal{F}, \delta)\} \geqslant 1 - \delta, \quad \forall f \in \mathcal{F}. \tag{5.4}$$

Then if we want to choose $\mathcal{F}^* \in \{\mathcal{F}_1, \cdots, \mathcal{F}_{n_\mathcal{F}}\}$, namely perform the MS phase, and estimate the generalization performance of $f^* = \mathscr{A}_{\mathcal{F}^*}(\mathcal{D}_n)$, namely perform the EE phase, we have to follow the procedure summarized in Algorithm 2. Note that the generalization of the final model is bounded by

$$L(f^*) \leqslant \mathit{\Delta}\left(f^*, \mathcal{D}_n, \mathcal{F}^*, \frac{\delta}{n_\mathcal{F}}\right), \quad \forall f^* \in \mathcal{F}^*, \quad \forall \mathcal{F}^* \in \{\mathcal{F}_1, \cdots, \mathcal{F}_{n_\mathcal{F}}\}, \tag{5.5}$$

with probability $(1 - \delta)$, since we have applied the Bonferroni correction [45] over the $n_\mathcal{F}$ choices for the space of functions [240, 265]. Note that, in Algorithm 2, contrarily to the Resampling methods, the whole data are used both for training, MS, and EE purposes.

5.1 Union and Shell Bounds

Union and Shell bounds deal with the case of finite $|\mathcal{F}|$. In this case bounding the generalization error of a function chosen in $\mathcal{F}$ based on $\mathcal{D}_n$ is trivial,

Algorithm 2: Complexity-based Methods: MS and EE Strategy.

```
Input: ${\mathcal{F}_1, \cdots, \mathcal{F}_{n_\mathcal{F}}}$, $\mathcal{D}_n$, and $\delta$
Output: Optimal Model $f^*$ and its estimated generalization error $L(f^*)$
$L^*_{\mathrm{MS}} = +\infty$;
for $\mathcal{F} \in \{\mathcal{F}_1, \cdots, \mathcal{F}_{n_\mathcal{F}}\}$ do
    $f = \mathscr{A}_\mathcal{F}(\mathcal{D}_n)$;
    $L_{\mathrm{MS}} = \mathit{\Delta}(f, \mathcal{D}_n, \mathcal{F}, \delta)$;
    if $L^*_{MS} > L_{MS}$ then
        $L^*_{\mathrm{MS}} = L_{\mathrm{MS}}$;
        $f^* = \mathscr{A}_\mathcal{F}(\mathcal{D}_n)$;
        $L(f^*) = \mathit{\Delta}\left(f^*, \mathcal{D}_n, \mathcal{F}, \frac{\delta}{n_\mathcal{F}}\right)$;
```

basically one has to apply any probability inequality (Hoeffding [118] or Bennett [35] or Berstein [38] or Chernoff [65] or Clopper-Pearson [67, 193] type inequalities) together with the Bonferroni correction [45].

By applying this procedure we can state with probability $(1-\delta)$ that [45, 118, 265]

$$\left|L(f) - \widehat{L}(f)\right| \leqslant \sqrt{\frac{\ln\left(|\mathcal{F}|\frac{2}{\delta}\right)}{2n}}, \quad \forall f \in \mathcal{F}, \tag{5.6}$$

which is a bound which shows slow rate of convergence $O\left({}^1\!/\!{}_n\right)$.

In order to improve the rate of convergence we can use a Chernoff-type bound [65], which is sharper when the empirical error is small and it can exhibit a fast convergence rate $O\left({}^1\!/\!{}_n\right)$

$$\left|L(f) - \widehat{L}(f)\right| \leqslant \sqrt{\widehat{L}(f)\frac{3\ln\left(|\mathcal{F}|\frac{2}{\delta}\right)}{n}} + \frac{3\ln\left(|\mathcal{F}|\frac{2}{\delta}\right)}{n}, \quad \forall f \in \mathcal{F}, \tag{5.7}$$

where the bound holds with probability $(1-\delta)$. Another option is to use a Bennet-type bound [35, 175], which is sharper when the variance of the empirical error is small and, as the Chernoff-type bound, it can exhibit a fast convergence rate $O\left({}^1\!/\!{}_n\right)$

$$\left|L(f) - \widehat{L}(f)\right| \leqslant \sqrt{\widehat{V}(f)\frac{2\ln\left(|\mathcal{F}|\frac{3}{\delta}\right)}{n}} + \frac{7\ln\left(|\mathcal{F}|\frac{3}{\delta}\right)}{3(n-1)}, \quad \forall f \in \mathcal{F}, \tag{5.8}$$

where the bound holds with probability $(1-\delta)$.

The state-of-art option is to use the Clopper-Pearson bound [67] recently extended [59, 193] in order to be applied to the case of $[0,1]$-bounded losses. Let

u be a random variable uniformly distributed over $[0,1]$ and let $\{u_1, \cdots, u_n\}$ be n variables sampled i.i.d. from u. Then we can state that

$$L(f) \in \left[\begin{array}{l} \mathtt{Q}\left[\frac{\delta}{2|\mathcal{F}|}; n\widehat{L}^u(f), n - n\widehat{L}^u(f) + 1\right], \\ \mathtt{Q}\left[1 - \frac{\delta}{2|\mathcal{F}|}; n\widehat{L}^u(f) + 1, n - n\widehat{L}^u(f)\right] \end{array} \right], \quad \forall f \in \mathcal{F}, \tag{5.9}$$

with probability $(1-\delta)$, where $\mathtt{Q}[p; v, w]$ is the p-th quantile of the Beta distribution with shape parameters v and w and $\widehat{L}^u(f) = \frac{1}{n}\sum_{i=1}^{n}[\ell(f(x_i), y_i) \geqslant u_i]$ (the Iverson bracket notation [121] is exploited).

The above mentioned bounds, even if tight and with good rates of convergence, are not sound from a learning point of view. In fact they take into account the whole set of functions inside $\mathcal{F}$, while, in practical application, only the function with small empirical error will be selected by the learning algorithm. Therefore, just the functions in $\mathcal{F}$ with small empirical error should be taken into account in the above mentioned bounds.

In order to address this issue we will show the approach of the Shell Bound [148, 149]. First we will present a naive version and then we will report the state-of-the-art bounds.

We can start by defining a subset of the original space of functions $\mathcal{F}$

$$\mathcal{F}(p) = \left\{ f : f \in \mathcal{F}, L(f) \in \left[\frac{1}{n}\lceil np - 1\rceil, \frac{1}{n}\lceil np\rceil\right]\right\}, \tag{5.10}$$

which are called shells of $\mathcal{F}$. Note that the shell may be empty, hence we need to take care of this problem as we will see later. Moreover the number of distinct shells are n, one for $L(f) \in \left[0, \frac{1}{n}\right]$, one for $L(f) \in \left[\frac{1}{n}, \frac{2}{n}\right]$, and so on until $L(f) \in \left[\frac{n-1}{n}, 1\right]$.

Then if we choose a function in one of these shells we can say that with probability $(1-\delta)$

$$L(f) \leqslant \widehat{L}(f) + \sqrt{\frac{\ln\left(|\mathcal{F}(p)|\frac{1}{\delta}\right)}{2n}}, \quad \forall f \in \mathcal{F}(p). \tag{5.11}$$

Note that $|\mathcal{F}(p)| \geqslant 1$ since we selected a function in it. Nevertheless, since μ is unknown, and then also $L(f)$ is unknown, we do not know what shell f belongs to and then we have to take the worst case scenario of $L(f)$ and also apply the Bonferroni Correction [45] over the number of shells (n). Then we can state [149] that with probability $(1-\delta)$

$$L(f) \leqslant \max_{p \in [0,1]} \left\{ p : p \leqslant \widehat{L}(f) + \sqrt{\frac{\ln\left(\max\left(1, |\mathcal{F}(p)|\right)\frac{n}{\delta}\right)}{2n}} \right\}, \quad \forall f \in \mathcal{F}, \tag{5.12}$$

note that the $\max(1,\cdot)$ takes care of the fact that some shells may be empty. Unfortunately the bound of Eq. (5.12) cannot be adopted in practice since $|\mathcal{F}(p)|$ cannot be computed because $L(f)$ is unknown. With a rather technical but simple proof it is possible to prove [149, 177] that with probability $(1-\delta)$

$$|\mathcal{F}(p)| \leqslant \left|\widehat{\mathcal{F}}(p,\delta)\right| = 2\left|\left\{ f : f \in \mathcal{F}, \left|\widehat{L}(f) - p\right| \leqslant \frac{1}{n} + \sqrt{\frac{\ln\left(\frac{8n}{\delta}\right)}{2n-1}}\right\}\right|, \tag{5.13}$$

which means that the number of functions in a shell defined based on the generalization error is upper bounded by the number of functions in another shell defined based on the empirical error. By combining Eqns. (5.12) and (5.13) we obtain that with probability $(1-\delta)$

$$L(f) \leqslant \max_{p\in[0,1]} \left\{ p : p \leqslant \widehat{L}(f) + \sqrt{\frac{\ln\left(\max\left(1, \left|\widehat{\mathcal{F}}\left(p, \frac{\delta}{2n}\right)\right|\right)\frac{2n}{\delta}\right)}{2n}} \right\}, \quad \forall f \in \mathcal{F}, \tag{5.14}$$

which is the fully empirical Shell bound. Note that the bound of Eq. (5.14) takes into account only the functions in $\mathcal{F}$ with small empirical error. Unfortunatelly the bound of Eq. (5.14) shows slow rate of convergence.

In order to improve the rate of convergence we can use its Chernoff-type [65] version, which is sharper when the empirical error is small and it can exhibit a fast convergence rate $O\left({}^{1}\!/_{n}\right)$

$$L(f) \leqslant \max_{p\in[0,1]} \left\{ p : p \leqslant \widehat{L}(f) + \sqrt{\widehat{L}(f)\frac{3\ln\left(\frac{2n\max\left(1,\left|\widehat{\mathcal{F}}\left(p,\frac{\delta}{2n}\right)\right|\right)}{\delta}\right)}{n}} \right. \tag{5.15}$$

$$\left. + \frac{3\ln\left(\frac{2n\max\left(1,\left|\widehat{\mathcal{F}}\left(p,\frac{\delta}{2n}\right)\right|\right)}{\delta}\right)}{n} \right\}, \quad \forall f \in \mathcal{F},$$

where the bound holds with probability $(1-\delta)$. Another option is to use a Bennet-type bound [35, 175], which is sharper when the variance of the empirical error is small and, as the Chernoff-type bound, it can exhibit a fast convergence rate $O\left({}^{1}\!/_{n}\right)$

$$L(f) \leqslant \max_{p\in[0,1]} \left\{ p : p \leqslant \widehat{L}(f) + \sqrt{\widehat{V}(f)\frac{2\ln\left(\frac{3n\max\left(1,\left|\widehat{\mathcal{F}}\left(p,\frac{\delta}{2n}\right)\right|\right)}{\delta}\right)}{n}} + \frac{7\ln\left(\frac{3n\max\left(1,\left|\widehat{\mathcal{F}}\left(p,\frac{\delta}{2n}\right)\right|\right)}{\delta}\right)}{3(n-1)} \right\}, \quad \forall f \in \mathcal{F}, \tag{5.16}$$

where the bound holds with probability $(1-\delta)$.

The state-of-art option is to use the Clopper-Pearson bound [67] recently extended [59, 193] in order to be applied to the case of $[0,1]$-bounded losses. Let u be a random variable uniformly distributed over $[0,1]$ and let $\{u_1, \cdots, u_n\}$ be n variables sampled i.i.d. from u. Then we can state that

$$L(f) \leqslant \max_{p\in[0,1]} \left\{ p : p \leqslant \mathtt{Q}\left[\frac{\delta}{2n\max\left(1,\left|\widehat{\mathcal{F}}\left(p,\frac{\delta}{2n}\right)\right|\right)}; n\widehat{L}^u(f), n - n\widehat{L}^u(f) + 1 \right] \right\}, \quad \forall f \in \mathcal{F}, \tag{5.17}$$

with probability $(1-\delta)$, where $\mathtt{Q}[p; v, w]$ is the p-th quantile of the Beta distribution with shape parameters v and w and and $\widehat{L}^u(f) = \frac{1}{n}\sum_{i=1}^{n}[\ell(f(x_i), y_i) \geqslant u_i]$ (the Iverson bracket notation [121] is exploited).

5.2 V. N. Vapnik and A. Chernovenkis Theory

Measuring the complexity of a set of rules is of crucial importance for a learning system, because it allows effective controlling of the learning process itself and careful trade-off of possible under- and over-fitting effects in model inference. Starting from the 1960s, Information Theory [246] and SLT [266] opened deep insights in this respect and clearly showed that naïve complexity measures, such as the number of parameters of a model or the number of rules in a set, are not able to guide a learning process toward the selection of rules with good generalization capabilities. Simple one-parameter rules exist that can over-fit any dataset, while some infinite-parameter models can achieve good generalization [265].

The two classical approaches mentioned earlier resulted in the definition of advanced complexity measures, capable of better evaluating the actual hypothesis space learning capacity. Kolmogorov, Chaitin, and Solomonoff independently introduced the concept of Kolmogorov Complexity [184], a general and powerful notion of complexity that, unfortunately, is not computable. Later, approximations of the Kolmogorov Complexity were proposed, such as the Minimum Description Length [109], which has several practical applications although the connection with the Kolmogorov Complexity is not rigorous. Following another path, i.e. taking inspiration from Popper's ideas [214], V. N. Vapnik and A. Chernovenkis developed a complete and computable theoretical framework for characterizing the learning process [68] and suggested several measures of complexity, such as the Vapnik-Chervonenkis (VC) Entropy, the Growth Function and the VC-Dimension [62, 64, 188, 265, 279]. Original VC-Theory mainly deals with binary classification problems and not the general SL framework. Since then several extensions have been proposed in the literature. For example, Kearns and Schapire [132] introduced a generalization of the VC-Dimension to real-valued functions, which is known as the Fat-Shattering Dimension [29].

Unfortunately, the VC-Theory has a global approach, because it takes into account all the rules in the sets of rules, and it is data-independent, because it does not take into account the actual distribution of the data available for learning. As a consequence of targeting this worst-case learning scenario, the VC-Dimension leads to very pessimistic generalization bounds. In order to deal with one of these issues, effective data-dependent complexity measures have been developed, which allow to take into account the actual distribution of the data and produce tighter estimates of the complexity of the class based on the actual learning problem. As an example, data-dependent versions of the VC-Theory have been developed in [46, 240]. In recent years researchers have also succeeded in developing local data-dependent complexity measures [26, 27, 40, 70, 140, 192, 198]. Local measures improve over global ones thanks to their ability of taking into account only those rules of the rules class that will be most likely chosen by the learning procedure, i.e. the models with small error. In particular, a localized version of a complexity measure based on the VC-Theory, called Local VC-Entropy, can be introduced [192]. The localization of the VC-Entropy allows us to introduce the same improvements achieved by Shell Bound into the VC-Theory as well, like, for example, the derivation of refined generalization bounds with respect to their global counterparts. Based on this new localized notion of complexity, it is also possible to derive a generalization bound that does not take into account all the functions in the set but only the ones with small error.

In this section we will start by presenting first the original VC-Theory and then we will present the novel Local VC-Theory.

In order to present the VC-theory let us define some preliminary quantities. Let us consider only the case when $\mathcal{Y} = \{\pm 1\}$ and that a $\{0, 1\}$-loss function is exploited $\ell(f(x), y) = [yf(x) \leqslant 0]$ (the Iverson bracket notation [121] is exploited). Then we define

$$\mathcal{F}|_{\mathcal{D}_n} = \left\{ \{f_1, \cdots, f_i, \cdots, f_n\} \,\middle|\, f \in \mathcal{F} \right\}, \tag{5.18}$$

where $f_i = f(x_i)$, and we recall the definition of VC-Entropy $\mathsf{H}_n(\mathcal{F})$, Annealed VC-Entropy $\mathsf{A}_n(\mathcal{F})$ and Growth Function $\mathsf{G}_n(\mathcal{F})$, along with their empirical versions $\widehat{\mathsf{H}}_n(\mathcal{F})$, $\widehat{\mathsf{A}}_n(\mathcal{F})$ and $\widehat{\mathsf{G}}_n(\mathcal{F})$, respectively [265]

$$\mathsf{H}_n(\mathcal{F}) = \mathbb{E}_{\mathcal{D}_n} \ln\left(|\mathcal{F}|_{D_n}|\right), \tag{5.19}$$

$$\mathsf{A}_n(\mathcal{F}) = \ln\left(\mathbb{E}_{\mathcal{D}_n} |\mathcal{F}|_{D_n}|\right), \tag{5.20}$$

$$\mathsf{G}_n(\mathcal{F}) = \max_{\mathcal{D}_n} \ln\left(|\mathcal{F}|_{D_n}|\right), \tag{5.21}$$

$$\widehat{\mathsf{H}}_n(\mathcal{F}) = \ln\left(|\mathcal{F}|_{\mathcal{D}_n}|\right), \tag{5.22}$$

$$\widehat{\mathsf{A}}_n(\mathcal{F}) = \widehat{\mathsf{H}}_n(\mathcal{F}), \tag{5.23}$$

$$\widehat{\mathsf{G}}_n(\mathcal{F}) = \widehat{\mathsf{H}}_n(\mathcal{F}), \tag{5.24}$$

where $|\cdot|$ is the cardinality of a set. In practice, $\widehat{\mathsf{H}}_n(\mathcal{F})$, $\widehat{\mathsf{A}}_n(\mathcal{F})$ and $\widehat{\mathsf{G}}_n(\mathcal{F})$ count the number of distinct functions on the available data. Moreover by applying Jensen's inequality [126] $\mathsf{H}_n(\mathcal{F}) \leqslant \mathsf{A}_n(\mathcal{F})$ and $\mathsf{A}_n(\mathcal{F}) \leqslant \mathsf{G}_n(\mathcal{F})$ because the worst-case scenario is selected. Finally we can recall the definition of the V. N. Vapnik and A. Chernovenkis dimension [265]

$$d_{\mathrm{VC}}(\mathcal{F}) = \max_{n \in \{0,1,2,\cdots\}} \{n : \mathsf{G}_n(\mathcal{F}) = \ln(2^n)\}. \tag{5.25}$$

Basically the d_{VC} is the maximum number of samples coming from μ that a space of function $\mathcal{F}$ is able to perfectly classify, no matter the configuration of the labels [265].

The Annealed VC-Entropy is the milestone of the Vapnik's results [269] since it allows to bound the generalization error given the empirical one (and vice versa). Given a space of functions $\mathcal{F}$ and a dataset $\mathcal{D}_n$ it is possible to state that

$$\mathbb{P}_{\mathcal{D}_n} \left\{ \sup_{f \in \mathcal{F}} \left[\frac{L(f) - \widehat{L}(f)}{\sqrt{L(f)}} \right] \geqslant t \right\} \leqslant 4 \exp\left[\left(\frac{\mathsf{A}_{2n}(\mathcal{F})}{n} - \frac{t^2}{4} \right) n \right], \tag{5.26}$$

$$\mathbb{P}_{\mathcal{D}_n} \left\{ \sup_{f \in \mathcal{F}} \left| L(f) - \widehat{L}(f) \right| \geqslant t \right\} \leqslant 4 \exp\left[\left(\frac{\mathsf{A}_{2n}(\mathcal{F})}{n} - t^2 \right) n \right]. \tag{5.27}$$

Unfortunately, this bound is not computable since $\mathsf{A}_{2n}(\mathcal{F})$ cannot be compute because μ is unknown, but thanks to a recent result appeared in literature it is possible to derive its computable version. In particular it is possible to prove that the Annealed VC-Entropy is concentrated around its expected value [46, 192] with probability at least $(1-\delta)$

$$\mathsf{A}_{2n}(\mathcal{F}) \leqslant 8\widehat{\mathsf{A}}_n(\mathcal{F}) + 16\ln\left(\frac{1}{\delta}\right) = 8\widehat{\mathsf{H}}_n(\mathcal{F}) + 16\ln\left(\frac{1}{\delta}\right). \tag{5.28}$$

By combining Eq. (5.28) with the bounds of Eqns. (5.26) and (5.27) it is possible to derive the fully empirical VC-based bounds [192]. Given a space of functions $\mathcal{F}$ and a dataset $\mathcal{D}_n$ it is possible to state with probability $(1-2\delta)$ that

$$\frac{L(f)-\widehat{L}(f)}{\sqrt{L(f)}} \leqslant 6\sqrt{\frac{\widehat{\mathsf{H}}_n(\mathcal{F}) + 2\ln\left(\frac{4}{\delta}\right)}{n}}, \quad \forall f \in \mathcal{F}, \tag{5.29}$$

$$\left|L(f)-\widehat{L}(f)\right| \leqslant 3\sqrt{\frac{\widehat{\mathsf{H}}_n(\mathcal{F}) + 2\ln\left(\frac{4}{\delta}\right)}{n}}, \quad \forall f \in \mathcal{F}. \tag{5.30}$$

Note that the bound of Eq. (5.29) is fully empirical and can show fast rate of convergence $O\left(1/n\right)$ when $\widehat{L}(f) \to 0$, while the bound of Eq. (5.30) always shows slow convergence rate $O\left(\sqrt{1/n}\right)$.
Even if the bounds of Eqns. (5.29) and (5.30) are the only ones that can be used in practice, the most known VC-Bounds are the ones based on the d_{VC}. In order to present them we need to recall the Saurer's Lemmas [228, 243]. Given the $d_{\mathrm{VC}}(\mathcal{F})$ then

$$\mathsf{G}_n(\mathcal{F}) \leqslant \ln\left[\sum_{i=0}^{d_{\mathrm{VC}}(\mathcal{F})} \binom{n}{i}\right] \tag{5.31}$$

If $n < d_{\mathrm{VC}}(\mathcal{F})$, then $\mathsf{G}_n(\mathcal{F}) = \ln\left(2^n\right)$. If $n \geqslant d_{\mathrm{VC}}(\mathcal{F})$, then

$$\mathsf{G}_n(\mathcal{F}) \leqslant \ln\left[\left(\frac{en}{d_{\mathrm{VC}}(\mathcal{F})}\right)^{d_{\mathrm{VC}}(\mathcal{F})}\right]. \tag{5.32}$$

Moreover if $d_{\mathrm{VC}}(\mathcal{F}) > e$, then

$$\mathsf{G}_n(\mathcal{F}) \leqslant \ln\left(n^{d_{\mathrm{VC}}(\mathcal{F})}\right) = d_{\mathrm{VC}}(\mathcal{F})\ln(n). \tag{5.33}$$

By combining Eq. (5.33) (for simplicity) with the bounds of Eqns. (5.26) and (5.27) it is possible to state with probability $(1-\delta)$ that [265]

$$\frac{L(f) - \widehat{L}(f)}{\sqrt{L(f)}} \leqslant 2\sqrt{\frac{d_{\mathrm{VC}}(\mathcal{F}) \ln{(n)} + \ln{\left(\frac{4}{\delta}\right)}}{n}}, \quad \forall f \in \mathcal{F}, \tag{5.34}$$

$$\left| L(f) - \widehat{L}(f) \right| \leqslant \sqrt{\frac{d_{\mathrm{VC}}(\mathcal{F}) \ln{(n)} + \ln{\left(\frac{4}{\delta}\right)}}{n}}, \quad \forall f \in \mathcal{F}. \tag{5.35}$$

The bounds of Eqns. (5.34) and (5.35) are the most well known results of the VC-Theory. We report them just for historical reasons even if not fully empirical.

The problem of the fully empirical bounds of Eqns. (5.29) and (5.30) is that they take into account all the functions inside $\mathcal{F}$. This is not reasonable since, during the learning process, the functions in $\mathcal{F}$ with high empirical error will be never selected by the learning algorithm, hence they should not be taken into account when computing the complexity of $\mathcal{F}$. In order to adress this issue a Local VC-theory need to be exploited [192].

In order to present Local VC-theory we need some preliminary definitions and results.

Let us define the localized version of the set of functions defined in Eq. (5.18) by introducing a constraint on the error, controlled by a parameter r:

$$\widehat{\mathcal{F}}\Big|_{(\mathcal{D}_n,r)} = \left\{ \{f_1, \cdots, f_n\} \,\Big|\, f \in \mathcal{F}, \widehat{L}_n(f) \leqslant r \right\}, \tag{5.36}$$

$$\mathcal{F}\Big|_{(\mathcal{D}_n,r)} = \left\{ \{f_1, \cdots, f_n\} \,\Big|\, f \in \mathcal{F}, L(f) \leqslant r \right\}, \tag{5.37}$$

then, the empirical Local VC-Entropy, and its expected counterpart, can be defined as:

$$\widehat{\mathsf{LH}}_n(\mathcal{F}, r) = \ln\left(\left| \widehat{\mathcal{F}}\Big|_{(\mathcal{D}_n,r)} \right| \right), \tag{5.38}$$

$$\mathsf{LH}_n(\mathcal{F}, r) = \mathbb{E}_{\mathcal{D}_n} \ln\left(\left| \mathcal{F}\Big|_{(\mathcal{D}_n,r)} \right| \right). \tag{5.39}$$

As we will show in the next section, it is possible to derive a fully empirical generalization bound on the true error of a classifier, based on these complexities.

The first result, needed to derive the Local VC-Theory, allows to bound the generalization error of a function based on a property of the entire class. This is a rather technical and complex result which allows to normalize the distance between the true and empirical error of the functions. More formally the result states that it is possible to upper-bound the generalization of an $f \in \mathcal{F}$ as follows [192]

$$\begin{aligned} L(f) \leqslant \min_{K \in (1,\infty)} \frac{K}{K-1} \widehat{L}(f) + \frac{r}{K}, \quad f \in \mathcal{F} \\ \text{s.t.} \quad \sup_{\alpha \in [0,1]} \alpha \sup_{f \in \left\{ f \,\middle|\, f \in \mathcal{F},\ L(f) \leqslant \frac{r}{\alpha} \right\}} \left[L(f) - \widehat{L}_n(f) \right] \leqslant \frac{r}{K}, \quad r > 0. \end{aligned} \tag{5.40}$$

Then by exploiting the results of Eqns. (5.29) and (5.30) we can state with probability at least $(1-\delta)$ that [192]

$$\sup_{f \in \mathcal{F}} \left[L(f) - \widehat{L}(f) \right] \leqslant 6 \sqrt{\frac{\widehat{\mathsf{H}}_n(\mathcal{F}) + 2 \ln\left(\frac{4}{\delta}\right)}{n}} \sup_{f \in \mathcal{F}} \sqrt{L(f)}. \tag{5.41}$$

Moreover with probability at least $(1-\delta)$ we can state that [192]

$$\widehat{L}(f) \leqslant L(f) + 3 \sqrt{\frac{\widehat{\mathsf{H}}_n(\mathcal{F}) + 2 \ln\left(\frac{4}{\delta}\right)}{n}}, \quad \forall f \in \mathcal{F}. \tag{5.42}$$

By exploiting Eq. (5.42) it is possible to state that with probability at least $(1-\delta)$ the subset of the class of functions $\mathcal{F}$ characterized by small generalization error is concentrated around the one with small empirical error [192]

$$\begin{aligned} & \left\{ f \,\middle|\, f \in \mathcal{F}, L(f) \leqslant r \right\} \\ & \subseteq \left\{ f \,\middle|\, f \in \mathcal{F}, \widehat{L}(f) \leqslant r + 3 \sqrt{\frac{\widehat{H}_n\left(\left\{ f \,\middle|\, f \in \mathcal{F}, L(f) \leqslant r \right\} \right) + 2 \ln\left(\frac{4}{\delta}\right)}{n}} \right\}. \end{aligned} \tag{5.43}$$

Finally by combining the results of Eqns. (5.40), (5.41), and (5.43) it is possible to obtain the fully empirical Local VC-Theory based bound [192]

$$\begin{aligned} L(f) \leqslant \min_{K \in (1,\infty)} \quad & \frac{K}{K-1} \widehat{L}(f) + \frac{r}{K}, \quad \forall f \in \mathcal{F} \\ \text{s.t.} \quad & \sup_{\alpha \in [0,1]} 6 \sqrt{\frac{r \alpha \left[\mathsf{T}(r,\alpha) + 2 \ln\left(\frac{9}{\delta}\right) \right]}{n}} \leqslant \frac{r}{K}, \quad r > 0 \\ & \mathsf{T}(r,\alpha) \leqslant \widehat{\mathsf{LH}}_n \left(\mathcal{F}, \frac{r}{\alpha} + 3 \sqrt{\frac{\mathsf{T}(r,\alpha) + 2 \ln\left(\frac{9}{\delta}\right)}{n}} \right), \end{aligned} \tag{5.44}$$

which holds with probability at least $(1-\delta)$. Note that the bound of Eq. (5.44), apart from being fully empirical, takes into account only the functions with small empirical error contrarily to the bounds of Eqns. (5.29) and (5.30).

5.3 Rademacher Complexity Theory

Measuring the performance of a learned model is a key topic in the SLT framework [265], as it allows to get more insights about the behavior of the model and to propose effective learning procedures [30, 63, 140, 147, 161, 184, 265]. While originally coping with asymptotic analysis, performance measurement is approached through recent advances in finite sample analysis, which allow to deal with both theoretical and practical issues and can be effectively exploited in real-world applications [4, 8, 98, 107, 209, 249].

The first data independent measures of complexity, i.e. the Growth Function and the VC-dimension, have been proposed [268], and subsequently refined [208, 239]; however, data dependent alternatives, such as the Rademacher Complexity (G)RC [30, 139, 194], have proved to address the limitations of data independent measures. As GRC is a global measure which contemplates the whole hypothesis space, improvements based on locality principles, namely Local Rademacher Complexity (L)RC bounds [26, 27, 140], have been proposed.

The superiority of LRC over the GRC based bounds is supported by a more deep connection with the learning process itself and their rate of convergence. While RC bounds are characterized by slow convergence [21, 24, 27, 61, 140, 166, 194, 265] $O\left(\sqrt{1/n}\right)$, LRC inequalities feature a fast rate [27] $O\left(1/n\right)$. Nevertheless, some conditions must hold in order to enable fast rates in LRC, for example with kernel classes [70, 136] the eigenvalues of the Gram matrix must decrease exponentially (Theorem 5.2 [27]). Another example is when a bounded loss function is adopted: in this case the hypothesis space must contain a function with generalization error equal to zero (Lemma 6.6 [27]). These conditions are seldom verified in practice (e.g., refer to the discussion following Theorem 5.2[27]): for example, the first hypothesis does not hold when Gaussian kernels are employed. Moreover, LRC bounds have also proved to be loose [194], mostly because of the size of the constants which characterize them [198]. Recently, because of the drawbacks of LRC bounds, some effort has been spent in order to leverage some of the basic ideas, driving LRC, in GRC bounds as well, targeted towards shrinking the hypothesis space and consequently reduce the overall impact of the complexity term [10, 14, 197]. Moreover new tight GRC bounds are derived [196], which exploit the paramount results pursued in [48, 156, 171, 257] in the framework of concentration inequalities: they show that it is possible to achieve a fast convergence rate $O\left(1/n\right)$ in the optimistic case, i.e. when the class is characterized by a complexity tending to zero and it contains a perfect classifier, analogously to [206]. Fast rates are then shown, for the first time, in the case of GRC, even though, in the general scenario, the slow rate $O\left(\sqrt{1/n}\right)$ is still valid [196].

Recent works in literature also showed that the Rademacher Complexity is an effective statistical measure, which can be exploited to analyze the learning performance of a model in learning frameworks other than the supervised one. For example, RC has been exploited in the transductive [92] and semi-supervised [130, 143, 226, 253, 281] learning frameworks. In the latter setting, in particular, previous works showed how the tightness of RC bounds can remarkably benefit from the exploitation of unlabeled samples [14, 34, 128, 196, 198].

In this section we will start by presenting first the GRC-Theory and then we will present the novel LRC-Theory.

In order to present the GRC-Theory let us define some preliminary quantities. GRC end LRC theories, contrarily to the VC-Theory, deal with the general SL framework where a $[0,1]$-bounded loss is exploited. Let us recall the definition of Uniform Deviation (UD) $\widehat{\mathsf{U}}_n(\mathcal{F})$ and RC $\widehat{\mathsf{R}}_n(\mathcal{F})$ [10, 30]:

$$\widehat{\mathsf{U}}_n(\mathcal{F}) = \sup_{f\in\mathcal{F}} \left[L(f) - \widehat{L}(f) \right], \tag{5.45}$$

$$\widehat{\mathsf{R}}_n(\mathcal{F}) = \mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \frac{2}{n} \sum_{i=1}^{n} \sigma_i \ell(f(x_i), y_i), \tag{5.46}$$

where $\sigma_1, \dots, \sigma_n$ are independent uniform $\{\pm 1\}$-valued random Rademacher variables. Note that, when Rademacher variables are allowed to assume only a subset of possible 2^n configurations, i.e. $\sum_{i=1}^{n} \sigma_i = 0$, another complexity measure is obtained, called the Maximal Discrepancy (MD) $\widehat{\mathsf{M}}_n(\mathcal{F})$ [10, 30]:

$$\widehat{\mathsf{M}}_n(\mathcal{F}) = \sup_{f\in\mathcal{F}} \frac{2}{n} \left[\sum_{i=1}^{\frac{n}{2}} \ell(f(x_i), y_i) - \sum_{i=\frac{n}{2}+1}^{n} \ell(f(x_i), y_i) \right]. \tag{5.47}$$

Since $\widehat{\mathsf{U}}_n(\mathcal{F})$, $\widehat{\mathsf{R}}_n(\mathcal{F})$, and $\widehat{\mathsf{M}}_n(\mathcal{F})$ are random quantities, let us define their deterministic counterparts: the Expected UD, the Expected RC, and the Expected Maximal Discrepancy

$$\mathsf{U}_n(\mathcal{F}) = \mathbb{E}_{\mathcal{D}_n} \widehat{\mathsf{U}}_n(\mathcal{F}), \tag{5.48}$$

$$\mathsf{R}_n(\mathcal{F}) = \mathbb{E}_{\mathcal{D}_n} \widehat{\mathsf{R}}_n(\mathcal{F}), \tag{5.49}$$

$$\mathsf{M}_n(\mathcal{F}) = \mathbb{E}_{\mathcal{D}_n} \widehat{\mathsf{M}}_n(\mathcal{F}). \tag{5.50}$$

These quantities allow to obtain an effective upper bound of the unknown $L(f)$ [25, 30, 139, 265] in terms of empirical quantities only. In particular, it is possible to study the difference between $L(f)$ and $\widehat{L}(f)$ through the UD $\widehat{\mathsf{U}}_n(\mathcal{F})$:

$$L(f) \leqslant \widehat{L}(f) + \widehat{\mathsf{U}}_n(\mathcal{F}), \quad \forall f \in \mathcal{F}. \tag{5.51}$$

However, $\widehat{\mathsf{U}}_n(\mathcal{F})$ depends on μ and is not computable as well. We can thus upper bound the UD through RC (or, equivalently, MD):

$$\begin{aligned}
&\mathbb{P}_{\mathcal{D}_n}\left\{\widehat{\mathsf{U}}_n(\mathsf{F}) \geqslant \widehat{\mathsf{C}}_n(\mathcal{F}) + t\right\} && (5.52)\\
&\leqslant \mathbb{P}_{\mathcal{D}_n}\left\{\widehat{\mathsf{U}}_n(\mathcal{F}) \geqslant \mathsf{C}_n(\mathcal{F}) + t_1\right\} + \mathbb{P}_{\mathcal{D}_n}\left\{\mathsf{C}_n(\mathcal{F}) \geqslant \widehat{\mathsf{C}}_n(\mathcal{F}) + t_2\right\} && (5.53)\\
&\leqslant \mathbb{P}_{\mathcal{D}_n}\left\{\widehat{\mathsf{U}}_n(\mathcal{F}) \geqslant \mathsf{U}_n(\mathcal{F}) + t_1\right\} + \mathbb{P}_{\mathcal{D}_n}\left\{\mathsf{C}_n(\mathcal{F}) \geqslant \widehat{\mathsf{C}}_n(\mathcal{F}) + t_2\right\}, \\
&\quad t_1 + t_2 = t, && (5.54)
\end{aligned}$$

where we exploited the following inequality [30]:

$$\mathsf{U}_n(\mathcal{F}) \leqslant \mathsf{C}_n(\mathcal{F}), \tag{5.55}$$

and C can be either R or M. Since $\widehat{\mathsf{U}}_n(\mathcal{F})$ satisfies the hypothesis of McDiarmid inequality [30, 181], it is possible to prove that, with probability $(1-\delta)$:

$$\begin{aligned}
&L(f) \leqslant \widehat{L}(f) + \mathsf{U}_n(\mathcal{F}) + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2n}} \leqslant \widehat{L}(f) + \mathsf{C}_n(\mathcal{F}) + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2n}}, \\
&\quad \forall f \in \mathcal{F}, && (5.56)
\end{aligned}$$

which unfortunately shows a slow convergence rate $O\left(\sqrt{1/n}\right)$. Moreover, it cannot be computed in practice, as it requires the knowledge of μ. By exploiting the fact that $\widehat{\mathsf{C}}_n(\mathcal{F})$ satisfies the McDiarmid inequality [30, 181] too, it is possible to provide a fully empirical bound that holds with probability $(1-\delta)$:

$$L(f) \leqslant \widehat{L}(f) + \widehat{\mathsf{C}}_n(\mathcal{F}) + 3\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2n}}, \quad \forall f \in \mathcal{F}. \tag{5.57}$$

The bound of Eq. (5.57) is characterized by the same slow convergence rate $O\left(\sqrt{1/n}\right)$, and the constants lead to a looser bound than the one of Eq. (5.56). However, both $\widehat{\mathsf{R}}_n(\mathcal{F})$ and $\widehat{\mathsf{M}}_n(\mathcal{F})$ can be easily computed. If $\widehat{\mathsf{R}}_n(\mathcal{F})$ is used, 2^n maximization problems must be solved: approximations through Monte Carlo approaches, rapidly converging to effective solutions, can be exploited

in practice [25, 30]. Instead, if an MD measure $\widehat{\mathsf{M}}_n(\mathcal{F})$ is used, a single minimization problem must be solved, although this process is usually replicated in order to avoid "unlucky" complexity estimations [15]. The advantages and disadvantages of using $\widehat{\mathsf{R}}_n(\mathcal{F})$ or $\widehat{\mathsf{M}}_n(\mathcal{F})$ are investigated in [17]: RC measures $\widehat{\mathsf{R}}_n(\mathcal{F})$ are preferred as they satisfy the self bounding property [46, 196], and consequently we will focus on RC.
As shown in [26, 194, 196], it is also possible to improve the constants in Eq. (5.57) by removing the factor 3 in the best case scenarios (when $\widehat{\mathsf{R}}_n(\mathcal{F}) \to 0$). Thanks to this result it is possible to obtain the explicit sharper version of the bound of Eq. (5.57), that holds with probability $(1-\delta)$:

$$\begin{aligned} L(f) \leqslant \widehat{L}(f) + \widehat{\mathsf{R}}_n(\mathcal{F}) + \frac{2}{n}\sqrt{n \ln\left(\frac{2}{\delta}\right) \widehat{\mathsf{R}}_n(\mathcal{F}) + \left[\ln\left(\frac{2}{\delta}\right)\right]^2} \\ + \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2n}} + \frac{2\ln\left(\frac{2}{\delta}\right)}{n}, \quad \forall f \in \mathcal{F}. \end{aligned} \tag{5.58}$$

Unfortunately, despite its tightness, the previous bound still shows a slow convergence rate $O\left(\sqrt{1/n}\right)$ even in the best case scenario, namely when the class of functions is small $\widehat{\mathsf{R}}_n(\mathcal{F}) \to 0$ and contains a function characterized by $\widehat{L}(f) \to 0$.
In order to improve the constants in the bound of Eq. (5.58), let us define the following quantity:

$$\phi(a) = (1+a)\log(1+a) - a, a > -1, \tag{5.59}$$

$$\widehat{\phi}(a) = 1 - \exp\left[1 + W_{-1}\left(\frac{a-1}{e}\right)\right], \quad \phi\left[-\widehat{\phi}(a)\right] = a, \quad a \in [0,1], \tag{5.60}$$

$$\breve{\phi}(a) = \exp\left[1 + W_0\left(\frac{a-1}{e}\right)\right] - 1, \quad \phi\left[\breve{\phi}(a)\right] = a, \quad a \in [0,+\infty), \tag{5.61}$$

where W_{-1} and W_0 are, respectively, two solutions of the Lambert W function [69]. It is now possible to show that the bound of Eq. (5.58) can be further improved by giving a closed form expression [194], i.e. by formulating an implicit bound which requires a numerical procedure to be solved

$$\begin{aligned} L(f) \leqslant \widehat{L}(f) + r^* + \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2n}}, \quad \forall f \in \mathcal{F}, \\ r^* = \arg\max_{r \in [0,1]} r \quad \text{s.t.} \quad r = \widehat{\mathsf{R}}_n(\mathcal{F}) + r\widehat{\phi}\left[\frac{2\ln\left(\frac{2}{\delta}\right)}{nr}\right], \end{aligned} \tag{5.62}$$

which holds with probability $(1-\delta)$, but still shows a slow convergence rate.

In order to derive and improved version of the bounds of Eqns. (5.58) and (5.62) we need to define another quantity, together with its empirical counterpart

$$\mathsf{V}(\mathcal{F}) = \sup_{f \in \mathcal{F}} \mathbb{E}_{(x,y)} \ell(f(x), y), \tag{5.63}$$

$$\widehat{\mathsf{V}}_n(\mathcal{F}) = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i). \tag{5.64}$$

From a statistical point of view, $\mathsf{V}(\mathcal{F})$ is an upper bound of the variance of $\ell(f(x), y)$ and is not computable: we can only derive its empirical estimate $\widehat{\mathsf{V}}_n(\mathcal{F})$. From a cognitive point of view, $\mathsf{V}(\mathcal{F})$ measures the ability of our hypothesis space $\mathcal{H}$ of "non-learning" μ: in other words, $\mathsf{V}(\mathcal{F})$ evaluates the worst case scenario which the procedure could have to cope with during the learning process.

At this point, since $\widehat{\mathsf{U}}_n(\mathcal{F})$ satisfies the hypothesis of the Bousquet inequality [48] and $\widehat{\mathsf{R}}_n(\mathcal{F})$ is a self bounding function [46, 196], it is possible to improve the explicit bound of Eq. (5.58), by stating that with probability $(1-\delta)$ the following bound holds [196]

$$\begin{aligned} L(f) &\leqslant \widehat{L}(f) + \widehat{\mathsf{R}}_n(\mathcal{F}) + 5\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{n}\widehat{\mathsf{R}}_n(\mathcal{F})} + 2\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{n}\widehat{\mathsf{V}}_n(\mathcal{F})} \\ &\quad + 4\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{n}\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{n}\widehat{\mathsf{R}}_n(\mathcal{F})}} + \frac{11\ln\left(\frac{2}{\delta}\right)}{n} \qquad (5.65) \\ &\leqslant \widehat{L}(f) + 3\widehat{\mathsf{R}}_n(\mathcal{F}) + \widehat{\mathsf{V}}_n(\mathcal{F}) + \frac{24\ln\left(\frac{2}{\delta}\right)}{n}. \qquad (5.66) \end{aligned}$$

The bound of Eq. (5.65) contains only empirical quantities. By analyzing the bound of Eq. (5.65), it is worth noting that the convergence rate can vary between a slow $O\left(\sqrt{1/n}\right)$ and a fast $O\left(1/n\right)$ value. As the complexity terms differ from zero, the slow convergence prevails; on the contrary, the fast rate is achievable when all the complexity terms tend to zero, as it can be more easily noted in the formulation of Eq. (5.66). Thus, the previous bound is characterized (though in a non-typical optimistic scenario) by fast convergence: this is the first time for GRC measures, since they only showed slow convergence so far [26, 30, 194]. A similar result can be derived when the distribution of the data depends on n. Let us consider $\mathcal{Y} \in \{\pm 1\}$ and $\mathbb{P}\{Y = +1 | X\} = 1/n$, and let also $\mathcal{H}$ include models that assign labels $+1$ to every sample: in this scenario, we have that $\widehat{\mathsf{V}}_n(\mathcal{F}) \approx 1/n$ and $\widehat{\mathsf{R}}_n(\mathcal{F}) \to 0$, thus we derive fast rates as well. One can also resort to approaches allowing to reduce the complexity of the space: for example, by reserving part of the data for building a data dependent hypothesis space, as described in [9, 12, 14], or by adding some hypothesis, like the low noise condition [24, 248]. These results show that the complexity of the class of function can be sharpened without resorting to LRC measures.

Despite being appealing in terms of convergence rate, the constants included in the bound of Eq. (5.65) are not optimal: we have to give up the closed form formulation in order to circumvent this issue.

Analogously to the bound of Eq. (5.62), for the bound of Eq. (5.58) the implicit form with optimal constants of the bounds of Eq. (5.65), which holds with probability $(1-\delta)$, is the following one [196]

$$L(f) \leqslant \widehat{L}(f) + s_1^*, \quad \forall f \in \mathcal{F}, \tag{5.67}$$

$$s_1^* = \arg \max_{s_1 \in [0,1]} \ s_1$$

$$\text{s.t.} \quad s_1 = r_1^* + \left(2r_1^* + \widehat{\mathsf{V}}_n(\mathcal{F}) + s_1\right) \breve{\phi}\left(\frac{\ln\left(\frac{2}{\delta}\right)}{n\left(2r_1^* + \widehat{\mathsf{V}}_n(\mathcal{F}) + s_1\right)}\right),$$

$$r_1^* = \arg \max_{r_1 \in [0,1]} \ r_1$$

$$\text{s.t.} \quad r_1 = \widehat{\mathsf{R}}_n(\mathcal{F}) + r_1 \widehat{\phi}\left(\frac{2\ln\left(\frac{2}{\delta}\right)}{n r_1}\right).$$

The problem that still affects the fully empirical bounds of Eqns. (5.58), (5.62), (5.65), and (5.67) is that they take into account all the functions inside $\mathcal{F}$. This is not reasonable since, during the learning process, the functions in $\mathcal{F}$ with high empirical error will be never selected by the learning algorithm, hence they should not be taken into account when computing the complexity of $\mathcal{F}$ by applying the localization principle. In order to address this issue the LRC-Theory need to be exploited [27, 198].

In order to present Local LRC-Theory we need some preliminary definitions and results.

First, we switch from the space of functions $\mathcal{F}$ to the space of loss functions. Given a space of functions $\mathcal{F}$ with its associated loss function $\ell(f(x), y)$, the space of loss functions $\mathcal{L}$ is defined as

$$\mathcal{L} = \left\{ \ell(f(x), y) \,\middle|\, f \in \mathcal{F} \right\}. \tag{5.68}$$

Let us also consider the corresponding star-shaped space of functions [27, 198]. Given the space of loss functions $\mathcal{L}$, its star-shaped version is

$$\mathcal{L}^s = \left\{ \alpha\ell \,\middle|\, \alpha \in [0,1],\ \ell \in \mathcal{L} \right\}. \tag{5.69}$$

Then, the generalization error and the empirical error can be rewritten in terms of the space of loss functions

$$L(f) \equiv L(\ell), \quad \widehat{L}(f) \equiv \widehat{L}(\ell). \tag{5.70}$$

Moreover we can define, respectively, the expected square error and the empirical square error:

$$L(\ell^2) = \mathbb{E}_{x,y}\left[\ell(f(x),y)\right]^2, \tag{5.71}$$

$$\widehat{L}(\ell^2) = \frac{1}{n}\sum_{i=1}^{n}\left[\ell\left(f\left(x_i\right),y_i\right)\right]^2. \tag{5.72}$$

Since we do not know in advance which $f \in \mathcal{F}$ will be chosen during the learning phase, in order to estimate $L(\ell)$ we have to study the behavior of the difference between the generalization error and the empirical error. Given $\mathcal{L}$, the UD of the loss $\widehat{\mathsf{U}}_n(\mathcal{L})$ and square loss $\widehat{\mathsf{U}}_n^2(\mathcal{L})$ are

$$\widehat{\mathsf{U}}_n(\mathcal{L}) = \sup_{\ell\in\mathcal{L}}\left[L(\ell) - \widehat{L}(\ell)\right], \tag{5.73}$$

$$\widehat{\mathsf{U}}_n^2(\mathcal{L}) = \sup_{\ell\in\mathcal{L}}\left[\widehat{L}(\ell^2) - L(\ell^2)\right], \tag{5.74}$$

while their deterministic counterparts are:

$$\mathsf{U}_n(\mathcal{L}) = \mathbb{E}_{\mathcal{D}_n}\widehat{\mathsf{U}}_n(\mathcal{L}), \tag{5.75}$$

$$\mathsf{U}_n^2(\mathcal{L}) = \mathbb{E}_{\mathcal{D}_n}\widehat{\mathsf{U}}_n^2(\mathcal{L}). \tag{5.76}$$

The UD is not computable, but we can upper bound its value through some computable quantities. One possibility is to use the RC. The RC of the loss and of the square loss are:

$$\widehat{\mathsf{R}}_n(\mathcal{L}) = \mathbb{E}_{\boldsymbol{\sigma}}\sup_{\ell\in\mathcal{L}}\frac{2}{n}\sum_{i=1}^{n}\sigma_i\ell\left(f\left(x_i\right),y_i\right), \tag{5.77}$$

$$\widehat{\mathsf{R}}_n^2(\mathcal{L}) = \mathbb{E}_{\boldsymbol{\sigma}}\sup_{\ell\in\mathcal{L}}\frac{2}{n}\sum_{i=1}^{n}\sigma_i\left[\ell\left(f\left(x_i\right),y_i\right)\right]^2. \tag{5.78}$$

Their deterministic counterparts are:

$$\mathsf{R}_n(\mathcal{L}) = \mathbb{E}_{\mathcal{D}_n}\widehat{\mathsf{R}}_n(\mathcal{L}), \tag{5.79}$$

$$\mathsf{R}_n^2(\mathcal{L}) = \mathbb{E}_{\mathcal{D}_n}\widehat{\mathsf{R}}_n^2(\mathcal{L}). \tag{5.80}$$

Finally, we will also make use of the notion of sub-root function [27, 198]. A function is a sub-root function if and only if: (I) $\psi(r)$ is positive, (II) $\psi(r)$ is non-decreasing, and (III) ${\psi(r)}/{\sqrt{r}}$ is non-increasing with $r > 0$. The properties of the sub-root functions are reported in many works [27, 198].

In this first part, we propose a proof of the LRC bound on the generalization error of a model [27, 140], which is simplified with respect to the original proof in literature and allows us also to obtain optimal constants [198]. Later we will report the state-of-the-art bound.

In order to improve the readability of this part, an outline of the main steps of the the presentation is reported. As a first step, we will show that it is possible to bound the generalization error of a function chosen in $\mathcal{F}$, through an assumption over the Expected UD of a normalized and slightly enlarged version of $\mathcal{F}$. As a second step, we will show how to relate the Expected UD and the Expected RC through the use of a sub-root function. The fixed point of this sub-root function is used to bound the generalization error of a function chosen in $\mathcal{F}$. As a third step, we will show that, instead of using any sub-root function, we can directly use the Expected RC of a local space of functions, where functions therein are the ones with low expected square error. As a fourth step, we will substitute the non-computable expected quantities mentioned above with their empirical counterpart, which can be computed from the data. Then, we finally derive the main result, which is a fully empirical LRC bound on the generalization error of a function chosen in the original hypothesis space $\mathcal{F}$.

The following result is needed for normalizing the original hypothesis space: this allows to bound the generalization error of a function chosen in $\mathcal{F}$. Let us consider the normalized loss space $\mathcal{L}_r$:

$$\mathcal{L}_r = \left\{ \frac{r}{L(\ell^2) \vee r} \ell \,\middle|\, \ell \in \mathcal{L} \right\}, \tag{5.81}$$

and let us suppose that, $\forall K > 1$:

$$\widehat{\mathsf{U}}_n(\mathcal{L}_r) \leqslant \frac{r}{K}. \tag{5.82}$$

Then, $\forall f \in \mathcal{F}$, the following inequality holds [198]

$$L(f) \leqslant \max\left\{ \left(\frac{K}{K-1} \widehat{L}(f) \right), \left(\widehat{L}(f) + \frac{r}{K} \right) \right\} \leqslant \frac{K}{K-1} \widehat{L}(f) + \frac{r}{K}. \tag{5.83}$$

The next step shows that the normalized hypothesis space defined in Eq. (5.81) is a subset of a new star-shaped space. Let

$$\mathcal{L}_r^s = \left\{ \alpha\ell \,\middle|\, \alpha \in [0,1],\ \ell \in \mathcal{L},\ L[(\alpha\ell)^2] \leqslant r \right\}, \tag{5.84}$$

then [198]

$$\mathcal{L}_r \subseteq \mathcal{L}_r^s. \tag{5.85}$$

If we consider a sub-root function that upper-bounds the Expected RC of the hypothesis space defined in Eq. (5.84), we can exploit the bound of Eq. (5.83) for bounding the generalization error of a function chosen in the original hypothesis space $\mathcal{F}$. Let us consider a sub-root function $\psi_n(r)$, with fixed point r_n^*, and suppose that, $\forall r > r_n^*$

$$\mathsf{R}_n\left(\mathcal{L}_r^s\right) \leqslant \psi_n(r). \tag{5.86}$$

Then, $\forall f \in \mathcal{F}$ and $\forall K > 1$ we have that [198], with probability $(1-\delta)$:

$$L(f) \leqslant \max\left\{ \left(\frac{K}{K-1} \widehat{L}(f) \right), \left(\widehat{L}(f) + K r_n^* + 2\sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2n}} \right) \right\}. \tag{5.87}$$

The previous result holds for any sub-root function which satisfies Eq. (5.86). The next lemma shows that the RC, defined in Eq. (5.86), is indeed a sub-root function, which means that the inequality of Eq. (5.86) is indeed an equality. Let us consider $\mathsf{R}_n\left(\mathcal{L}_r^s\right)$, namely the Expected RC computed on $\mathcal{L}_r^s$. Then

$$\psi_n(r) = \mathsf{R}_n\left(\mathcal{L}_r^s\right) \tag{5.88}$$

is a sub-root function [27, 198].

The next two results allow to substitute the non-computable expected quantities, $\mathcal{L}_r^s$ and R_n, with their empirical counterparts, which can be computed from the data. Let us suppose that

$$r \geqslant \mathsf{R}_n\left(\mathcal{L}_r^s\right), \tag{5.89}$$

and let us define

$$\widehat{\mathcal{L}}_r^s = \left\{ \alpha\ell \,\middle|\, \alpha \in [0,1],\ \ell \in \mathcal{L},\ \widehat{L}[(\alpha\ell)^2] \leqslant \left(3r + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2n}} \right) \right\}. \tag{5.90}$$

Then, $\forall \ell_s^r \in \mathcal{L}_r^s$, the following inequality holds [198] with probability $(1-\delta)$:

$$\mathcal{L}_r^s = \widehat{\mathcal{L}}_r^s. \tag{5.91}$$

Moreover let us consider two sub-root functions and their fixed points:

$$\psi_n(r) = \mathsf{R}_n\left(\mathcal{L}_r^s\right), \quad \psi_n(r_n^*) = r_n^* \tag{5.92}$$

$$\widehat{\psi}_n(r) = \widehat{\mathsf{R}}_n\left(\widehat{\mathcal{L}}_r^s\right) + \sqrt{\frac{2x}{n}}, \quad \psi_n(\widehat{r}_n^*) = \widehat{r}_n^*. \tag{5.93}$$

The following inequalities hold [198] with probability $(1 - 2\delta)$:

$$\psi_n(r) \leqslant \widehat{\psi}_n(r), \quad r_n^* \leqslant \widehat{r}_n^*. \tag{5.94}$$

Finally, we derive the main result of this part, namely a fully empirical LRC bound on the generalization error of a function, chosen in the original hypothesis space $\mathcal{F}$. Let us consider a space of functions $\mathcal{F}$ and the fixed point $\widehat{r}_n^*$ of the following sub-root function:

$$\widehat{\psi}_n(r) = \widehat{\mathsf{R}}_n\left(\widehat{\mathcal{L}}_r^s\right) + \sqrt{\frac{2\ln\left(\frac{1}{\delta}\right)}{n}}, \tag{5.95}$$

where

$$\widehat{\mathcal{L}}_r^s = \left\{ \alpha\ell \,\middle|\, \alpha \in [0,1],\ \ell \in \mathcal{L},\ \widehat{L}(\ell^2) \leqslant \frac{1}{\alpha^2}\left(3r + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2n}}\right)\right\}. \tag{5.96}$$

Then, $\forall f \in \mathcal{F}$ and $\forall K > 1$ the following inequality holds [198] with probability $(1 - 3\delta)$:

$$L(f) \leqslant \max\left\{\left(\frac{K}{K-1}\widehat{L}(f)\right), \left(\widehat{L}(f) + K\widehat{r}_n^* + 2\sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2n}}\right)\right\}. \tag{5.97}$$

The bound of Eq. 5.97 is mainly based on the exploitation of McDiarmid's inequalities [181]. In order to improve the tightness of the bound of Eq. (5.97), more refined concentration inequalities [46–48], based on the milestone results of Talagrand [156], need to be exploited. This approach improves the technique proposed by [27] and obtains optimal constants for the bounds [198] by giving up the closed form solution [198].

Then, we exploit the more refined concentration [46–48] inequalities in the results of Eqns. (5.91) and (5.94). By combining these different pieces, the desired bound can be derived.

The first step is to obtain the counterpart of the bound of Eq. (5.87). Let us consider a sub-root function $\psi_n(r)$ and its fixed point r_n^*, and suppose that $\forall r > r_n^*$:

$$\mathsf{R}_n\left(\mathcal{L}_r^s\right) \leqslant \psi_n(r). \tag{5.98}$$

Let us define r^{U} as the largest solution, with respect to r, of the following equation:

$$\sqrt{rr_n^*} + \left[2\sqrt{rr_n^*} + r\right]\breve{\phi}\left\{\frac{\ln\left(\frac{1}{\delta}\right)}{\left[n\left(2\sqrt{rr_n^*} + r\right)\right]}\right\} = \frac{r}{K}. \tag{5.99}$$

Then $\forall f \in \mathcal{F}$ and $\forall K > 1$ we have that [198], with probability $(1-\delta)$:

$$L(f) \leqslant \max\left\{\left(\frac{K}{K-1}\widehat{L}(f)\right), \left(\widehat{L}(f) + \frac{r^{\mathsf{U}}}{K}\right)\right\}. \tag{5.100}$$

The next two results are the counterparts of Eqns. (5.91) and (5.94). Let us suppose that:

$$r \geqslant \mathsf{R}_n\left(\mathcal{L}_r^s\right). \tag{5.101}$$

Let us define $\widehat{\mathcal{L}}_r^s$ as:

$$\widehat{\mathcal{L}}_r^s = \left\{\alpha\ell \,\middle|\, \alpha \in [0,1],\ \ell \in \mathcal{L},\ \widehat{L}[(\alpha\ell)^2] \leqslant 3r + 5r\breve{\phi}\left(\frac{\ln\left(\frac{1}{\delta}\right)}{n5r}\right)\right\}. \tag{5.102}$$

Then [198] $\forall \ell_s^r \in \mathcal{L}_r^s$ and with probability $(1-\delta)$:

$$\mathcal{L}_r^s \subseteq \widehat{\mathcal{L}}_r^s. \tag{5.103}$$

Moreover let us consider two sub-root functions and their fixed points:

$$\psi_n(r) = \mathsf{R}_n\left(\mathcal{L}_r^s\right), \quad \psi_n(r_n^*) = r_n^* \tag{5.104}$$

$$\widehat{\psi}_n(r) = \widehat{\mathsf{R}}_n(\widehat{\mathcal{L}}_r^s) + r\widehat{\phi}\left(\frac{2\ln\left(\frac{1}{\delta}\right)}{nr}\right), \quad \widehat{\psi}_n(\widehat{r}_n^*) = \widehat{r}_n^*. \tag{5.105}$$

Then, the following inequalities hold [198] with probability $(1-2\delta)$:

$$\psi_n(r) \leqslant \widehat{\psi}_n(r), \quad r_n^* \leqslant \widehat{r}_n^*. \tag{5.106}$$

Finally, we can derive the fully empirical and tighter version of the bound of Eq. (5.97) LRC state of the art bound [198]. Let us consider a space of functions $\mathcal{F}$ and the fixed point $\widehat{r}_n^*$ of the following sub-root function:

$$\widehat{\psi}_n(r) = \widehat{\mathsf{R}}_n\left(\widehat{\mathcal{L}}_r^s\right) + r\widehat{\phi}\left(\frac{2\ln\left(\frac{1}{\delta}\right)}{nr}\right), \tag{5.107}$$

where

$$\widehat{\mathcal{L}}_r^s = \left\{ \alpha\ell \,\middle|\, \alpha{\in}[0,1],\ \ell \in \mathcal{L},\ \widehat{L}[(\alpha\ell)^2] \leqslant 3r + 5r\breve{\phi}\left(\frac{\ln\left(\frac{1}{\delta}\right)}{n5r}\right).\right\} \tag{5.108}$$

Let us define r^{U} as the largest solution of the following identity:

$$\sqrt{r\widehat{r}_n^*} + \left[2\sqrt{r\widehat{r}_n^*} + r\right]\breve{\phi}\left\{\frac{\ln\left(\frac{1}{\delta}\right)}{\left[n\left(2\sqrt{r\widehat{r}_n^*} + r\right)\right]}\right\} = \frac{r}{K}. \tag{5.109}$$

Then, $\forall f \in \mathcal{F}$ and $\forall K > 1$, the following inequality holds [198] with probability $(1 - 3\delta)$:

$$L(f) \leqslant \max\left\{\left(\frac{K}{K-1}\widehat{L}(f)\right), \left(\widehat{L}(f) + \frac{r^{\mathsf{U}}}{K}\right)\right\}. \tag{5.110}$$

Note that both the bounds of Eqns. (5.97) and (5.110) are in implicit form. However, the bound of Eq. (5.110) requires to look for a fixed point and to find the largest solution of an equation; the bound of Eq. (5.97), instead, only requires to find a fixed point.

6

Compression Bound

Compression bound is probably the simplest yet theoretically grounded approach to MS and EE. The Compression bound [96, 147, 162] relies on a simple idea: if an algorithm is able to compress the data provided to learn a rule then the algorithm will generalize. This idea has a long history in the literature from the seminal groundbreaking work about the Kolmogorov Complexity [184] to the Minimum Description Length principle [109]. The fact that compressing is related to learning is nowadays a well known and firm principle which guides many researchers in the design of new learning schema.

The compression bound applies to deterministic algorithms and both deterministic and probabilistic rules. Just the i.i.d. hypothesis over the sampled data is necessary. Mainstream learning algorithms do not optimize data compression metric, so the compression bound is seldom used. Nonetheless, there do exist some reasonably competitive learning algorithms (e.g. Support Vector Machines [71, 265]) for which the sample compression bound produces significant results.

In order to present the compression bound more formally let us introduce the notation. Let $\mathcal{X}$ and $\mathcal{Y}$ be, respectively, an input and an output space. We consider a set of labeled independent and identically distributed (i.i.d.) data $\mathcal{D}_n : \{z_1, \cdots, z_n\}$ of size n, where $z_{i \in \{1,\cdots,n\}} = (x_i, y_i)$, sampled from an unknown distribution μ where $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $z \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. A learning algorithm $\mathscr{A}$ maps $\mathcal{D}_n$ into a rule $\mathfrak{R} : \mathscr{A}(\mathcal{D}_n)$, which maps elements in $\mathcal{X}$ to elements in $\mathcal{Y}$. In particular, $\mathscr{A}$ allows designing $\mathfrak{R} \in \mathcal{R}$ and defining the set of rules $\mathcal{R}$, that is generally unknown.

We suppose here that the choice of the final rule $\mathfrak{R}$ does not depend on the entire dataset $\mathcal{D}_n$ but just on a subset of it $\mathcal{D}'_{n'} \subset \mathcal{D}_n$. In other words we assume that the algorithm is able to intrinsically compress the data. Formally we assume that $\mathfrak{R} = \mathscr{A}(\mathcal{D}_n) = \mathscr{A}(\mathcal{D}'_{n'})$.

The accuracy of $\mathfrak{R}$ in representing the hidden relationship μ is measured with reference to a $[0, 1]$-bounded loss function $\ell : \mathcal{R} \times \mathcal{Z} \to [0, 1]$. Consequently, the quantity of interest is defined as the generalization error, namely the error that a model will perform on new data generated by μ and previously unseen

$$L(\mathfrak{R}) = \mathbb{E}_z \ell(\mathfrak{R}, z) . \tag{6.1}$$

$L(\mathfrak{R})$ cannot be computed since μ is unknown and, consequently, must be estimated. The empirical error [10, 26, 30, 266], its empirical estimator, can be computed

$$\widehat{L}(\mathfrak{R}, \mathcal{D}_n) = \frac{1}{n} \sum_{z \in \mathcal{D}_n} \ell(\mathfrak{R}, z) , \tag{6.2}$$

together with the empirical variance

$$\widehat{V}(\mathfrak{R}, \mathcal{D}_n) = \frac{1}{n(n-1)} \sum_{z' \in \mathcal{D}_n} \sum_{z'' \in \mathcal{D}_n} [\ell(\mathfrak{R}, z') - \ell(\mathfrak{R}, z'')]^2. \tag{6.3}$$

Deriving the Compression bound is rather simple because we just need to count all the possible choices that the algorithm $\mathscr{A}$ made in order to define $\mathfrak{R} = \mathscr{A}(\mathcal{D}_n)$ based on just $\mathcal{D}'_{n'}$. Then, since, by definition, some data $\mathcal{D}_n \backslash \mathcal{D}'_{n'}$ have not been exploited by the algorithm during the training phase we can use them as an hold out set. Therefore, analogously to the resampling method, we can use the i.i.d. data in $\mathcal{D}_n \backslash \mathcal{D}'_{n'}$ together with the union bound [45] (the Bonferroni correction) over all the choices made by the algorithm in order to bound the generalization error.

Hence let us count how many choices have been made by the algorithm. Since the $\mathfrak{R} = \mathscr{A}(\mathcal{D}_n)$ depends on $\mathcal{D}'_{n'} \subset \mathcal{D}_n$,

1. the algorithm has chosen the cardinality of $\mathcal{D}'_{n'}$ namely $n' \in \{0, 1, \cdots, n\}$;
2. the algorithm has chosen a subset of samples of cardinality n' in $\mathcal{D}_n$. But how many choices have the algorithm had for selecting a subset a cardinality n' in $\mathcal{D}_n$? The answer is rather simple and it is $\binom{n}{n'}$.

Then we can state that the algorithm $\mathscr{A}$ made $(n+1)\binom{n}{n'}$ choices.

At this point we have to recall that the samples in $\mathcal{D}_n$ are sampled i.i.d. from μ. Then since $\mathfrak{R} = \mathscr{A}(\mathcal{D}_n) = \mathscr{A}(\mathcal{D}'_{n'})$ we have that the errors that $\mathfrak{R}$ makes on $\mathcal{D}_n \backslash \mathcal{D}'_{n'}$ are i.i.d. and then we can use the Hoeffding Inequality [118] together with the Bonferroni over the whole choices made by $\mathscr{A}$ in order to state that:

$$\mathbb{P}_{\mathcal{D}_n \backslash \mathcal{D}'_{n'}} \left\{ L(\mathscr{A}(\mathcal{D}_n)) \geqslant \widehat{L}(\mathscr{A}(\mathcal{D}_n), \mathcal{D}_n \backslash \mathcal{D}'_{n'}) + t \right\}$$
$$\leqslant (n+1) \binom{n}{n'} e^{-2(n-n')t^2}, \tag{6.4}$$
$$\mathbb{P}_{\mathcal{D}_n \backslash \mathcal{D}'_{n'}} \left\{ \left| L(\mathscr{A}(\mathcal{D}_n)) - \widehat{L}(\mathscr{A}(\mathcal{D}_n), \mathcal{D}_n \backslash \mathcal{D}'_{n'}) \right| \geqslant t \right\}$$
$$\leqslant 2(n+1) \binom{n}{n'} e^{-2(n-n')t^2}, \tag{6.5}$$

or, alternatively, that with probability $(1-\delta)$

$$L(\mathscr{A}(\mathcal{D}_n))$$
$$\leqslant \widehat{L}(\mathscr{A}(\mathcal{D}_n), \mathcal{D}_n \backslash \mathcal{D}'_{n'}) + \sqrt{\frac{\ln\left((n+1)\binom{n}{n'}\right)}{2(n-n')}} + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2(n-n')}}, \tag{6.6}$$
$$\left| L(\mathscr{A}(\mathcal{D}_n)) - \widehat{L}(\mathscr{A}(\mathcal{D}_n), \mathcal{D}_n \backslash \mathcal{D}'_{n'}) \right|$$
$$\leqslant \sqrt{\frac{\ln\left((n+1)\binom{n}{n'}\right)}{2(n-n')}} + \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2(n-n')}}. \tag{6.7}$$

Note that from these bounds it is rather clear that they lead to meaningful and useful results just when n' is rather small, otherwise $\binom{n}{n'}$ becomes soon huge. This is the principal limitation of the compression bound: the need for huge rates of compression.
Obviously it is possible to retrieve the Chernoff-type [65] version of the bound of Eq. (6.7) which is sharper when the empirical error is small and it can exhibit a fast convergence rate $O\left({}^{1}\!/\!{}_{n\,-\,n'}\right)$

$$\left| L(\mathscr{A}(\mathcal{D}_n)) - \widehat{L}(\mathscr{A}(\mathcal{D}_n), \mathcal{D}_n \backslash \mathcal{D}'_{n'}) \right| \tag{6.8}$$
$$\leqslant \sqrt{\widehat{L}(\mathscr{A}(\mathcal{D}_n), \mathcal{D}_n \backslash \mathcal{D}'_{n'}) \frac{3\ln\left(\frac{2}{\delta}(n+1)\binom{n}{n'}\right)}{n-n'}} + \frac{3\ln\left(\frac{2}{\delta}(n+1)\binom{n}{n'}\right)}{n-n'},$$

where the bound holds with probability $(1-\delta)$.
It is also possible to derive a Bennet-type bound [35, 175] which is sharper when the variance of the empirical error is small and, as the Chernoff-type bound, it can exhibit a fast convergence rate $O\left({}^{1}\!/\!{}_{n\,-\,n'}\right)$

$$\left| L(\mathscr{A}(\mathcal{D}_n)) - \widehat{L}(\mathscr{A}(\mathcal{D}_n), \mathcal{D}_n \backslash \mathcal{D}'_{n'}) \right| \tag{6.9}$$
$$\leqslant \sqrt{\widehat{V}(\mathscr{A}(\mathcal{D}_n), \mathcal{D}_n \backslash \mathcal{D}'_{n'}) \frac{2\ln\left(\frac{3}{\delta}(n+1)\binom{n}{n'}\right)}{n-n'}} + \frac{7\ln\left(\frac{3}{\delta}(n+1)\binom{n}{n'}\right)}{3(n-n'-1)}, \tag{6.10}$$

where the bound holds with probability $(1-\delta)$.

The state-of-art option is to use the Clopper-Pearson bound [67] extended [59, 193] in order to be applied to the case of $[0,1]$-bounded losses. Let u be a random variable uniformly distributed over $[0,1]$ and let $\{u_1, \cdots, u_n\}$ be n variables sampled i.i.d. from u. Then we can state that

$$\begin{aligned}
&L(\mathscr{A}(\mathcal{D}_n)) && (6.11)\\
&\in \left[\begin{array}{l} \mathtt{Q}\left[\frac{\delta}{2(n+1)\binom{n}{n'}}; \begin{array}{l} (n-n')\widehat{L}^u(\mathscr{A}(\mathcal{D}_n), \mathcal{D}_n \backslash \mathcal{D}'_{n'}), \\ (n-n')-(n-n')\widehat{L}^u(\mathscr{A}(\mathcal{D}_n), \mathcal{D}_n \backslash \mathcal{D}'_{n'})+1 \end{array} \right], \\ \mathtt{Q}\left[1-\frac{\delta}{2(n+1)\binom{n}{n'}}; \begin{array}{l} (n-n')\widehat{L}^u(\mathscr{A}(\mathcal{D}_n), \mathcal{D}_n \backslash \mathcal{D}'_{n'})+1, \\ (n-n')-(n-n')\widehat{L}^u(\mathscr{A}(\mathcal{D}_n), \mathcal{D}_n \backslash \mathcal{D}'_{n'}) \end{array} \right] \end{array} \right],
\end{aligned}$$

with probability $(1-\delta)$ and where $\mathtt{Q}[p; v, w]$ is the p-th quantile of the Beta distribution with shape parameters v and w and $\widehat{L}^u(\mathscr{A}(\mathcal{D}_n), \mathcal{D}_n \backslash \mathcal{D}'_{n'}) = \frac{1}{n-n'} \sum_{\boldsymbol{z} \in \mathcal{D}_n \backslash \mathcal{D}'_{n'}} [\ell(\mathscr{A}(\mathcal{D}_n)) \geqslant u_i]$ (each u_i is associated with a different $z \in \mathcal{D}_n \backslash \mathcal{D}'_{n'}$ and the the Iverson bracket notation [121] is exploited).

Finally note that all the compression bounds are fully empirical bounds and can be used both for MS and EE purposes by exploiting the approach described in the preliminaries of this book [147]. In fact all the Compression-based bounds have the following form:

$$\mathbb{P}_{\mathcal{D}_n \backslash \mathcal{D}'_n} \left\{ L(\mathscr{A}(\mathcal{D}_n)) \leqslant \Delta(\mathcal{D}_n, \mathcal{D}_n \backslash \mathcal{D}'_{n'}, \mathscr{A}, n', \delta) \right\} \geqslant 1-\delta. \qquad (6.12)$$

Then if we want to choose $\mathscr{A}^* \in \{\mathscr{A}_1, \cdots, \mathscr{A}_{n_{\mathscr{A}}}\}$, namely perform the MS phase, and estimate the generalization performance of $\mathscr{A}^*(\mathcal{D}_n)$, namely perform the EE phase, we have to follow the procedure summarized in Algorithm 3. Note that the generalization of the final model is bounded by

$$\begin{aligned}
L(\mathscr{A}^*(\mathcal{D}_n)) \leqslant & \Delta\left(\mathcal{D}_n, \mathcal{D}_n \backslash \mathcal{D}'_{n'}(\mathscr{A}^*), \mathscr{A}^*, n'(\mathscr{A}^*), \frac{\delta}{n_{\mathscr{A}}} \right), \\
& \forall \mathscr{A}^* \in \{\mathscr{A}_1, \cdots, \mathscr{A}_{n_{\mathscr{A}}}\}, \qquad (6.13)
\end{aligned}$$

with probability $(1-\delta)$, since we have applied the Bonferroni correction [45] over the $n_{\mathscr{A}}$ choices for the algorithm. Note that $\mathcal{D}'_{n'}$ and n' depend on the particular $\mathscr{A}$.

Algorithm 3: Compression bound: MS and EE Strategy.

```
Input: $\{\mathscr{A}_1, \cdots, \mathscr{A}_{n_{\mathscr{A}}}\}$, $\mathcal{D}_n$, and $\delta$
Output: Optimal Model $\mathscr{A}^*(\mathcal{D}_n)$ and its estimated generalization error
        $L(\mathscr{A}^*(\mathcal{D}_n))$
$L^*_{\mathrm{MS}} = +\infty$;
for $\mathscr{A} \in \{\mathscr{A}_1, \cdots, \mathscr{A}_{n_{\mathscr{A}}}\}$ do
    Compute $\mathcal{D}'_{n'}$ and $n'$ based on $\mathscr{A}(\mathcal{D}_n)$;
    $L_{\mathrm{MS}} = \Delta(\mathcal{D}_n, \mathcal{D}_n \backslash \mathcal{D}'_{n'}, \mathscr{A}, n', \delta)$;
    if $L^*_{MS} > L_{MS}$ then
        $L^*_{\mathrm{MS}} = L_{\mathrm{MS}}$;
        $\mathscr{A}^*(\mathcal{D}_n) = \mathscr{A}(\mathcal{D}_n)$;
        $L(\mathscr{A}^*(\mathcal{D}_n)) = \Delta\left(\mathcal{D}_n, \mathcal{D}_n \backslash \mathcal{D}'_{n'}, \mathscr{A}, n', \frac{\delta}{n_{\mathscr{A}}}\right)$;
```

7

Algorithmic Stability Theory

The notion of Stability [49, 186, 212] allows to answer a fundamental question in learning theory: which are the properties that a learning algorithm $\mathscr{A}$ should fulfill in order to achieve good generalization performance? Stability answers this question in a very intuitive way: if $\mathscr{A}$ selects similar models, even if the training data are (slightly) modified, then we can be confident that the learning algorithm is stable. In other words, if Stability shows that $\mathscr{A}$ does not excessively fit the noise that afflicts the available data, therefore $\mathscr{A}$ is able to achieve good generalization. Note that, using this approach, there is no need to aprioristically fix a set of models to be explored by the learning algorithm, which represents a novelty with respect to the complexity-based approaches [265]. The groundbreaking idea of Stability, in fact, allowed to overcome some computational and theoretical issues of the complexity-based approaches, where it is necessary to fix a class of functions $\mathcal{F}$ in a data-independent way, and to measure its complexity for obtaining valid generalization bounds[1] [10, 26, 30, 240, 266].

However, several successful learning algorithms, like, for example, the k-Nearest Neighbors (k-NN) [135], are developed from (eventually heuristic) training procedures or strategies, without explicitly defining a fixed hypothesis space. The k-NN idea is to group similar objects into the same class, but the hypothesis is only defined as soon as the data become available: no set of functions, from which we pick up the best one fitting the available data, is defined [135].

[1] This is true not only in a frequentist setting, but in the Bayesian learning framework as well, where a prior distribution of models must be specified before seeing the data [117, 178, 207].

When the hypothesis space cannot be defined in advance, the complexity-bases approaches fails and it becomes mandatory to resort to resampling techniques. Some attempts have also been made for extending the complexity-based approaches to data-dependent hypothesis spaces, but no practical and general results have been obtained so far [52, 240, 277].

Stability works in the same hypothesis under which the resampling methods work, apart from the fact that it handles just deterministic algorithms and rules. Recently extensions to probabilistic rules and algorithms have been proposed [93] but results are quite preliminary and overcomplicated with respect to the scope of this book and to other tools which better handle these kinds of algorithms and rules like Differential Privacy. Stability, as we will see later, unfortunately still brings to loose bounds or almost data and algorithm independent bounds and these limitations do not allow its adoption in more contexts.

At this point we can start to describe the Stability framework in a more formal way. Let $\mathcal{X}$ and $\mathcal{Y}$ be, respectively, an input and an output space. We consider a set of labeled independent and identically distributed (i.i.d.) data $\mathcal{D}_n : \{z_1, \cdots, z_n\}$ of size n, where $z_{i \in \{1,\cdots,n\}} = (x_i, y_i)$, sampled from an unknown distribution μ where $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $z \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. A learning algorithm $\mathscr{A}$, characterized by a configuration of its hyperparameters h that must be tuned, it maps $\mathcal{D}_n$ into a function $f : \mathscr{A}_h(\mathcal{D}_n)$, which maps elements in $\mathcal{X}$ to elements in $\mathcal{Y}$. In particular, $\mathscr{A}_h$ allows designing $f \in \mathcal{F}_h$ and defining the hypothesis space $\mathcal{F}_h$, that is generally unknown (and depends on h). We assume that $\mathscr{A}_h$ satisfies some minor properties detailed in [49]: namely, we consider only deterministic algorithms which produce deterministic rules and that are symmetric with respect to $\mathcal{D}_n$ (then they do not depend on the order of the elements in the training set); moreover, all the functions are measurable and all the sets are countable. We also define two modified training sets: $\mathcal{D}_n^{\setminus i} : \{z_1, \cdots, z_{i-1}, z_{i+1}, \cdots, z_n\}$, where the i-th element is removed $\mathcal{D}_n^i : \{z_1, \cdots, z_{i-1}, z_i', z_{i+1}, \cdots, z_n\}$ and $\mathcal{D}_n^i$, where the i-th element is replaced and where z_i' is an i.i.d. pattern, sampled from μ. The accuracy of $\mathscr{A}_h(\mathcal{D}_n)$ in representing the hidden relationship μ is measured with reference to a $[0,1]$-bounded loss function $\ell : \mathcal{F}_h \times \mathcal{Z} \to [0,1]$. Consequently, the quantity of interest is defined as the generalization error, namely the error that a model will perform on new data generated by μ and previously unseen

$$L\left(\mathscr{A}_h(\mathcal{D}_n)\right) = \mathbb{E}_z \ell\left(\mathscr{A}_h(\mathcal{D}_n), z\right). \tag{7.1}$$

$L\left(\mathscr{A}_h(\mathcal{D}_n)\right)$ is a random variable that depends on $\mathcal{D}_n$, that unfortunately cannot be computed since μ is unknown and, consequently, must be estimated. Two of its most exploited estimators are the empirical error [10, 26, 30, 266]

$$\widehat{L}_{\mathrm{emp}}\left(\mathscr{A}_h(\mathcal{D}_n), \mathcal{D}_n\right) = \frac{1}{n} \sum_{z \in \mathcal{D}_n} \ell\left(\mathscr{A}_h(\mathcal{D}_n), z\right), \tag{7.2}$$

and the Leave-One-Out (LOO) error [100, 157]

$$\widehat{L}_{\mathrm{loo}}\left(\mathscr{A}_h(\mathcal{D}_n), \mathcal{D}_n\right) = \frac{1}{n} \sum_{z \in \mathcal{D}_n} \ell\left(\mathscr{A}_h(\mathcal{D}_n \backslash z), z\right). \tag{7.3}$$

With the complexity-based approaches, an upper bound of $L\left(\mathscr{A}_h(\mathcal{D}_n)\right)$ is derived by studying the supremum of the uniform deviation of the generalization error from the empirical error of Eq. (7.2) or, alternatively, from the LOO error of Eq. (7.3)

$$\sup_{f \in \mathcal{F}_h} \left| L\left(\mathscr{A}_h(\mathcal{D}_n)\right) - \widehat{L}_{\mathrm{emp}}\left(\mathscr{A}_h(\mathcal{D}_n), \mathcal{D}_n\right) \right|, \tag{7.4}$$

$$\sup_{f \in \mathcal{F}_h} \left| L\left(\mathscr{A}_h(\mathcal{D}_n)\right) - \widehat{L}_{\mathrm{loo}}\left(\mathscr{A}_h(\mathcal{D}_n), \mathcal{D}_n\right) \right|. \tag{7.5}$$

When the complexity-based approaches are adopted, it is hypothesized that the class of functions $\mathcal{F}_h$ is defined in a data-independent fashion and, then, is known. When dealing with Stability, we suppose that $\mathcal{F}_h$ is not aprioristically designed, thus studying the uniform deviation is not possible since $\mathcal{F}_h$ is unknown. The deviation $\widehat{D}\left(\mathscr{A}_h(\mathcal{D}_n), \mathcal{D}_n\right)$ of the generalization error from the empirical or the LOO errors is analyzed, instead

$$\widehat{D}_{\mathrm{emp}}\left(\mathscr{A}_h(\mathcal{D}_n), \mathcal{D}_n\right) = \left| L\left(\mathscr{A}_h(\mathcal{D}_n)\right) - \widehat{L}_{\mathrm{emp}}\left(\mathscr{A}_h(\mathcal{D}_n), \mathcal{D}_n\right) \right|, \tag{7.6}$$

$$\widehat{D}_{\mathrm{loo}}\left(\mathscr{A}_h(\mathcal{D}_n), \mathcal{D}_n\right) = \left| L\left(\mathscr{A}_h(\mathcal{D}_n)\right) - \widehat{L}_{\mathrm{loo}}\left(\mathscr{A}_h(\mathcal{D}_n), \mathcal{D}_n\right) \right|. \tag{7.7}$$

Note that the deterministic counterpart of the above mentioned sets can be defined as

$$D^2_{\mathrm{emp}}\left(\mathscr{A}_h, n\right) = \mathbb{E}_{\mathcal{D}_n} \widehat{D}^2_{\mathrm{emp}}\left(\mathscr{A}_h(\mathcal{D}_n), \mathcal{D}_n\right), \tag{7.8}$$

$$D^2_{\mathrm{loo}}\left(\mathscr{A}_h, n\right) = \mathbb{E}_{\mathcal{D}_n} \widehat{D}^2_{\mathrm{loo}}\left(\mathscr{A}_h(\mathcal{D}_n), \mathcal{D}_n\right). \tag{7.9}$$

In order to study $\widehat{D}\left(\mathscr{A}_h(\mathcal{D}_n), \mathcal{D}_n\right)$, we can adopt different approaches. The first one consists in using the hypothesis Stability $H\left(\mathscr{A}_h, n\right)$

$$H_{\mathrm{emp}}\left(\mathscr{A}_h, n\right) = \mathbb{E}_{\mathcal{D}_n, z'_i} \left| \ell\left(\mathscr{A}_h(\mathcal{D}_n), z_i\right) - \ell\left(\mathscr{A}_h(\mathcal{D}^i_n), z_i\right) \right| \leqslant \beta_{\mathrm{emp}}, \tag{7.10}$$

$$H_{\mathrm{loo}}\left(\mathscr{A}_h, n\right) = \mathbb{E}_{\mathcal{D}_n, z} \left| \ell\left(\mathscr{A}_h(\mathcal{D}_n), z\right) - \ell\left(\mathscr{A}_h(\mathcal{D}^{\backslash i}_n), z\right) \right| \leqslant \beta_{\mathrm{loo}}. \tag{7.11}$$

Lemma 3 in [49] proves that:

$$D^2_{\mathrm{emp}}\left(\mathscr{A}_h, n\right) \leqslant \frac{1}{2n} + 3H_{\mathrm{emp}}\left(\mathscr{A}_h, n\right), \tag{7.12}$$

$$D^2_{\mathrm{loo}}\left(\mathscr{A}_h, n\right) \leqslant \frac{1}{2n} + 3H_{\mathrm{loo}}\left(\mathscr{A}_h, n\right). \tag{7.13}$$

By exploiting the Chebyshev inequality [54], Eqns. (7.6), (7.10) and (7.12) (or, analogously, Eqns. (7.7), (7.11) and (7.13)) we obtain that, with probability $(1-\delta)$:

$$L\left(\mathscr{A}_h(\mathcal{D}_n)\right) \leqslant \widehat{L}_{\mathrm{emp}}\left(\mathscr{A}_h(\mathcal{D}_n), \mathcal{D}_n\right) + \sqrt{\frac{1}{2n\delta} + \frac{3\beta_{\mathrm{emp}}}{\delta}}, \tag{7.14}$$

$$L\left(\mathscr{A}_h(\mathcal{D}_n)\right) \leqslant \widehat{L}_{\mathrm{loo}}\left(\mathscr{A}_h(\mathcal{D}_n), \mathcal{D}_n\right) + \sqrt{\frac{1}{2n\delta} + \frac{3\beta_{\mathrm{loo}}}{\delta}}, \tag{7.15}$$

which are the polynomial bounds previously derived in [49] and based on hypothesis stabilty[2].

Another approach, targeted towards deriving a Stability bound on the generalization error, consists in exploiting the uniform Stability $U(\mathscr{A}_h)$:

$$U^i\left(\mathscr{A}_h, n\right) = \left|\ell\left(\mathscr{A}_h(\mathcal{D}_n), \cdot\right) - \ell\left(\mathscr{A}_h(\mathcal{D}_n^i), \cdot\right)\right|_\infty \leqslant \beta^i, \tag{7.16}$$

$$U^{\setminus i}\left(\mathscr{A}_h, n\right) = \left|\ell\left(\mathscr{A}_h(\mathcal{D}_n), \cdot\right) - \ell\left(\mathscr{A}_h(\mathcal{D}_n^{\setminus i}), \cdot\right)\right|_\infty \leqslant \beta^{\setminus i}. \tag{7.17}$$

Note that:

$$H_{\mathrm{emp}}\left(\mathscr{A}_{\mathcal{H}}, n\right) \leqslant U^i\left(\mathscr{A}_{\mathcal{H}}, n\right), \tag{7.18}$$

$$H_{\mathrm{loo}}\left(\mathscr{A}_{\mathcal{H}}, n\right) \leqslant U^{\setminus i}\left(\mathscr{A}_{\mathcal{H}}, n\right). \tag{7.19}$$

By exploiting the McDiarmid's Inequality [181] it is possible to derive the following exponential bounds [49], that hold with probability $(1-\delta)$:

$$L\left(\mathscr{A}_h(\mathcal{D}_n)\right) \leqslant \widehat{L}_{\mathrm{emp}}\left(\mathscr{A}_h(\mathcal{D}_n), \mathcal{D}_n\right) + 2\beta^i + \left(4n\beta^i + 1\right)\sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2n}}, \tag{7.20}$$

$$L\left(\mathscr{A}_h(\mathcal{D}_n)\right) \leqslant \widehat{L}_{\mathrm{loo}}\left(\mathscr{A}_h(\mathcal{D}_n), \mathcal{D}_n\right) + \beta^{\setminus i} + \left(4n\beta^{\setminus i} + 1\right)\sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2n}}. \tag{7.21}$$

Although the bounds are exponential, Stability must decrease with n in order to obtain a non-trivial result. Unfortunately, this is seldom the case: for

[2] For the sake of precision, note that the bounds slightly differ from the results proposed in [49]: as a matter of fact, as also underlined in [186], the original work on Stability [49] contains one error, which motivates the exploitation of the two notions of hypothesis Stability of Eqns. (7.12) and (7.13).

example, when a hard $\{0,1\}$-loss function is exploited in binary classification to count the number of misclassifications, it is possible to prove that $\beta^i = \beta^{\setminus i} = 1$ for many well-known and widely used algorithms [75, 76] (such as k-Local Rules [224] or Support Vector Machines [71]). Moreover, in those cases where non-trivial results can be derived (e.g., in bounded Support Vector Regression [81]), strong conditions on μ must hold, which are rarely satisfied in practice. Finally, also note that the previous results are all algorithmic-dependent, but they are not data-dependent: as remarked in the introduction, this represents a drawback in practical applications.

In order to cope with these blind spots of Stability it is necessary to derive a fully empirical and data-dependent result.

Let us consider the LOO error: we will devote a paragraph later to the motivations of this choice. We have to start by making an assumption on the learning algorithm $\mathscr{A}$. In particular, we suppose that the hypothesis Stability does not increase with the cardinality of the training set:

$$D_{\text{loo}}\left(\mathscr{A}_h, n\right) \leqslant D_{\text{loo}}\left(\mathscr{A}_h, \frac{\sqrt{n}}{2}\right). \tag{7.22}$$

We point out that Property (7.22) is a desirable requirement for any learning algorithm: in fact, the impact on the learning procedure of removing samples from $\mathcal{D}_n$ should decrease, on average, as n grows. Alternatively, we can hypothesize that:

$$H_{\text{loo}}\left(\mathscr{A}_h, n\right) \leqslant H_{\text{loo}}\left(\mathscr{A}_h, \frac{\sqrt{n}}{2}\right). \tag{7.23}$$

Note that:

$$\begin{aligned} & H_{\text{loo}}\left(\mathscr{A}_{\mathcal{H}}, n\right) \leqslant H_{\text{loo}}\left(\mathscr{A}_{\mathcal{H}}, \frac{\sqrt{n}}{2}\right) \\ \to \quad & D_{\text{loo}}\left(\mathscr{A}_{\mathcal{H}}, n\right) \leqslant D_{\text{loo}}\left(\mathscr{A}_{\mathcal{H}}, \frac{\sqrt{n}}{2}\right). \end{aligned} \tag{7.24}$$

Note also that Property (7.22) (or, alternatively, Property (7.23)) has already been studied by many researchers in the past. In particular, these properties are related to the concept of consistency [75, 251]. However, connections can also be identified with the trend of the learning curves of an algorithm [78, 187, 204, 205]. Moreover, such quantities are strictly linked to the concept of Smart Rule [75]. The purpose of these works is to prove that an algorithm performs better as the cardinality of the learning set increases: then, the more data we have, the more concentrated the empirical or the LOO errors should be around the generalization error. It is worth underlining that, in many of the above-referenced works, Property (7.22) (or, alternatively, Property (7.23)) is proved to be satisfied by many well known algorithms (Support Vector Machines, Kernelized Regularized Least Squares, k-Local Rules with $k > 1$, etc.)

In the following, we start by considering and using the assumption of Eq. (7.22). In this case, we exploit the Chebyshev inequality [54] and derive that, with probability $(1-\delta)$:

$$\widehat{D}_{\text{loo}}(\mathscr{A}_h(\mathcal{D}_n),\mathcal{D}_n) \leqslant \sqrt{\frac{D^2_{\text{loo}}\left(\mathscr{A}_h,n\right)}{\delta}} \leqslant \sqrt{\frac{D^2_{\text{loo}}\left(\mathscr{A}_h,\frac{\sqrt{n}}{2}\right)}{\delta}}. \tag{7.25}$$

By exploiting Eq. (7.13), we have that:

$$D^2_{\text{loo}}\left(\mathscr{A}_h,\frac{\sqrt{n}}{2}\right) \leqslant \frac{1}{\sqrt{n}} + 3H_{\text{loo}}\left(\mathscr{A}_h,\frac{\sqrt{n}}{2}\right). \tag{7.26}$$

We focus now on $H_{\text{loo}}\left(\mathscr{A}_h,\frac{\sqrt{n}}{2}\right)$. For this purpose, let us introduce the following empirical quantity:

$$\begin{aligned}&\widehat{H}_{\text{loo}}\left(\mathscr{A}_h\left(\mathcal{D}_{\frac{\sqrt{n}}{2}}\right),\mathcal{D}_{\frac{\sqrt{n}}{2}}\right) \\ &= \frac{8}{n\sqrt{n}}\sum_{k=1}^{\frac{\sqrt{n}}{2}}\sum_{j=1}^{\frac{\sqrt{n}}{2}}\sum_{i=1}^{\frac{\sqrt{n}}{2}}\left|\ell\left(\mathscr{A}\left(\breve{\mathcal{D}}^k_{\frac{\sqrt{n}}{2}}\right),\breve{z}^k_j\right) - \ell\left(\mathscr{A}\left(\left(\breve{\mathcal{D}}^k_{\frac{\sqrt{n}}{2}}\right)^{\setminus i}\right),\breve{z}^k_j\right)\right|,\end{aligned} \tag{7.27}$$

where:

$$\breve{\mathcal{D}}^k_{\frac{\sqrt{n}}{2}}: \quad \left\{z_{(k-1)\sqrt{n}+1},\cdots,z_{(k-1)\sqrt{n}+\frac{\sqrt{n}}{2}}\right\}, \quad k \in \left\{1,\cdots,\frac{\sqrt{n}}{2}\right\}, \tag{7.28}$$

$$\breve{z}^k_j: \quad z_{(k-1)\sqrt{n}+\frac{\sqrt{n}}{2}+j}, \quad k \in \left\{1,\cdots,\frac{\sqrt{n}}{2}\right\}. \tag{7.29}$$

Note that the quantity of Eq. (7.27) is the empirical unbiased estimator of $H_{\text{loo}}\left(\mathscr{A}_h,\frac{\sqrt{n}}{2}\right)$ and then:

$$H_{\text{loo}}\left(\mathscr{A}_h,\frac{\sqrt{n}}{2}\right) = \mathbb{E}_{\mathcal{D}_{\frac{\sqrt{n}}{2}}}\widehat{H}_{\text{loo}}\left(\mathscr{A}_h\left(\mathcal{D}_{\frac{\sqrt{n}}{2}}\right),\mathcal{D}_{\frac{\sqrt{n}}{2}}\right). \tag{7.30}$$

It is worth noting that, when dealing with $\widehat{H}_{\text{loo}}\left(\mathscr{A}_h\left(\mathcal{D}_{\frac{\sqrt{n}}{2}}\right),\mathcal{D}_{\frac{\sqrt{n}}{2}}\right)$, all the samples z_i are i.i.d. and sampled from μ. Thus

$$\left|\ell\left(\mathscr{A}\left(\breve{\mathcal{D}}^k_{\frac{\sqrt{n}}{2}}\right),\breve{z}^k_j\right) - \ell\left(\mathscr{A}\left(\left(\breve{\mathcal{D}}^k_{\frac{\sqrt{n}}{2}}\right)^{\setminus i}\right),\breve{z}^k_j\right)\right| \in [0,1], \quad \forall i,j,k \tag{7.31}$$

will be i.i.d., and we can bound in probability the difference between the $H_{\text{loo}}\left(\mathscr{A}_h, \frac{\sqrt{n}}{2}\right)$ and the $\widehat{H}_{\text{loo}}\left(\mathscr{A}_h\left(\mathcal{D}_{\frac{\sqrt{n}}{2}}\right), \mathcal{D}_{\frac{\sqrt{n}}{2}}\right)$ by exploiting, for example, the Hoeffding inequality [118]:

$$\mathbb{P}\left[H_{\text{loo}}\left(\mathscr{A}_h, \frac{\sqrt{n}}{2}\right) - \widehat{H}_{\text{loo}}\left(\mathscr{A}_h\left(\mathcal{D}_{\frac{\sqrt{n}}{2}}\right), \mathcal{D}_{\frac{\sqrt{n}}{2}}\right) > t\right] \leqslant e^{-\sqrt{n}t^2}. \tag{7.32}$$

Then, with probability $(1-\delta)$:

$$H_{\text{loo}}\left(\mathscr{A}_h, \frac{\sqrt{n}}{2}\right) \leqslant \widehat{H}_{\text{loo}}\left(\mathscr{A}_h\left(\mathcal{D}_{\frac{\sqrt{n}}{2}}\right), \mathcal{D}_{\frac{\sqrt{n}}{2}}\right) + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{\sqrt{n}}}. \tag{7.33}$$

Combining Eqns. (7.25), (7.26) and (7.33) we get that, with probability $(1-\delta)$:

$$\begin{aligned} L\left(\mathscr{A}_h(\mathcal{D}_n)\right) \leqslant & \widehat{L}_{\text{loo}}\left(\mathscr{A}_h(\mathcal{D}_n), \mathcal{D}_n\right) \\ & + \sqrt{\frac{2}{\delta}\left[\frac{1}{\sqrt{n}} + 3\left[\widehat{H}_{\text{loo}}\left(\mathscr{A}_h\left(\mathcal{D}_{\frac{\sqrt{n}}{2}}\right), \mathcal{D}_{\frac{\sqrt{n}}{2}}\right) + \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{\sqrt{n}}}\right]\right]}. \end{aligned} \tag{7.34}$$

When exploiting Property (7.23), the proof is analogous. We can make use of Eqns. (7.13) and the Chebyshev inequality [54], and state that, with probability $(1-\delta)$:

$$\widehat{D}_{\text{loo}}(\mathscr{A}_{(\mathcal{D}_n, \mathcal{H})}, \mathcal{D}_n) \leqslant \sqrt{\frac{1}{\delta}\left[\frac{1}{2n} + 3H_{\text{loo}}\left(\mathscr{A}_{\mathcal{H}}, \frac{\sqrt{n}}{2}\right)\right]}. \tag{7.35}$$

Then, with probability $(1-\delta)$:

$$\begin{aligned} L\left(\mathscr{A}_h(\mathcal{D}_n)\right) \leqslant & \widehat{L}_{\text{loo}}\left(\mathscr{A}_h(\mathcal{D}_n), \mathcal{D}_n\right) \\ & + \sqrt{\frac{2}{\delta}\left[\frac{1}{2n} + 3\left(\widehat{H}_{\text{loo}}\left(\mathscr{A}_h\left(\mathcal{D}_{\frac{\sqrt{n}}{2}}\right), \mathcal{D}_{\frac{\sqrt{n}}{2}}\right) + \sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{\sqrt{n}}}\right)\right]}. \end{aligned} \tag{7.36}$$

Let us roll back to the choice of using the LOO error in place of the empirical error for the previous proofs. One can imagine that the empirical estimator can be exploited as well, e.g., by defining two properties analogous to the ones of Eqns. (7.22) and (7.23):

$$D_{\text{emp}}\left(\mathscr{A}_h, n\right) \leqslant D_{\text{emp}}\left(\mathscr{A}_h, \frac{\sqrt{n}}{2}\right), \tag{7.37}$$

$$H_{\text{emp}}\left(\mathscr{A}_h, n\right) \leqslant H_{\text{emp}}\left(\mathscr{A}_h, \frac{\sqrt{n}}{2}\right). \tag{7.38}$$

Consequently, we can study the empirical estimator of $H_{\text{emp}}\left(\mathscr{A}_h, \frac{\sqrt{n}}{2}\right)$. For this purpose we have to introduce the following empirical quantity:

$$\begin{aligned}
&\widehat{H}_{\text{emp}}\left(\mathscr{A}_h\left(\mathcal{D}_{\frac{\sqrt{n}}{2}}\right), \mathcal{D}_{\frac{\sqrt{n}}{2}}\right)\\
&= \frac{4}{m}\sum_{k=1}^{\frac{\sqrt{n}}{2}}\sum_{i=1}^{\frac{\sqrt{n}}{2}}\left|\ell\left(\mathscr{A}\left(\breve{\mathcal{D}}^k_{\frac{\sqrt{n}}{2}}\right), \breve{z}^k_i\right) - \ell\left(\mathscr{A}\left(\left(\breve{\mathcal{D}}^k_{\frac{\sqrt{n}}{2}}\right)^i\right), \breve{z}^k_i\right)\right|, \qquad (7.39)
\end{aligned}$$

where:

$$\breve{\mathcal{D}}^k_{\frac{\sqrt{n}}{2}}: \quad \left\{z_{(k-1)\sqrt{n}+1}, \cdots, z_{(k-1)\sqrt{n}+\frac{\sqrt{n}}{2}}\right\}, \quad k \in \left\{1, \cdots, \frac{\sqrt{n}}{2}\right\}, \qquad (7.40)$$

$$\breve{z}^k_i: \quad z_{(k-1)\sqrt{n}+i}, \quad k \in \left\{1, \cdots, \frac{\sqrt{n}}{2}\right\}, \qquad (7.41)$$

$$\begin{aligned}
\left(\breve{\mathcal{D}}^k_{\frac{\sqrt{n}}{2}}\right)^i: \quad &\left\{z_{(k-1)\sqrt{n}+1}, \cdots, \breve{z}'^k_i, \cdots, z_{(k-1)\sqrt{n}+\frac{\sqrt{n}}{2}}\right\},\\
&k \in \left\{1, \cdots, \frac{\sqrt{n}}{2}\right\}, \qquad (7.42)
\end{aligned}$$

$$\breve{z}'^k_i: \quad z_{(k-1)\sqrt{n}+\frac{\sqrt{n}}{2}+i}, \quad k \in \left\{1, \cdots, \frac{\sqrt{n}}{2}\right\}. \qquad (7.43)$$

Unfortunately, although all the patterns z_i are i.i.d. and sampled from μ, the terms in the summations of Eq. (7.39)

$$\left|\ell\left(\mathscr{A}\left(\breve{\mathcal{D}}^k_{\frac{\sqrt{n}}{2}}\right), \breve{z}^k_i\right) - \ell\left(\mathscr{A}\left(\left(\breve{\mathcal{D}}^k_{\frac{\sqrt{n}}{2}}\right)^i\right), \breve{z}^k_i\right)\right| \in [0,1], \quad \forall i, k \qquad (7.44)$$

are not i.i.d. Thus, a bound, analogous to the one for the LOO error, cannot be derived.

Note that it does not make much sense to exploit sharper bounds in Eq. (7.33) since the sharpness of the fully empirical bounds of Eqns. (7.34) and (7.34) mainly depends on the result of Eq. (7.15) for which faster bounds do not exist. Recently [174] some attempts have been made to obtain a Bernstein-type bound [38] but the results do not take into account the empirical variance of the error, but the empirical variance of the data distribution which is seldom small.

Finally note that the fully empirical bounds of Eqns. (7.34) and (7.36) can be used both for MS and EE purposes by exploiting the approach described in the preliminaries of this book [195, 199]. In fact all the Compression-based bounds have the following form:

$$\mathbb{P}_{\mathcal{D}_n}\left\{L(\mathscr{A}_h(\mathcal{D}_n)) \leqslant \Delta(\mathcal{D}_n, \mathscr{A}_h, \delta)\right\} \geqslant 1 - \delta. \qquad (7.45)$$

Algorithm 4: Algorithmic Stability: MS and EE Strategy.

```
Input: $\mathcal{A}_{\mathcal{H}}$, $\mathcal{D}_n$, and $\delta$
Output: Optimal Model $\mathscr{A}_h^*(\mathcal{D}_n)$ and its estimated generalization error
        $L(\mathscr{A}_h^*(\mathcal{D}_n))$
$L^*_{\mathrm{MS}} = +\infty$;
for $\mathscr{A}_h \in \mathcal{A}_{\mathcal{H}}$ do
    $L_{\mathrm{MS}} = \Delta(\mathcal{D}_n, \mathscr{A}_h, \delta)$;
    if $L^*_{MS} > L_{MS}$ then
        $L^*_{\mathrm{MS}} = L_{\mathrm{MS}}$;
        $\mathscr{A}_h^*(\mathcal{D}_n) = \mathscr{A}_h(\mathcal{D}_n)$;
        $L(\mathscr{A}_h^*(\mathcal{D}_n)) = \Delta\left(\mathcal{D}_n, \mathscr{A}_h, \frac{\delta}{|\mathcal{A}_{\mathcal{H}}|}\right)$;
```

Then if we want to choose $\mathscr{A}_h^* \in \mathcal{A}_{\mathcal{H}} = \{\mathscr{A}_h : \mathscr{A} \in \mathcal{A}, h \in \mathcal{H}_{\mathscr{A}}\}$, namely perform the MS phase, and estimate the generalization performance of $\mathscr{A}_h^*(\mathcal{D}_n)$, namely perform the EE phase, we have to follow the procedure summarized in Algorithm 4. Note that the generalization of the final model is bounded by

$$\mathbb{P}_{\mathcal{D}_n}\left\{L(\mathscr{A}_h^*(\mathcal{D}_n)) \leqslant \Delta\left(\mathcal{D}_n, \mathscr{A}_h^*, \frac{\delta}{|\mathcal{A}_{\mathcal{H}}|}\right)\right\} \geqslant 1-\delta, \quad \forall \mathscr{A}_h^* \in \mathcal{A}_{\mathcal{H}} \tag{7.46}$$

with probability $(1-\delta)$, since we have applied the Bonferroni correction [45] over the $|\mathcal{A}_{\mathcal{H}}|$ choices for the algorithm and hyperparameters configurations.

8

PAC-Bayes Theory

It is well known that combining the output of several rules results in much better performance than using any one of them alone. In fact many state-of-the-art algorithms search for a weighted combination of simpler rules [104]: Bagging [50, 51], Boosting [229, 230] and Bayesian approaches [102] or even Kernel methods [265] and Neural Networks [39]. The major open problem in this scenario is how to weight the different rules in order to obtain good performance [37, 55, 159, 160, 190, 207], how these performances can be assessed [55, 79, 103, 104, 146, 150, 152, 153, 159, 160, 163, 176, 178, 179, 242, 260, 264], and how this theoretical framework can be exploited for deriving new learning approaches or for applying it in other contexts [22, 23, 32, 105, 178, 185, 217, 227, 231–236, 241]. The Probably Approximately Correct Bayes (PAC-Bayes) approach is one of the sharpest analysis frameworks in this context, since it can provide tight bounds on the risk of the Gibbs Classifier (GC), also called Randomised (or probabilistic) Classifier, and the Bayes Classifier (BC), also called Weighted Majority Vote Classifier [104]. The GC chooses a classifier in the set of classifiers according to the posterior distribution each time a new sample has to be classified [160] while the BC takes the decision based on the expected value of the GC over the posterior distribution [104]. In particular, in the PAC-Bayes framework a prior distribution over the different classifiers must be defined before seeing the data, then, based on the available data, a posterior distribution can be chosen, and the risk of the associate GC and BC is computed, based on the empirical risk and the divergence between the prior and posterior distributions [176].

Note that GC and BC are classifiers and not rules since the PAC-Bayes Theory deals mainly with classifiers, in particular binary classifiers, and most of the results are not yet extended to the general SL framework. For this reason, with the PAC-Bayes Theory, we will only deal with the binary classification framework.

In the conventional PAC-Bayes framework, a posterior distribution that minimizes the divergence between prior and posterior distributions must be chosen, since this divergence forms part of the bound. This choice is critical: in some cases this choice results to be too generic and not suited for the particular problem [160], other times some data are kept apart from the learning process and exploited to derive a generally good prior [147, 207]. Consequently in the first case the divergence term in the PAC-Bayes analysis can typically be large, while in the second case the bound tends to be loose since some data are wasted in order to design the prior. In order to address this issue in [55] a localized PAC-Bayes analysis is proposed, which uses a Boltzmann prior distribution defined in terms of the distribution that has generated the data. Note that, since the prior depends on the distribution, the PAC-Bayes analysis is still valid because the prior is defined before observing the data [55]. By tuning the prior to the distribution, Catoni was able to remove the divergence term from the bound, hence significantly reducing the complexity penalty. More recently this approach has been extended in [160, 193] by deriving some new sharper bounds and by combining these results with the recent development in the analysis of the GC reported in [104]. Note that other approaches for removing the divergence exist. One approach is to design a prior and a posterior such that they are aligned [104, 105, 227]. The second one is to design a so called expectation-prior which does not require any separate set of data to build a prior which will be probably close to the posterior [207]. Every approach has its own strengths and weaknesses but the approach of Catoni seems to be the most promising one [55, 160] even if using Boltzmann distributions in some contexts can be seen as a limitation [160]. In fact, keeping the divergence term allowed many researchers to design new MS methods and learning algorithms [3, 103, 241].

As it should be clear by the general description reported above, the PAC-Bayes Theory deals with deterministic algorithms which choose a distribution over a set of known deterministic rules. Moreover data must be sampled i.i.d. Finally, even if the PAC-Bayes Theory is one of the sharpest analysis for probabilistic rules, a lot of research is still ongoing for the definition of appropriate prior and posterior distributions and for sharpening the already quite effective generalization bounds.

In order to present the state-of-the-art, we first recall some common definitions [49, 104, 265]. Let us consider a set of labeled samples defined as $\mathcal{D}_n = \{(x_1, y_1), \cdots, (x_n, y_n)\} = \{z_1, \cdots, z_n\}$ drawn i.i.d. according to an unknown probability distribution μ over the cartesian product between the input space $\mathcal{X}$ and the output space $\mathcal{Y} = \{-1, +1\}$, defined as $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Let us consider a function f in a set of possible ones $\mathcal{F}$, where $f : \mathcal{X} \to \overline{\mathcal{Y}} = [-1, +1]$. The error of f in approximating μ is measured with reference to some $[0, 1]$-bounded loss function $\ell : \mathcal{F} \times \mathcal{Z} \to [0, 1]$. Then the risk of f can be defined as

$$L^{\ell}(f) = \mathbb{E}_z \left\{\ell(f, z)\right\}, \tag{8.1}$$

together with its variance

$$V^{\ell}(f) = \mathbb{E}_z\{(\ell(f, z) - \mathbb{E}_z \ell(f, z))^2\}. \tag{8.2}$$

Since μ is unknown $L^{\ell}(f)$ and $V^{\ell}(f)$ cannot be computed, but we can compute its empirical estimators, the empirical error

$$\widehat{L}^{\ell}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f, z_i), \tag{8.3}$$

and the empirical variance

$$\widehat{V}^{\ell}(f) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\ell(f, z_i) - \widehat{L}^{\ell}(f)\right)^2. \tag{8.4}$$

Since in this part we will deal with binary classification problems, we will often make use of the Hard loss function $\ell_H(f, z) = [y f(x) \leqslant 0]$, but also the Linear loss function $\ell_S(f, z) = {}^1\!/\!_2(1 - y f(x))$ will be exploited. Note that, if $f_B \in \mathcal{F}_B$ is a binary classifier $f_B : \mathcal{X} \to \{\pm 1\}$, we have that $\ell_S(f_B, z) = \ell_H(f_B, z)$.

The GC draws an $f \in \mathcal{F}$, according to a probability distribution Q over $\mathcal{F}$, each time a label for an input $x \in \mathcal{X}$ is required. For the GC, that we will call G_{Q}, we can define its risk together with its empirical counterpart [160]

$$L^{\ell}(G_{\mathsf{Q}}) = \mathbb{E}_{f \sim \mathsf{Q}}\{L^{\ell}(f)\}, \quad \widehat{L}^{\ell}(G_{\mathsf{Q}}) = \mathbb{E}_{f \sim \mathsf{Q}}\{\widehat{L}^{\ell}(f)\}. \tag{8.5}$$

Analogously it is possible to define the average expected variance and the average empirical variance

$$V^{\ell}(G_{\mathsf{Q}}) = \mathbb{E}_{f \sim \mathsf{Q}} V^{\ell}(f), \quad \widehat{V}^{\ell}(G_{\mathsf{Q}}) = \mathbb{E}_{f \sim \mathsf{Q}} \widehat{V}^{\ell}(f). \tag{8.6}$$

The BC [104], instead, can be defined as

$$B_{\mathsf{Q}}(x) = \operatorname{sign}\left[\mathbb{E}_{f\sim\mathsf{Q}}\left\{f(x)\right\}\right]. \tag{8.7}$$

Consequently, it is possible to define the generalization error of the BC [104, 160] as

$$L^{\ell}(B_{\mathsf{Q}}) = \mathbb{E}_{z}\left\{\ell(B_{\mathsf{Q}}, z)\right\}. \tag{8.8}$$

Let us define the Expected Disagreement (ED) [104] of $f \in \mathcal{F}$ with respect to Q, together with its empirical counterpart computed over s samples:

$$d_{\mathsf{Q}} = {}^{1}\!/_{2} - {}^{1}\!/_{2}\mathbb{E}_{x}\left\{\left[\mathbb{E}_{f\sim\mathsf{Q}}\left\{f(x)\right\}\right]^{2}\right\}, \tag{8.9}$$

$$\widehat{d}^{s}_{\mathsf{Q}} = {}^{1}\!/_{2} - {}^{1}\!/_{2s}\sum_{i=1}^{s}\left[\mathbb{E}_{f\sim\mathsf{Q}}\left\{f(x_i)\right\}\right]^{2}. \tag{8.10}$$

The Expected Joint Error (EJE) [104] of $f \in \mathcal{F}$ with respect to Q, together with its empirical counterpart, is:

$$e^{\ell}_{\mathsf{Q}} = \mathbb{E}_{z,f_1\sim\mathsf{Q},f_2\sim\mathsf{Q}}\left\{\ell(f_1,z)\ell(f_2,z)\right\}, \tag{8.11}$$

$$\widehat{e}^{\ell}_{\mathsf{Q}} = \mathbb{E}_{f_1\sim\mathsf{Q},f_2\sim\mathsf{Q}}\left\{{}^{1}\!/_{n}\sum_{i=1}^{n}\left[\ell(f_1,z_i)\ell(f_2,z_i)\right]\right\}. \tag{8.12}$$

Note that in [104] it is proved that $d_{\mathsf{Q}} \leqslant 2\left(\sqrt{e_{\mathsf{Q}}} - e_{\mathsf{Q}}\right)$. We denote with $\mathtt{KL}[\mathsf{Q}||P]$ the Kullback-Leibler Divergence (KLD) [262] between P and Q, while $\mathtt{kl}[q||p]$ and $\mathtt{kl}[\begin{smallmatrix}q_1\\q_2\end{smallmatrix}||\begin{smallmatrix}p_1\\p_2\end{smallmatrix}]$ are, respectively, the KLD for the Binomial and Trinomial distributions [103]:

$$\mathtt{kl}[q||p] = q\ln\left[\frac{q}{p}\right] + [1-q]\ln\left[\frac{1-q}{1-p}\right], \tag{8.13}$$

$$\mathtt{kl}\left[\begin{matrix}q_1\\q_2\end{matrix}\middle\|\begin{matrix}p_1\\p_2\end{matrix}\right] = q_1\ln\left[\frac{q_1}{p_1}\right] + q_2\ln\left[\frac{q_2}{p_2}\right] + [1-q_1-q_2]\ln\left[\frac{1-q_1-q_2}{1-p_1-p_2}\right].$$

Thanks to the Pinsker's Inequality [262], we can state that $|q-p| \leqslant \sqrt{{}^{1}\!/_{2}\mathtt{kl}(q||p)}$. Finally, let us recall the definition of the last fundamental quantity in the PAC-Bayes framework [103, 172]:

$$\xi_n = \sum_{k=0}^{n}\binom{n}{k}\left(\frac{k}{n}\right)^{k}\left(1-\frac{k}{n}\right)^{n-k} \in [\sqrt{n}, 2\sqrt{n}], \tag{8.14}$$

which will appear in many of the subsequent results.

A lot of work has been done in order to bound the risk of the BC and GC. The first one bounds the risk of the GC in terms of its empirical estimate.

For any probability distribution P over $\mathcal{F}$, chosen before seeing $\mathcal{D}_n$, $\forall \mathsf{Q}$ we have[1] [231, 241]

$$\mathbb{P}\left\{ \mathtt{kl}\left[\widehat{L}^\ell(G_\mathsf{Q})||L^\ell(G_\mathsf{Q})\right] \geqslant \frac{\mathtt{KL} + \ln\left[\frac{\xi_n}{\delta}\right]}{n} \right\} \leqslant \delta. \tag{8.15}$$

Consequently, with probability at least $(1-\delta)$, we have that:

$$L^\ell(G_\mathsf{Q}) \in \left[\inf \mathcal{I}_0^\ell(\delta, \mathsf{P}, \mathsf{Q}, n), \sup \mathcal{I}_0^\ell(\delta, \mathsf{P}, \mathsf{Q}, n)\right], \tag{8.16}$$

where

$$\mathcal{I}_0^\ell(\delta, \mathsf{P}, \mathsf{Q}, n) = \left\{ r : r \in [0,1], \mathtt{kl}\left[\widehat{L}^\ell(G_\mathsf{Q})||r\right] \leqslant \frac{\mathtt{KL} + \ln\left[\frac{\xi_n}{\delta}\right]}{n} \right\}. \tag{8.17}$$

Recently new versions of the PAC-Bayes bounds of Eq. (8.15) based on the Rényi Divergence [33] have been proposed. These bounds, even if tighter in some cases, do not improve the convergence rate but only the constants involved in the bound and they are quite complicate and out of the scope of this book.

The second result bounds the risk of the BC and it is commonly known as the $\mathtt{C}$-bound. Unfortunately, this bound involves only quantities that cannot be computed from the data.

Given the risk of the GC, the ED and EJE of Q over $\mathcal{F}$, we have that [104, 146]

$$L^\ell(B_\mathsf{Q}) \leqslant 1 - \frac{[1 - 2L^\ell(G_\mathsf{Q})]^2}{1 - 2d_\mathsf{Q}} = 1 - \frac{[1 - (2e_\mathsf{Q}^\ell + d_\mathsf{Q})]^2}{1 - 2d_\mathsf{Q}}. \tag{8.18}$$

This bound holds only for $\ell = \ell_S$. Moreover, $L^\ell(B_\mathsf{Q}) \leqslant 2L^\ell(G_\mathsf{Q})$ which holds for both $\ell = \ell_S$ and $\ell = \ell_H$.

By exploiting the bounds of Eqns. (8.15) and (8.18) it is possible to obtain the third milestone result, which is an empirical bound over the risk of the BC.

For any probability distribution P over $\mathcal{F}$, chosen before seeing $\mathcal{D}_n$, with probability at least $(1-\delta)$ and $\forall \mathsf{Q}$, we have that:

$$L^\ell(B_\mathsf{Q}) \leqslant 2 \min\left[1/2, \sup \mathcal{I}_0^\ell(\delta, \mathsf{P}, \mathsf{Q}, n)\right]. \tag{8.19}$$

This bound holds for both $\ell = \ell_S$ and $\ell = \ell_H$.

[1] In the following we will sometimes indicate $\mathtt{KL} = \mathtt{KL}[\mathsf{Q}||\mathsf{P}]$ for brevity.

The next result bounds the ED in terms of its empirical counterparts and its proof is very similar to the one of the bound of Eq. (8.15).

For any probability distribution P over $\mathcal{F}$, chosen before seeing $\mathcal{D}_n$, with probability at least $(1-\delta)$ and $\forall \mathsf{Q}$, we have that [104]

$$d_{\mathsf{Q}} \in \left[\inf \mathcal{I}_1(\delta, \mathsf{P}, \mathsf{Q}, n), \sup \mathcal{I}_1(\delta, \mathsf{P}, \mathsf{Q}, n)\right], \tag{8.20}$$

where

$$\mathcal{I}_1(\delta, \mathsf{P}, \mathsf{Q}, n)= \left\{ r : r \in \left[0, \frac{1}{2}\right], \mathtt{kl}\left[\widehat{d}^n_{\mathsf{Q}} || r\right] \leqslant \frac{2\mathtt{KL} + \ln\left[\frac{\xi_n}{\delta}\right]}{n} \right\}. \tag{8.21}$$

Based on the bounds of Eqns. (8.15), (8.18), and (8.20), it is possible to prove the fourth milestone result, which bounds the risk of the BC.

For any probability distribution P over $\mathcal{F}$, chosen before seeing $\mathcal{D}_n$, with probability at least $(1-2\delta)$ and $\forall \mathsf{Q}$, we have that [104]

$$L^{\ell}(B_{\mathsf{Q}}) \leqslant 1 - \frac{\left(1 - 2\min\left[{}^1\!/\!{}_2, \sup \mathcal{I}^{\ell}_0(\delta, \mathsf{P}, \mathsf{Q}, n)\right]\right)^2}{1 - 2\inf \mathcal{I}_1(\delta, \mathsf{P}, \mathsf{Q}, n)}. \tag{8.22}$$

This bound holds only for $\ell = \ell_S$.

Finally, the fifth milestone result, which is also the most recent one in PAC analysis, can be reported.

For any probability distribution P over $\mathcal{F}$, chosen before seeing $\mathcal{D}_n$, with probability at least $(1-\delta)$ and $\forall \mathsf{Q}$, we have that [104, 278]

$$L^{\ell}(B_{\mathsf{Q}}) \leqslant \sup_{(e,d) \in \mathcal{I}^{\ell}_2(\delta, \mathsf{P}, \mathsf{Q}, n)} 1 - \frac{\left(1 - \min[1, (2e+d)]\right)^2}{1-2d}, \tag{8.23}$$

where

$$\begin{aligned} &\mathcal{I}^{\ell}_2(\delta, \mathsf{P}, \mathsf{Q}, n) \qquad\qquad (8.24) \\ &= \left\{ (e,d) : e, d \in [0,1],\ d \leqslant 2\left(\sqrt{e} - e\right), \mathtt{kl}\left[\begin{matrix}\widehat{d}^n_{\mathsf{Q}} \\ \widehat{e}^{\ell_S}_{\mathsf{Q}}\end{matrix} \middle\| \begin{matrix} d \\ e \end{matrix}\right] \leqslant \frac{2\mathtt{KL} + \ln\left[\frac{\xi_n + n}{\delta}\right]}{n} \right\}. \end{aligned}$$

This bound holds only for $\ell = \ell_S$.

Note that all the previous bounds do not take into account the variance of the error. Recently [260], a new bound has been derived, which takes the variance into account, and in many natural situations their PAC-Bayes-Empirical-Bernstein inequality can be much tighter than the bound of Eq. (8.15).

For any probability distribution P over $\mathcal{F}$, chosen before seeing $\mathcal{D}_n$, with probability at least $(1-2\delta)$, $\forall \mathsf{Q}$, and for any $c_1, c_2 > 1$, we have that [260]

$$L^\ell(G_\mathsf{Q}) \leqslant \widehat{L}^\ell(G_\mathsf{Q}) + (1+c_1)\sqrt{\frac{(e-2)\nu_3\left[\mathtt{KL} + \ln\left[\frac{2\nu_1}{\delta}\right]\right]}{n}}, \tag{8.25}$$

when $\sqrt{[\mathtt{KL}[\mathsf{Q}||\mathsf{P}] + \ln[2\nu_1/\delta]]/[(e-2)\nu_3]} \leqslant \sqrt{n}$, otherwise

$$L^\ell(G_\mathsf{Q}) \leqslant \widehat{L}^\ell(G_\mathsf{Q}) + 2\frac{\mathtt{KL} + \ln\left[\frac{2\nu_1}{\delta}\right]}{n}, \tag{8.26}$$

where

$$\nu_1 = \left\lceil \frac{1}{\ln(c_1)} \ln\left(\sqrt{\frac{(e-2)n}{4\ln\left(\frac{1}{\delta}\right)}}\right)\right\rceil, \tag{8.27}$$

$$\nu_2 = \left\lceil \frac{1}{\ln(c_2)} \ln\left(\frac{1}{2}\sqrt{\frac{n-1}{\ln\left(\frac{1}{\delta}\right)}+1}+\frac{1}{2}\right)\right\rceil, \tag{8.28}$$

$$\nu_3 = \widehat{V}^\ell(G_\mathsf{Q}) + (1+c_2)\sqrt{\frac{\widehat{V}^\ell(G_\mathsf{Q})\left[\mathtt{KL} + \ln\left[\frac{\nu_2}{\delta}\right]\right]}{2(n-1)}} + \frac{2c_2\left[\mathtt{KL} + \ln\left[\frac{\nu_2}{\delta}\right]\right]}{n-1}. \tag{8.29}$$

Note that all the bounds involve the KLD between P and Q and for this reason the choice of P and Q in the PAC-Bayes Theory can be critical. From one side Q should fit our observations, but from another side Q should be close to P, in order to minimize the KLD term. The milestone result of [55], later extended by [160], proposes to use a Boltzmann prior distribution P which depends on the data generating distribution μ. In particular, let us suppose that the density function associated to the prior P is:

$$p(f) = c_p e^{-\gamma L^\ell(f)}, \tag{8.30}$$

where $\gamma \in [0, \infty)$ and $1/c_p = \int_\mathcal{F} e^{-\gamma L^\ell(f)} df$ is a normalization term. Moreover, let us suppose that the density function associated to the posterior Q is:

$$q(f) = c_q e^{-\gamma \widehat{L}^\ell(f)}, \tag{8.31}$$

where $1/c_q = \int_\mathcal{F} e^{-\gamma \widehat{L}^\ell(f)} df$ is a normalization term. Hence, we give more importance to functions with small risk.

Note that, in order to sample $f \in \mathcal{F}$ according to this particular Q, there are two main cases. In the first case the cardinality of $\mathcal{F}$ is finite and reasonably small to compute exactly $p(f)$. In the second case $\mathcal{F}$ contains too many functions (or even an infinite number), and consequently we have to resort to a subsampling of $\mathcal{F}$ via Monte Carlo search in order to make the problem treatable and then compute $p(f)$. Note that this last approach may produce numerical problems when γ is large.

Based on the previous definitions the following result has been derived.

Given P defined in Eq. (8.30) and Q defined in Eq. (8.31), with probability at least $(1-\delta)$, we have that [160]

$$\mathtt{KL}[\mathsf{Q}||\mathsf{P}] \leqslant \mathtt{KL}_1(\gamma,\delta,n) = \frac{\gamma^2}{2n} + \gamma\sqrt{\frac{2\ln\left[\frac{\xi_n}{\delta}\right]}{n}}. \tag{8.32}$$

Consequently, with probability at least $(1-2\delta)$, we have that [160]

$$\mathtt{kl}\left[\widehat{L}^\ell(G_\mathsf{Q})||L^\ell(G_\mathsf{Q})\right] \leqslant \frac{\mathtt{KL}_1(\gamma,\delta,n)+\ln\left[\frac{\xi_n}{\delta}\right]}{n}. \tag{8.33}$$

The loss ℓ that we use for P and Q can be different from the one that we use for $\widehat{L}^\ell(G_\mathsf{Q})$ and $L^\ell(G_\mathsf{Q})$.

The bound of Eq. (8.32) has been recently further improved [193] by exploiting the Clopper-Pearson bound [67].

Given P defined in Eq. (8.30) and Q defined in Eq. (8.31), with probability at least $(1-2\delta)$, we have that [193]

$$\begin{aligned}
&\mathtt{KL}[\mathsf{Q}||\mathsf{P}] \leqslant \mathtt{KL}_2(\gamma,\delta,n)\\
&= \frac{\gamma^2}{4n} - \gamma\mathtt{Q}\left[\delta;\frac{n}{2},\frac{n}{2}+1\right] + \frac{\gamma}{2} + \gamma\sqrt{\frac{\ln\left[\frac{\xi_n}{\delta}\right]}{2n} + \frac{\gamma^2}{16n^2} - \frac{\gamma\mathtt{Q}\left[\delta;\frac{n}{2},\frac{n}{2}+1\right]}{2n} + \frac{\gamma}{4n}}\\
&\leqslant \mathtt{KL}_1(\gamma,2\delta,n),
\end{aligned} \tag{8.34}$$

where $\mathtt{Q}[t;v,w]$ is the t-th quantile of the Beta distribution with shape parameters v and w. Moreover, with probability at least $(1-3\delta)$, we have that:

$$\mathtt{kl}\left[\widehat{L}^\ell(G_\mathsf{Q})||L^\ell(G_\mathsf{Q})\right] \leqslant \frac{\mathtt{KL}_2(\gamma,\delta,n)+\ln\left[\frac{\xi_n}{\delta}\right]}{n}. \tag{8.35}$$

The loss ℓ that we use for P and Q can be different from the one that we use for $\widehat{L}^\ell(G_\mathsf{Q})$ and $L^\ell(G_\mathsf{Q})$. If the losses used to define P, Q, $\widehat{L}^\ell(G_\mathsf{Q})$ and $L^\ell(G_\mathsf{Q})$ are the same, the bound of Eq. (8.35) can be further improved and, with probability at least $(1-2\delta)$, we have that:

$$\begin{aligned}
&\mathtt{kl}\left[\widehat{L}^\ell(G_\mathsf{Q})||L^\ell(G_\mathsf{Q})\right]\\
&\leqslant \frac{\gamma\left|L^\ell(G_\mathsf{Q})-\widehat{L}^\ell(G_\mathsf{Q})\right|}{n} + \frac{-\gamma\mathtt{Q}\left[\delta;\frac{n}{2},\frac{n}{2}+1\right]+\frac{\gamma}{2}+\ln\left[\frac{\xi_n}{\delta}\right]}{n}.
\end{aligned} \tag{8.36}$$

Note that the main result of the bound of Eq. (8.34) is that $\mathtt{KL}_2(\gamma, \delta, n) \leqslant \mathtt{KL}_1(\gamma, 2\delta, n)$. The results of the bound of Eq. (8.34) can be plugged in any of the bounds of Eqns. (8.15), (8.23), (8.25), and (8.20). In this way we obtain the sets $\mathcal{I}_0^\ell(\delta, \mathsf{P}, \mathsf{Q}, n)$, $\mathcal{I}_1(\delta, \mathsf{P}, \mathsf{Q}, m)$ and $\mathcal{I}_2^\ell(\delta, \mathsf{P}, \mathsf{Q}, n)$, when P and Q defined in this section according to [55, 160], are adopted. Then, the bounds on the risk of the GC and BC can be derived.

Finally, we would like to underline that using two different kinds of losses (one for P and Q and another one for $\widehat{L}^\ell(G_\mathsf{Q})$ and $L^\ell(G_\mathsf{Q})$) can be, in some cases, very useful. In fact, when one has to build a classifier with few high dimensional data [11], using ℓ_S during the learning phase, namely choosing the prior, instead of ℓ_H that we would like to actually minimize, can led to better generalization performances since we can better control our class of functions [30, 49, 225, 265].

A straightforward consequence of the bounds of Eqns. (8.34) and (8.19) is the following bound, which improves all the results reported in [160].

Given P defined in Eq. (8.30) and Q defined in Eq. (8.31), with probability at least $(1 - p\delta)$, we have that:

$$L^\ell(G_\mathsf{Q}) \in \left[\inf \mathcal{I}_3^\ell(\delta, \mathsf{P}, \mathsf{Q}, n), \sup \mathcal{I}_3^\ell(\delta, \mathsf{P}, \mathsf{Q}, n)\right], \tag{8.37}$$

$$L^\ell(B_\mathsf{Q}) \leqslant 2 \min\left[1/2, \sup \mathcal{I}_3^\ell(\delta, \mathsf{P}, \mathsf{Q}, n)\right]. \tag{8.38}$$

Note that if the loss exploited in P and Q is the same as the one used in $L^\ell(G_\mathsf{Q})$ and $\widehat{L}^\ell(G_\mathsf{Q})$, then $p = 2$ and

$$\begin{aligned} I_3^\ell(\delta, \mathsf{Q}, \mathcal{D}_n) = \{ r :\ & r \in [0,1], \\ \mathtt{kl}\left[\widehat{L}^\ell(G_\mathsf{Q}) || r\right] \leqslant & \frac{\gamma \left| r - \widehat{L}^\ell(G_\mathsf{Q}) \right|}{n} + \frac{-\gamma \mathsf{Q}\left[\delta; \frac{n}{2}, \frac{n}{2}+1\right] + \frac{\gamma}{2} + \ln\left[\frac{\xi_n}{\delta}\right]}{n} \Bigg\}. \end{aligned} \tag{8.39}$$

Otherwise $p = 3$ and:

$$\begin{aligned} & I_3^\ell(\delta, \mathsf{Q}, \mathcal{D}_n) \\ & = \left\{ r :\ r \in [0,1], \mathtt{kl}\left[\widehat{L}^\ell(G_\mathsf{Q}) || r\right] \leqslant \frac{\mathtt{KL}_2(\gamma, \delta, n) + \ln\left[\frac{\xi_n}{\delta}\right]}{n} \right\}. \end{aligned} \tag{8.40}$$

Analogously, it is possible to plug the result of the bound of (8.34) into the bound of Eq. (8.25) in order to improve this last result in the cases when the bound of Eq. (8.25) is sharper with respect to the bound of (8.34).

Finally note that all the above mentioned bounds that take into account only empirical quantities can be used both for MS and EE purposes by exploiting

Algorithm 5: PAC-Bayes Theory: MS and EE Strategy for the Gibbs Classifier (for the Bayes Classifier it is analogous).

```
Input: $\{(\mathsf{P},\mathcal{F})_1,\cdots,(\mathsf{P},\mathcal{F})_{n_c}\}$, $\mathcal{D}_n$, and $\delta$
Output: Optimal Model $\mathsf{Q}^*$ and the estimated generalization error of the
        associated Gibbs Classifier $L(G_{\mathsf{Q}^*})$
$L^*_{\mathrm{MS}} = +\infty$;
for $(\mathsf{P}^*,\mathcal{F}^*) \in \{(\mathsf{P},\mathcal{F})_1,\cdots,(\mathsf{P},\mathcal{F})_{n_c}\}$ do
    Choose $\mathsf{Q}$ with the preferred method;
    $L_{\mathrm{MS}} = \varDelta(\mathcal{D}_n,\mathsf{Q},\mathsf{P},\delta)$;
    if $L^*_{MS} > L_{MS}$ then
        $L^*_{\mathrm{MS}} = L_{\mathrm{MS}}$;
        $\mathsf{Q}^* = \mathsf{Q}$;
        $L(G_{\mathsf{Q}^*}) = \varDelta\left(\mathcal{D}_n,\mathsf{Q},\mathsf{P},\frac{\delta}{n_c}\right)$;
```

the approach described in the preliminaries of this book [104, 202, 203]. In fact all the PAC-Bayes based bounds have the following form:

$$\mathbb{P}_{\mathcal{D}_n}\left\{L(G_{\mathsf{Q}}) \leqslant \varDelta(\mathcal{D}_n,\mathsf{Q},\mathsf{P},\delta)\right\} \geqslant 1-\delta, \tag{8.41}$$

$$\mathbb{P}_{\mathcal{D}_n}\left\{L(B_{\mathsf{Q}}) \leqslant \varDelta(\mathcal{D}_n,\mathsf{Q},\mathsf{P},\delta)\right\} \geqslant 1-\delta. \tag{8.42}$$

Then if we want to choose $(\mathsf{P}^*,\mathcal{F}^*) \in \{(\mathsf{P},\mathcal{F})_1,\cdots,(\mathsf{P},\mathcal{F})_{n_c}\}$ and whatever Q^* we want, namely perform the MS phase, and estimate the generalization performance of G_{Q^*} and B_{Q^*}, namely perform the EE phase, we have to follow the procedure summarized in Algorithm 5. Note that the generalization of the final model is bounded by

$$\begin{aligned}&\mathbb{P}_{\mathcal{D}_n}\left\{L(G_{\mathsf{Q}^*})\leqslant\varDelta\left(\mathcal{D}_n,\mathsf{Q}^*,\mathsf{P}^*,\frac{\delta}{n_c}\right)\right\} \geqslant 1-\delta,\\ &\qquad\forall(\mathsf{P}^*,\mathcal{F}^*) \in \{(\mathsf{P},\mathcal{F})_1,\cdots,(\mathsf{P},\mathcal{F})_{n_c}\},\quad \forall\mathsf{Q}^*,\end{aligned} \tag{8.43}$$

$$\begin{aligned}&\mathbb{P}_{\mathcal{D}_n}\left\{L(B_{\mathsf{Q}^*})\leqslant\varDelta\left(\mathcal{D}_n,\mathsf{Q}^*,\mathsf{P}^*,\frac{\delta}{n_c}\right)\right\} \geqslant 1-\delta,\\ &\qquad\forall(\mathsf{P}^*,\mathcal{F}^*) \in \{(\mathsf{P},\mathcal{F})_1,\cdots,(\mathsf{P},\mathcal{F})_{n_c}\},\quad \forall\mathsf{Q}^*.\end{aligned} \tag{8.44}$$

with probability $(1-\delta)$, since we have applied the Bonferroni correction [45] over the n_c choices for the Prior and space of functions. Note that $\varDelta$ depends on the particular $\mathscr{A}$.

9

Differential Privacy Theory

The problem of learning from data while preserving the privacy of individual observations has a long history and spans over multiple disciplines [87, 108, 270]. One way to preserve privacy is to corrupt the learning procedure with noise without destroying the information that we want to extract. Differential Privacy (DP) is one of the most powerful tools in this context [82, 87]. DP addresses the apparently self-contradictory problem of keeping private the information about an individual observation while learning useful information about a population. In particular, a procedure is DP if and only if its output is almost independent from any of the individual observations (which is similar, is some sense, to the concept of Stability). In other words, the probability of a certain output should not change significantly if one individual is present or not, where the probabilities are taken over the noise introduced by the procedure. In the last years, DP has been deeply studied from a theoretical point of view [56, 86, 88, 124, 129, 144, 158, 191, 223, 247, 250, 273] and exploited to develop new learning strategies for solving real world problems [41, 41, 57, 58, 99, 122, 123, 245, 272].

DP allowed to reach a milestone result by connecting the field of privacy preserving data analysis and the generalization capability of a learning algorithm. From one side it proved that a learning algorithm which shows DP properties also generalizes [84]. From the other side, if an algorithm is not DP, it allowed to state the conditions under which a hold out set can be reused without risk of false discovery through a DP procedure called Thresholdout [83, 85].

As we will see later DP can be applied to probabilistic (randomized) algorithms, we need to know the space of rules from which the algorithms choose the final rule and the observed data must be sampled independently and identically distributed (i.i.d.). Even if very powerful, DP is a quite young and immature field of research which still needs to be carefully studied and deeply understood.

In order to present the DP Theory, we first recall some preliminary definitions [84, 87, 265]. Let us consider an input space $\mathcal{X}$ and an output space $\mathcal{Y}$. We indicate with $\mu_{\mathcal{X}}$, $\mu_{\mathcal{Y}}$, and $\mu_{\mathcal{Z}}$ respectively the distributions over $\mathcal{X}$, $\mathcal{Y}$, and the cartesian product between the input and the output space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. From $\mathcal{Z}$ we observe a series of n i.i.d. samples $s = \{z_1, \cdots, z_n\} = \{(x_1, y_1), \cdots, (x_n, y_n)\}$, where $\forall i \in \mathcal{I}_n = \{1, 2, 3, \cdots, n\}$ we have $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$, and $z_i \in \mathcal{Z}$. Moreover, $\boldsymbol{Z}$ is a random variable sampled from $\mathcal{Z}$ according to $\mu_{\mathcal{Z}}$, whereas s is a dataset inside the space of all the possible datasets $\mathcal{S} = \mathcal{Z}^n$ and $\mathfrak{P}_{\mathcal{S}}$ is the distribution of probability generated by $\mu_{\mathcal{Z}}$ over $\mathcal{S}$. Analogously to $\boldsymbol{Z}$, $\boldsymbol{S}$ is a random variable sampled from $\mathcal{S}$ according to $\mathfrak{P}_{\mathcal{S}}$. We denote with $\dot{s}$ the neighborhood dataset of s such that $s = \{z_1, \cdots, z_{i-1}, \dot{z}_i, z_{i+1}, \cdots, z_n\}$, where i may assume any value in $\mathcal{I}_n$ and $\dot{z}_i$ i.i.d. with z_i. We denote with $\breve{\mathcal{S}}$ a subset of the space of datasets $\mathcal{S}$: $\breve{\mathcal{S}} \subseteq \mathcal{S}$. Let us define with $f : \mathcal{X} \to \mathcal{Y}$ a function in a space $\mathcal{F}$ of all the possible functions and $\breve{\mathcal{F}} \subseteq \mathcal{F}$. The functions (or rules) in $\mathcal{F}$ can be deterministic or probabilistic. A randomized algorithm $\mathscr{A} : \mathcal{S} \to \mathcal{F}$ maps a dataset $s \in \mathcal{S}$ in a function $f \in \mathcal{F}$ in a nondeterministic way that can be encapsulated in a distribution $\mu_{\mathscr{A}}$ over $\mathcal{F}$ given $s \in \mathcal{S}$. We also define an operator $\breve{D}$ which maps a function $f \in \mathcal{F}$ into a subset of all the possible datasets $\breve{\mathcal{S}}$. For example, $\breve{D}$ can be seen as an inverse operator of $\mathscr{A}$ which, given an $f \in \mathcal{F}$, tries to retrieve the datasets $\breve{\mathcal{S}}$ that may have generated f. The accuracy of $f \in \mathcal{F}$ in representing $\mu_{\mathcal{Z}}$ is measured with reference to a loss function $\ell : \mathcal{F} \times \mathcal{Z} \to [0, 1]$. Hence, we can define the true risk of f, namely generalization error, as

$$L(f) = \mathbb{E}_{\boldsymbol{Z}} \ell(f, \boldsymbol{Z}), \tag{9.1}$$

together with its variance

$$V(f) = \mathbb{E}_{\boldsymbol{Z}} [\ell(f, \boldsymbol{Z}) - L(f)]^2. \tag{9.2}$$

Since $\mu_{\mathcal{Z}}$ is unknown, $L(f)$ and $V(f)$ cannot be computed. Therefore, we have to resort to their empirical estimators, respectively the empirical error [265]

$$\widehat{L}_n^s(f) = {}^1\!/\!{}_n \sum_{i=1}^{n} \ell(f, z_i), \tag{9.3}$$

and the empirical variance [175]

$$\widehat{V}_n^s(f) = {}^1\!/\!_{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} [\ell(f, z_i) - \ell(f, z_j)]^2. \tag{9.4}$$

Let us recall the definition of DP [87]: a randomized algorithm $\mathscr{A}$ is (ϵ, δ)-Differentially Private if

$$\mathbb{P}_{\mathscr{A}} \left\{ \mathscr{A}(s) \in \breve{\mathcal{F}} \right\} \leqslant e^{\epsilon} \mathbb{P}_{\mathscr{A}} \left\{ \mathscr{A}(\dot{s}) \in \breve{\mathcal{F}} \right\} + \delta, \quad \forall \breve{\mathcal{F}} \subseteq \mathcal{F}, \quad \forall s \in \mathcal{S}. \tag{9.5}$$

Note that in this book we will only deal with $(\epsilon, 0)$-Differentially Private algorithms that we will denote as ϵ-DP for brevity.

Since we are dealing with ϵ-DP algorithms it is useful to give an alternative simpler and more intuitive definition of ϵ-DP with respect to the one of Eq. (9.5). Basically this new definition says that if the probability of choosing a function does not change too much if the algorithm is fed with a dataset s or with its neighborhood $\dot{s}$ then the algorithm is private [202]: a randomized algorithm $\mathscr{A}$ is ϵ-DP if

$$\frac{\mathbb{P}_{\mathscr{A}}\{\mathscr{A}(s) = f\}}{\mathbb{P}_{\mathscr{A}}\{\mathscr{A}(\dot{s}) = f\}} \leqslant e^{\epsilon}, \quad \forall f \in \mathcal{F}, \quad \forall s \in \mathcal{S}. \tag{9.6}$$

The milestone result in DP Theory [84] shows that an ϵ-DP algorithm generalizes. In particular two main results are derived. The first one [84] is very general and shows that if a function $\breve{D}(f)$ is defined for each element $f \in \mathcal{F}$ and the probability that $\boldsymbol{S} \in \breve{D}(f)$ is small, then the probability remains small if f is chosen based on $\boldsymbol{S}$ and $\mathscr{A}$. In other words the probability that $\boldsymbol{S} \in \breve{D}(\mathscr{A}(\boldsymbol{S}))$ remains small[1].

This first result [84] can be formalized as follows. Let $\mathscr{A}$ be an ϵ-DP. Let us suppose that $\mathbb{P}_{\boldsymbol{S}}\{\boldsymbol{S} \in \breve{D}(f)\} \leqslant \beta, \quad \forall f \in \mathcal{F}$. Then

$$\epsilon \leqslant \sqrt{{}^{\ln({}^1\!/\!_\beta)}\!/\!_{2n}} \quad \rightarrow \quad \mathbb{P}_{\boldsymbol{S},\boldsymbol{F}}\{\boldsymbol{S} \in \breve{D}(\boldsymbol{F})\} \leqslant 3\sqrt{\beta}. \tag{9.7}$$

The second result, which builds upon the first one, shows that the empirical error of a function chosen with an ϵ-DP algorithm is concentrated around its generalization error [202]. In particular, let $\mathscr{A}$ be an ϵ-DP, then for any $t > 0$

$$\epsilon \leqslant \sqrt{t^2 - {}^{\ln(2)}\!/\!_{2n}} \quad \rightarrow \quad \mathbb{P}_{\boldsymbol{S},\boldsymbol{F}} \left\{ |L(\boldsymbol{F}) - \widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F})| \geqslant t \right\} \leqslant 3\sqrt{2}e^{-nt^2}. \tag{9.8}$$

This result can be reformulated in a more convenient expression, which is more suited for the subsequent analysis [202]. Let $\mathscr{A}$ be an ϵ-DP, then we can state that

[1] From now on with a little abuse of notation we will identify $\boldsymbol{F} = \mathscr{A}(\boldsymbol{S})$.

$$\mathbb{P}_{\boldsymbol{S},\boldsymbol{F}}\left\{|L(\boldsymbol{F})-\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F})| \geqslant \epsilon+\sqrt{1/n}\right\} \leqslant 3e^{-n\epsilon^2}. \tag{9.9}$$

The limitation of Bounds (9.8) and (9.9) is the slow convergence rate (i.e. $O(\sqrt{1/n})$). When the empirical error is small we would like to retrieve a Chernoff-type result [65]. Instead, when the variance is small, a Bernstein-type or Bennet-type bound would be preferred [35, 38].

By exploiting the result of Eq. (9.7) and the Chernoff bound [65] it is possible to prove the following results, which improve the rate of convergence of the bound of Eq. (9.8) when $\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F})$ is small [202]. Let $\mathscr{A}$ be an ϵ-DP, then for any $t > 0$

$$\begin{aligned}
&\epsilon \leqslant \sqrt{t^2 - \ln(2)/2n} \\
&\rightarrow \quad \mathbb{P}_{\boldsymbol{S},\boldsymbol{F}}\left\{|L(\boldsymbol{F})-\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F})| \geqslant \sqrt{6L(\boldsymbol{F})}t\right\} \leqslant 3\sqrt{2}e^{-nt^2}.
\end{aligned} \tag{9.10}$$

Note that the rate of convergence of the bound of Eq. (9.10) can be faster with respect to the one of Eq. (9.8). In fact when $\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F}) \to 0$ the convergence of the bound can reach $O\left(1/n\right)$. This is made more evident in the following reformulation [202] of the bound of Eq. (9.10): let $\mathscr{A}$ be an ϵ-DP, then we can state that

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{S},\boldsymbol{F}}\left\{|L(\boldsymbol{F})-\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F})| \geqslant \sqrt{6\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F})}\left(\epsilon+\sqrt{1/n}\right)+6\left(\epsilon^2+1/n\right)\right\} \\
&\qquad \leqslant 3e^{-n\epsilon^2}.
\end{aligned} \tag{9.11}$$

The bounds of Eqns. (9.10) and (9.11) can be further improved when $\ell : \mathcal{F} \times \mathcal{Z} \to \{0,1\}$ by exploiting the exact confidence interval for binomial tails [67]. Analogously to the bound of Eq. (9.10) which exploits the Chernoff bound [65], it is possible to prove the following bound based on the exact confidence interval for binomial tails [67]. Let $\mathscr{A}$ be an ϵ-DP and $\ell : \mathcal{F} \times \mathcal{Z} \to \{0,1\}$, then [202]

$$\begin{aligned}
&\epsilon \leqslant \sqrt{\ln\left(1/2\delta\right)/2n} \\
&\rightarrow \quad \mathbb{P}_{\boldsymbol{S},\boldsymbol{F}}\left\{L(\boldsymbol{F}) \leqslant \mathtt{Q}[\delta; n\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F}), n-n\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F})+1] \vee \right. \\
&\qquad\qquad \left. L(\boldsymbol{F}) \geqslant \mathtt{Q}[1-\delta; n\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F})+1, n-n\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F})]\right\} \leqslant 3\sqrt{2\delta},
\end{aligned} \tag{9.12}$$

where $\mathtt{Q}[p; v, w]$ is the p-th quantile of the Beta distribution with shape parameters v and w.

Analogously to what has been done for the bound of Eq. (9.10), the bound of Eq. (9.12) can be easily reformulated as follows. Let $\mathscr{A}$ be an ϵ-DP and $\ell : \mathcal{F} \times \mathcal{Z} \to \{0,1\}$, then we can state that

$$\mathbb{P}_{\boldsymbol{S},\boldsymbol{F}}\Big\{L(\boldsymbol{F}) \leqslant \mathtt{Q}[{}^{e^{-2n\epsilon^2}}\!/_2; n\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F}), n - n\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F}) + 1] \vee$$
$$L(\boldsymbol{F}) \geqslant \mathtt{Q}[1 - {}^{e^{-2n\epsilon^2}}\!/_2; n\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F}) + 1, n - n\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F})]\Big\} \leqslant 3e^{-n\epsilon^2}. \quad (9.13)$$

By exploiting a recent result on the Clopper-Pearson bound [59, 67, 193], it is possible to extend the results of Eq. (9.12) and (9.12) to the general case of $\ell : \mathcal{F} \times \mathcal{Z} \rightarrow [0, 1]$. Let u be a random variable uniformly distributed over $[0, 1]$ and let $\{u_1, \cdots, u_{n_t}\}$ be n_t variables sampled i.i.d. from u. Let $\mathscr{A}$ be an ϵ-DP, then

$$\epsilon \leqslant \sqrt{{}^{\ln({}^1\!/_{2\delta})}\!/_{2n}}$$
$$\rightarrow \quad \mathbb{P}_{\boldsymbol{S},\boldsymbol{F}}\Big\{L(\boldsymbol{F}) \leqslant \mathtt{Q}[\delta; n\widehat{L}_n^{\boldsymbol{S},u}(\boldsymbol{F}), n - n\widehat{L}_n^{\boldsymbol{S},u}(\boldsymbol{F}) + 1] \vee \quad (9.14)$$
$$L(\boldsymbol{F}) \geqslant \mathtt{Q}[1 - \delta; n\widehat{L}_n^{\boldsymbol{S},u}(\boldsymbol{F}) + 1, n - n\widehat{L}_n^{\boldsymbol{S},u}(\boldsymbol{F})]\Big\} \leqslant 3\sqrt{2\delta},$$
$$\rightarrow \quad \mathbb{P}_{\boldsymbol{S},\boldsymbol{F}}\Big\{L(\boldsymbol{F}) \leqslant \mathtt{Q}[{}^{e^{-2n\epsilon^2}}\!/_2; n\widehat{L}_n^{\boldsymbol{S},u}(\boldsymbol{F}), n - n\widehat{L}_n^{\boldsymbol{S},u}(\boldsymbol{F}) + 1] \vee \quad (9.15)$$
$$L(\boldsymbol{F}) \geqslant \mathtt{Q}[1 - {}^{e^{-2n\epsilon^2}}\!/_2; n\widehat{L}_n^{\boldsymbol{S},u}(\boldsymbol{F}) + 1, n - n\widehat{L}_n^{\boldsymbol{S},u}(\boldsymbol{F})]\Big\} \leqslant 3e^{-n\epsilon^2},$$

where $\widehat{L}_n^{s,u}(f) = {}^1\!/_n \sum_{i=1}^n [\ell(f, z_i) \geqslant u_i]$ (the Iverson bracket notation [121] is exploited).

The problem of the bounds of Eq. (9.10), (9.11), (9.12), (9.13), (9.14), and (9.15) is that, if $\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F})$ is large, the rate of convergence is always $O\left(\sqrt{{}^1\!/_n}\right)$. In order to improve this result we have to take into account the variance of $\boldsymbol{F}$. For this purpose we can use Bernstein or Bennet type inequalities [202]. Analogously to bound of Eq. (9.10), by exploiting the result of Eq. (9.7) and the Bernstein bound [175] it is possible to prove the following result, which improves both the bounds of Eqns. (9.8) and (9.10) when $\widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F})$ is large and $\widehat{V}_n^{\boldsymbol{S}}(\boldsymbol{F})$ is small. Let $\mathscr{A}$ be an ϵ-DP, then for any $t > 0$

$$\epsilon \leqslant \sqrt{t^2 - {}^{\ln(3)}\!/_{2n}} \quad (9.16)$$
$$\rightarrow \quad \mathbb{P}_{\boldsymbol{S},\boldsymbol{F}}\left\{\left|L(\boldsymbol{F}) - \widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F})\right| \geqslant \sqrt{4\widehat{V}_n^{\boldsymbol{S}}(\boldsymbol{F})}t + \frac{14nt^2}{3(n-1)}\right\} \leqslant 3\sqrt{3}e^{-nt^2}.$$

The result of Eq. (9.16) can be reformulated in order to better show the advances with respect to the bounds of Eqns. (9.8) and (9.10). Let $\mathscr{A}$ be an ϵ-DP, then we can state that

$$\mathbb{P}_{\boldsymbol{S},\boldsymbol{F}}\left\{\left|L(\boldsymbol{F}) - \widehat{L}_n^{\boldsymbol{S}}(\boldsymbol{F})\right| \geqslant \sqrt{4\widehat{V}_n^{\boldsymbol{S}}(\boldsymbol{F})}\left(\epsilon + \sqrt{{}^1\!/_n}\right) + \frac{5n}{n-1}\left(\epsilon^2 + {}^1\!/_n\right)\right\}$$
$$\leqslant 3e^{-n\epsilon^2}. \quad (9.17)$$

Algorithm 6: Differential Privacy Theory: MS and EE Strategy.

```
Input: $\{\mathscr{A}_1, \cdots, \mathscr{A}_{n_{\mathscr{A}}}\}$, $s$, and $\delta$
Output: Optimal Model $f^*$ and its estimated generalization error $L(f^*)$
$L^*_{\mathrm{MS}} = +\infty$;
for $\mathscr{A} \in \{\mathscr{A}_1, \cdots, \mathscr{A}_{n_{\mathscr{A}}}\}$ do
    Compute $\epsilon$ for $\mathscr{A}$ ;
    $L_{\mathrm{MS}} = \Delta(s, \mathscr{A}, \epsilon, \delta)$;
    if $L^*_{MS} > L_{MS}$ then
        $L^*_{\mathrm{MS}} = L_{\mathrm{MS}}$;
        $f^* = \mathscr{A}(s)$;
        $L(\mathscr{A}^*(f^*)) = \Delta(s, \mathscr{A}, \epsilon, \frac{\delta}{n_{\mathscr{A}}})$;
```

Obviously, these DP-based bounds can be used both for MS and EE purposes by exploiting the approach described in the preliminaries of this book. In fact all the DP-based bounds have the following form:

$$\mathbb{P}_{\boldsymbol{S},\boldsymbol{F}} \left\{ L(\boldsymbol{F}) \leqslant \Delta(\boldsymbol{S}, \mathscr{A}, \epsilon, \delta) \right\} \geqslant 1 - \delta. \tag{9.18}$$

Then if we want to choose $\mathscr{A}^* \in \{\mathscr{A}_1, \cdots, \mathscr{A}_{n_{\mathscr{A}}}\}$, namely perform the MS phase, and estimate the generalization performance of $f^* = \mathscr{A}^*(s)$, namely perform the EE phase, we have to follow the procedure summarized in Algorithm 6. Note that the generalization of the final model is bounded by

$$\begin{aligned} &\mathbb{P}_{\boldsymbol{S},\boldsymbol{F}^*} \left\{ L(\boldsymbol{F}^*) \leqslant \Delta\left(\boldsymbol{S}, \mathscr{A}^*, \epsilon_{\mathscr{A}^*}, \frac{\delta}{n_{\mathscr{A}}}\right) \right\} \geqslant 1 - \delta, \\ &\forall \mathscr{A}^* \in \{\mathscr{A}_1, \cdots, \mathscr{A}_{n_{\mathscr{A}}}\}, \end{aligned} \tag{9.19}$$

with probability $(1 - \delta)$, since we have applied the Bonferroni correction [45] over the $n_{\mathscr{A}}$ choices for the algorithm. Note that Δ depends on the particular $\mathscr{A}$.

Another important DP-based result is the possibility of using the same hold out set many times (contrarily to what can be done with the resampling methods) in adaptive data analysis by adding some noise over the measured error on the hold out set via Thresholdout algorithm [83, 85]. This is extremely useful for many reasons. The simplest one is that we need less data to assess the performance of our data analysis procedure [85]. The less intuitive one is that, by adding some noise, we reduce the risk of overfitting and false discovery [85, 87]. Moreover, Thresholdout pushed research into a shift of paradigm in approaching the problem of adaptive data analysis by introducing

Algorithm 7: Union Bound for the NAS

```
Input: $s_t, s_h$, and $\mathscr{P}_1, \cdots, \mathscr{P}_m$
Output: $\widehat{L}_n^{s_h}(f_1), \cdots, \widehat{L}_n^{s_h}(f_m)$
for $i \leftarrow 1$ to $m$ do
    $f_i = \mathscr{P}_i\,(s_t)$ and compute $\widehat{L}_n^{s_h}(f_i)$;
```

a new methodology which goes beyond the state-of-the-art EE techniques [10, 85, 147, 195]. Finally, Thresholdout does not require any DP property over the procedure for selecting our models. In particular, we have a training set with no access restrictions for building the model and we have to interact with the hold out set through Thresholdout, which instead is DP, for MS and EE purposes [83]. Basically in the MS phase we will select the model with the smallest generalization error guaranteed by the EE phase.

In the next paragraphs we will show the benefit of using the adaptive data analysis with respect to the conventional non-adaptive one specifically for the problems of MS and EE.

In order to proceed, we first need to recall some preliminary definitions. Let $s_t \in \mathcal{S}$ be a training set and $s_h \in \mathcal{S}$ an hold out set, both of size n and i.i.d. Let Lap(b) be a random variable sampled from a Laplace distribution of mean zero and variance $2b^2$, in other words with probability density function $l(x)$ such that $l(x) = {}^1\!/_{2b}\; e^{-|x|/b}$. In MS and EE we want to create many models f_i with $i \in \mathcal{I}_m$ based on s_t through a procedure $\mathscr{P}_i$ with $i \in \mathcal{I}_m$ and we want to select the best performing one based on s_h. This process can be performed in a non-adaptive setting (NAS) or in the adaptive one (AS).

The NAS setting is the case when the procedures for building the models $\mathscr{P}_i$ with $i \in \mathcal{I}_m$ exploit just the training set.

For the NAS the best approach [202] is to use the Union Bound [45] (or Bonferroni correction), since all the functions $f_i = \mathscr{P}_i(s_t)$ with $i \in \mathcal{I}_m$ are chosen without seeing s_h (see Algorithm 7). This allows to state that, if $\{f_1, \cdots, f_m\}$ are chosen in a NAS [202],

$$\mathbb{P}_{\boldsymbol{S}_h}\left\{\exists i \in \mathcal{I}_m : \left|L(f_i) - \widehat{L}_n^{\boldsymbol{S}_h}(f_i)\right| \geqslant \sqrt{\frac{\ln\left(\frac{2m}{\delta}\right)}{2n}}\right\} \leqslant \delta. \tag{9.20}$$

Instead, the AS is the case when the procedures for building our models $\mathscr{P}_i$ with $i \in \mathcal{I}_m$ exploit both the training set and the performance of $\mathscr{P}_1, \cdots, \mathscr{P}_{i-1}$ over the hold out set.

```
Algorithm 8: Hold out for the AS
  Input: $s_t, s_h$, and $\mathscr{P}_1, \cdots, \mathscr{P}_m$
  Output: $\widehat{L}_n^{s_h^1}(f_1), \cdots, \widehat{L}_n^{s_h^m}(f_m)$
  Split $s_h$ in $s_h^i$ with $i \in \mathcal{I}_m$;
  for $i \leftarrow 1$ to $m$ do
      $f_i = \mathscr{P}_i\left(s_t, \widehat{L}_n^{s_h^1}(f_1), \cdots, \widehat{L}_n^{s_h^{i-1}}(f_{i-1})\right)$ and compute $\widehat{L}_n^{s_h^i}(f_i)$;
```

In the AS, conversely to NAS, f_i with $i \in \mathcal{I}_m$ in general depends on s_h (see Algorithm 8), hence we cannot use the bound of Eq. (9.20). Then, one hold out set for each $\mathscr{P}_i$ with $i \in \mathcal{I}_m$ is needed. Since the data available are just the ones in s_h, we have to split it in m sets of size $^n/_m$ that we will call $s_h^1, \cdots, s_h^m$ (see Algorithm 8). This allows us to use the Hoeffding inequality [118] and state that, if $\{f_1, \cdots, f_m\}$ are chosen in an AS [202]:

$$\mathbb{P}_{\boldsymbol{S}_h^i}\left\{\exists i \in \mathcal{I}_m : \left|L(f_i) - \widehat{L}_n^{\boldsymbol{S}_h^i}(f_i)\right| \geqslant \sqrt{\frac{m \ln\left(\frac{2}{\delta}\right)}{2n}}\right\} \leqslant \delta. \tag{9.21}$$

Note that the bound of Eq. (9.21) has many drawbacks with respect to the bound of Eq. (9.20). The first one (I) is the slower rate of convergence: $O\left(\sqrt{^m/_n}\right)$ for the bound of Eq. (9.21) and $O\left(\sqrt{^{\ln(m)}/_n}\right)$ for the bound of Eq. (9.20). The second one (II) is that the datasets that we use for testing each f_i with $i \in \mathcal{I}_m$ are composed of a smaller number of samples: $^n/_m$ for the bound of Eq. (9.21) and n for the bound of Eq. (9.20).

Obviously the advantage of the bound of Eq. (9.21) is that it can be used in a AS, while the bound of Eq. (9.20) can be used just in the NAS.

The fact that the bound of Eq. (9.21) can be used in the AS can be exploited for solving drawback (I). Let us make an example. Let us suppose that we want to select the best hyperparameter of an algorithm. In the NAS the typical approach is to perform a grid search over g points defined before seeing the hold out set and select the value which gives the best performance in accordance with the bound of Eq. (9.20). Hence, we have to set $m = g$ in the bound of Eq. (9.20). In the AS, instead, we can employ, for example, a bisection method for finding the best value of the hyperparameter in accordance with the bound of Eq. (9.21). Obviously this will result, in general, in a search for a local minima but with the bisection method, in order to explore the same grid, we need to explore a number of values of the hyperparameter which is approximately $\ln(g)$. Hence, we have to set $m \approx \ln(g)$ in the bound

```
Algorithm 9: Thresholdout for the AS
Input: $s_t, s_h, T, \sigma, B$, and $\mathscr{P}_1, \cdots, \mathscr{P}_m$
Output: $a_1, \cdots, a_m$
$\gamma \sim \mathrm{Lap}(2\sigma)$,  $\widehat{T} = T + \gamma$;
for $i \leftarrow 1$ to $m$ do
    if $B < 1$ then
        $a_i = \perp$;
    else
        $f_i = \mathscr{P}_i(s_t, a_1, \cdots, a_{i-1})$,  $\eta \sim \mathrm{Lap}(4\sigma)$;
        if $|\widehat{L}_n^{s_h}(f_i) - \widehat{L}_n^{s_t}(f_i)| \geqslant \widehat{T} + \eta$ then
            $\xi \sim \mathrm{Lap}(\sigma)$,  $\gamma \sim \mathrm{Lap}(2\sigma)$,  $\widehat{T} = T + \gamma$,  $B = B - 1$;
            $a_i = \widehat{L}_n^{s_h}(f_i) + \xi$;
        else
            $a_i = \widehat{L}_n^{s_t}(f_i)$;
```

of Eq. (9.21). Consequently, in this example, the rate of convergence of the two bounds is $O\left(\sqrt{\ln(g)/n}\right)$ in both the bounds of Eqns. (9.20) and (9.21).
The bound of Eq. (9.21) can be further improved, under particular conditions, by using the Thresholdout algorithm reported in Algorithm 9. In Algorithm 9, on the contrary to the classical hold out method, we do not have access to the error on s_h directly but we have to corrupt it with a Laplace noise of variance proportional to σ. Then it is necessary to define a budget B, corrupt the distance between the error on s_t and s_h again with a Laplace noise of variance proportional to σ, and, only when it is above a defined threshold T, it is possible to access the corrupted test set error by consuming the budget B. Otherwise, we just look at the error on s_t. When the budget B is finished, we cannot test any additional function and we signal this event by returning $\perp$. Thanks to the Thresholdout algorithm we can both improve the rate of convergence and solve the drawback (II).

Thresholdout is an advanced combination of two main tools in the DP literature: the Laplace Mechanism and the Sparse Validate techniques [83, 87], and for this reason it can be proved that Thresholdout is a DP algorithm [83] with respect to s_h. More formally, Thresholdout is $2B/\sigma n$-DP with respect to s_h, where $B, \sigma, T > 0$ are user defined parameters of the Thresholdout.

Thresholdout basically perturbs the information over the error on the hold out set such that we are able to maintain the DP property. The fact that Thresholdout is $2B/\sigma n$-DP together with the previous DP-based generalization bounds allows to guarantee that this perturbed error is still concentrated around the generalization error [83]. Let $\beta, t > 0$ and $m \geqslant B > 0$. If $T = {}^{3t}\!/_{4}$, $\sigma = {}^{t}\!/_{96 \ln\left(\frac{4m}{\beta}\right)}$, and $t = 40\sqrt{{}^{B \ln({}^{12m}\!/_{\beta})}\!/_{n}}$, we have that

$$\mathbb{P}_{\boldsymbol{A}_i, \boldsymbol{F}_i} \left\{ \exists i \in \mathcal{I}_m : \boldsymbol{A}_i \neq \perp \wedge |\boldsymbol{A}_i - L(\boldsymbol{F}_i)| \geqslant t \right\} \leqslant \beta, \tag{9.22}$$

where $\boldsymbol{A}_i$ and $\boldsymbol{F}_i$ are the random variables associated respectively to a_i and f_i in Algorithm 9.

It is easy to note that the bound of Eq. (9.22) improves the rate of convergence of the bound of Eq. (9.21) and is also tighter when $B \ln(m) \ll m$. Moreover in the bound of Eq. (9.22) the whole hold out set can be used for testing each f_i with $i \in \mathcal{I}_m$.

Note that, in our hyperparameter search example, we can use the Thresholdout combined with the bisection method and select the best value of the hyperparameter in accordance with the bound of Eq. (9.22). Consequently, we have that Thresholdout can improve over both the bounds of Eqns. (9.20) and (9.21) when $B \ln(\ln(g)) \ll \ln(g)$.

Finally, we want to underline that the results of the bounds of Eqns. (9.20), (9.21), and (9.22) suffer from the problem of having slow convergence rates $O\left(\sqrt{{}^{1}\!/_{n}}\right)$ with respect to the number of samples.

The bounds of Eqns. (9.20) and (9.21) can be easily improved via multiplicative Chernoff inequalities [65, 202] and Bennet [35, 175] or Bernstein [38] inequalities. Then, the same has been done also for the bound of Eq. (9.22). In particular, let $\beta, t > 0$ and $m \geqslant B > 0$. If $T = {}^{3t^2}\!/_{4}$, $\sigma = {}^{t^2}\!/_{96 \ln\left(\frac{4m}{\beta}\right)}$, and $t = 40\sqrt{{}^{B \ln({}^{12m}\!/_{\beta})}\!/_{n}}$, we have that [202]

$$\mathbb{P}_{\boldsymbol{A}_i, \boldsymbol{F}_i} \left\{ \exists i \in \mathcal{I}_m : \boldsymbol{A}_i \neq \perp \wedge |\boldsymbol{A}_i - L(\boldsymbol{F}_i)| \geqslant 30\sqrt{\boldsymbol{A}_i} t + 50t^2 \right\} \leqslant \beta. \tag{9.23}$$

The Bennett version of the bound of Eq. (9.23) is not reported because it is overcomplicated.

10

Conclusions & Further Readings

In this book we tried to provide an intelligible overview of the problems of Model Selection and Error Estimation by focusing on the ideas behind the different Statistical Learning Theory based approaches and simplifying most of the technical aspects with the purpose of making them more accessible and usable in practice.
We have not obviously covered all the methods available in literature but we tried to give our point of view of this topic of research by focusing on methods that do not require any additional knowledge apart from the available data. This book is the result of many years of research and it can be useful both for young researchers in order to approach this field of research and for expert researchers in order to have a starting point for open new path of research.
We encourage the readers to also check other ME and EE promising methods like, for example, the work on learning without concentration [183], the works on learning with the low noise assumption [248], the Occam' Razor Bound [42], the Bayesian approaches [110], and many others which are to many to list in this book.
Finally we hope that we managed to transfer to the reader our passion and dedication for this intriguing, challenging, and multifaceted field of research.

A

Bayesian and Frequentist: an Example

In order to show the differences between the bayesian and frequentist inference let us solve the same problem with the two approaches. Let us suppose that we want to find the probability of a swan to be black since we have observed just white swans. The probability of a swan to be black, can be modeled as a Bernoulli random variable where the probability of success (probability of a swan to be black) is $P \in [0,1]$ and the observed probability (the number of black swan s that we have observed out of a total of n observations) is $\widehat{P} \in [0,1]$.

In the bayesian perspective the parameter P is not fixed but has an unknown probability distribution and, before observing $\widehat{P}$, we have to encapsulate our prior belief about P in the prior distribution. Since to do not have any prior information abut P we use an uninformative prior which assigns equal probabilities to all possibilities:

$$\mathbb{P}\{P = p\} = \begin{cases} 1 & \text{if } p \in [0,1] \\ 0 & \text{otherwise} \end{cases}. \tag{A.1}$$

Now we could open a debate about the quality of this uninformative prior respect to other ones like the Jeffreys prior [116] or the Haldane prior [112] but this is out of the scope of the current presentation and more details can be found in [53, 66]. Since we are dealing with n realization of a Binomial distribution we can compute the likelihood or, in other words, is the probability of observing s black swans in a group of n swans if the probability of observing a black swan P is known:

$$\mathbb{P}\left\{\widehat{P} = \frac{s}{n} \middle| P = p\right\} = \binom{n}{s} p^s (1-p)^{n-s}. \tag{A.2}$$

Thanks to the Bayes' theorem we have that:

$$\mathbb{P}\left\{P = p \middle| \widehat{P} = \frac{s}{n}\right\} = \frac{\mathbb{P}\left\{\widehat{P} = \frac{s}{n} \middle| P = p\right\} \mathbb{P}\{P = p\}}{\mathbb{P}\left\{\widehat{P} = \frac{s}{n}\right\}}. \tag{A.3}$$

Note that, by the axioms of probability and since $P \in [0,1]$, we have that:

$$\int_{-\infty}^{\infty} \mathbb{P}\left\{P = p \middle| \widehat{P} = \frac{s}{n}\right\} dp = \int_{0}^{1} \mathbb{P}\left\{P = p \middle| \widehat{P} = \frac{s}{n}\right\} dp = 1 \tag{A.4}$$

Consequently, we can state that:

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} \frac{\mathbb{P}\left\{\widehat{P} = \frac{s}{n} \middle| P = p\right\} \mathbb{P}\{P = p\}}{\mathbb{P}\left\{\widehat{P} = \frac{s}{n}\right\}} dp \\ &= \frac{1}{\mathbb{P}\left\{\widehat{P} = \frac{s}{n}\right\}} \int_{0}^{1} \mathbb{P}\left\{\widehat{P} = \frac{s}{n} \middle| P = p\right\} \mathbb{P}\{P = p\} dp. \end{aligned} \tag{A.5}$$

Based on this result and on Eqns. (A.1) and (A.2) we can compute the marginal likelihood which is the probability of observing particular exactly s black swans in n observations:

$$\begin{aligned} \mathbb{P}\left\{\widehat{P} = \frac{s}{n}\right\} &= \int_{0}^{1} \mathbb{P}\left\{\widehat{P} = \frac{s}{n} \middle| P = p\right\} \mathbb{P}\{P = p\} dp \\ &= \int_{0}^{1} \binom{n}{s} p^{s}(1-p)^{n-s} dp = \frac{1}{n+1}, \end{aligned} \tag{A.6}$$

where the Euler integral, which states that

$$\int_{0}^{1} x^{a-1}(1-x)^{b-1} dx = \frac{(a-1)!(b-1)!}{(a+b-1)!}, \quad a, b \in \mathbb{N}_0, \tag{A.7}$$

has been exploited. Finally by exploiting Eqns. (A.3), (A.2), (A.1) and (A.6) we can retrieve the probability of the probability of a swan to be black given the fact that we have observed s black swans in a group of n swans:

$$\mathbb{P}\left\{P = p \middle| \widehat{P} = \frac{s}{n}\right\} = (n+1)\binom{n}{s} p^{s}(1-p)^{n-s}. \tag{A.8}$$

From the posterior probability it is possible to retrieve any information about the probability of a swan to be black. For example we can compute the expected value:

$$\begin{aligned}\mathbb{E}_P\{P\} &= \int_{-\infty}^{\infty} p\mathbb{P}\left\{P = p\middle|\widehat{P} = \frac{s}{n}\right\} dp \\ &= \int_0^1 (n+1)\binom{n}{s} p^{s+1}(1-p)^{n-s} = \frac{s+1}{n+2}. \end{aligned} \tag{A.9}$$

which is the Laplace rule of succession [151]. Note, again, that the the Euler integral has been exploited. Another important information about P is it credible interval. In particular we can state that with probability $(1-\delta)$

$$P \in \left[\max\left(0, \frac{s}{n} - \epsilon\right), \min\left(1, \frac{s}{n} + \epsilon\right)\right], \tag{A.10}$$

if the following condition is satisfied:

$$\int_{\max(0,\frac{s}{n}-\epsilon)}^{\min(1,\frac{s}{n}+\epsilon)} (n+1)\binom{n}{s} p^s (1-p)^{n-s} dp \geqslant 1 - \delta. \tag{A.11}$$

Unfortunately an explicit form of the credible interval does not exists. The credible interval gives and information abut the possible values of the parameter of interest and the interval is true with high probability (see Figure A.1). Finally we would like to make two observations:

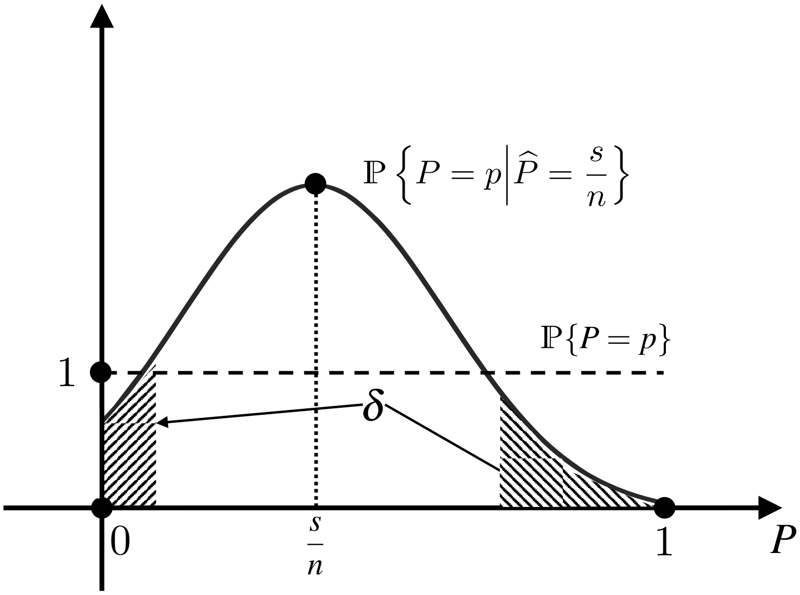
$\mathbb{P}\left\{P = p\middle|\widehat{P} = \frac{s}{n}\right\}$
$\mathbb{P}\{P = p\}$
1
δ
0
$\frac{s}{n}$
1
P

Fig. A.1. Bayesian interpretation of probability: credible interval.

- choosing a different prior would give a different posterior and consequently a different credible interval respect to the one of Eq. (A.10);
- in Eq. (A.10) we decided to report a credible interval symmetric respect to the observed ratio. This is a choice that we made but it could be possible to chose the credible interval which responds to a different criteria (e.g. the tight interval which contains the area $(1-\delta)$ of the total area of the posterior which is 1).

In the frequentist perspective, instead, the parameter P is fixed and unknown while $\widehat{P}$ is a random variable distributed according to a binomial distribution of parameter P. Consequently we can compute the probability of observing s black swans in n observations:

$$\mathbb{P}\left\{\widehat{P} = \frac{s}{n}\right\} = \binom{n}{s} P^s (1-P)^{n-i}. \tag{A.12}$$

We can also compute the probability of observing at most s black swans out n swans

$$\mathbb{P}\left\{\widehat{P} \leqslant \frac{s}{n}\right\} = \sum_{i=0}^{s} \binom{n}{i} P^i (1-P)^{n-i}, \tag{A.13}$$

and the probability of observing at least s black swans out n swans

$$\mathbb{P}\left\{\widehat{P} \geqslant \frac{s}{n}\right\} = \sum_{i=s}^{n} \binom{n}{i} P^i (1-P)^{n-i}. \tag{A.14}$$

Conversely to Bayesian statistic there is no meaning in searching for the expected value of P since P is a deterministic variable. What we can do is to find the confidence interval for the probability of a swan to be black P which is the the range of values of P that with probability $(1-\delta)$ could have generated our particular observation of s black swan out of n. In other words if we repeat again the observations of n swans, we check how many swans are black and we compute the confidence interval, P will fall in that interval at least with probability $(1-\delta)$ (see Figure A.2). For this reason we have to search for the largest and the smallest values of P such that the probability obtaining at most and at least s black swans out n swans is grater than $\delta/2$. Then we can state that if s black swans out n swans are observed than for the probability of a swan to be black the following statement must hold with probability $(1-\delta)$:

$$P \in \left[\begin{array}{l} \min_{p\in[0,1]} \left\{p : \sum_{i=s}^{n} \binom{n}{i} p^i (1-p)^{n-i} \geqslant \frac{\delta}{2}\right\}, \\ \max_{p\in[0,1]} \left\{p : \sum_{i=0}^{s} \binom{n}{i} p^i (1-p)^{n-i} \geqslant \frac{\delta}{2}\right\} \end{array}\right]. \tag{A.15}$$

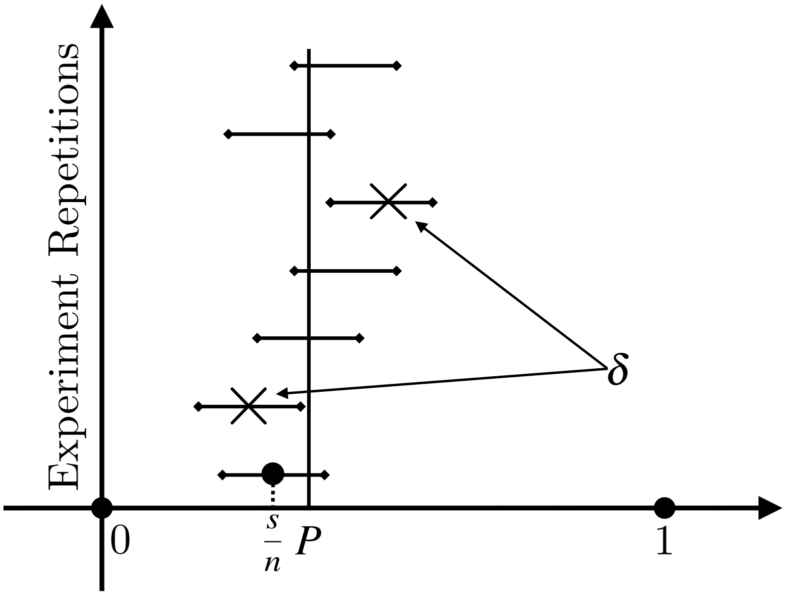

Fig. A.2. Frequentist interpretation of probability: confidence interval.

There are different forms of the quantities in Eq. (A.15), refer to Appendix C for more details.
In order have some insight on the the results of the two approaches let us suppose that we have observed $n = 30$ swans out of which just $s = 3$ are black. Based on the bayesian statistic (see Eq. (A.10)) we have that

$$P \in [0.00, 0.24] \text{ with probability } 0.95, \tag{A.16}$$

while, based on frequentist statistic (see Eq. (A.15)) we have that

$$P \in [0.02, 0.27] \text{ with probability } 0.95. \tag{A.17}$$

The code for retrieving these results of Eqns. (A.16) and (A.17) can be found in Listing A.1.

Listing A.1. Mathematica code for computing the credible and confidence interval of Eqns. (A.16) and (A.17)

```
(* Parameters *)
n = 30;
s = 3;
delta = .05;
(* Bayesians *)
```

```
creint[s_,n_,delta_]:=(
 For[epsilon=0,epsilon<=1,epsilon=epsilon+.001,
 tmp = NIntegrate[(n+1)Binomial[n,s]p^s(1-p)^(n-s),
 {p,Max[0,(s/n)-epsilon],Min[1,(s/n)+epsilon]}];
  If[tmp>=1-delta,
 Return[epsilon];
  ];
  ];
);
epsilon=creint[s,n,delta];
lb=Max[0,(s/n)-epsilon]
lb=Min[1,(s/n)+epsilon]
(* Frequantists *)
lf=Quantile[BetaDistribution[s,n-s+1],(delta/2)]
uf=Quantile[BetaDistribution[s+1,n-s],1-(delta/2)]
(* Clean *)
Clear[n,s,delta,creint,epsilon,lb,ub,lf,uf]
```

Note that the two intervals are not so different, but the two approaches gives to radically different informations. The credible interval is an interval in the domain of a posterior probability distribution. Bayesian intervals treat their bounds as fixed and the estimated parameter as a random variable. A frequentist $(1-\delta)$ confidence interval means that with a large number of repeated samples there is a probability of $(1-\delta)$ of such calculated confidence intervals to include the true value of the parameter. The frequentist confidence intervals treat their bounds as random variables and the parameter as a fixed value. Most of the times to two statistical approaches often behave, like in this case, pretty much the same, but in some cases one can be easier to apply respect to the other.

B

The Wrong Set of Rules could be the Right One

The problem of MS and EE is very tricky since some results are often counter intuitive. Let us make an example. Let us suppose that $x \in \mathcal{X} = [0,1]$ and $y \in \mathcal{Y} = \mathbb{R}$. Let us suppose that $\mathfrak{S}$ is the following one:

$$Y = X^2 + \epsilon, \tag{B.1}$$

where ϵ is a random variable distributed according to a Normal distribution of mean $\mu = 0$ and variance σ and $\mathbb{P}\{X = x\} = 1$ if $x \in [0,1]$ otherwise $\mathbb{P}\{X = x\} = 0$. We define our set of rules $\mathcal{F}_h$ as the polynomials of degree h:

$$f(x) = \sum_{i=0}^{h} a_i x^i, \tag{B.2}$$

where the unknown parameters which defines the final rule are $\{a_1, \dots, a_h\} \in \mathbb{R}^{h+1}$ and $h \in \mathbb{N}_0$ is the hyperparameter which defines the size of the set of rules (the degree of the polynomial). The error of f in approximation $\mathfrak{S}$ is measured according to the square loss function $\ell_2(f, z) = [y - f(x)]^2$. Consequently the generalization error of a function chosen in $\mathcal{F}_h$ can be written as:

$$\begin{aligned} L(f) &= \mathbb{E}_X \mathbb{E}_\epsilon \left\{ \left[X^2 + \epsilon - f(X) \right]^2 \right\} \\ &= \mathbb{E}_X \mathbb{E}_\epsilon \left\{ \left[X^2 + \epsilon - \sum_{i=0}^{h} a_i X^i \right]^2 \right\}, \quad f \in \mathcal{F}_h \end{aligned} \tag{B.3}$$

Consequently the Bayesian rule will be:

$$\begin{aligned} f^{\text{Bayes}} :\ & \arg\inf_f \ L(f) \\ & = \arg\inf_f \ \mathbb{E}_X \mathbb{E}_\epsilon \left\{ [X^2 + \epsilon - f(X)]^2 \right\} \\ & = \arg\inf_f \ \mathbb{E}_X \mathbb{E}_\epsilon \left\{ X^4 + 2X^2\epsilon + \epsilon^2 - 2f(X)[X^2 + \epsilon] + [f(X)]^2 \right\} \\ & = \arg\inf_f \ \mathbb{E}_X \left\{ [f(X)]^2 - 2f(X)[X^2 + \mathbb{E}_\epsilon\{\epsilon\}] + X^4 + \mathbb{E}_\epsilon\left\{\epsilon^2\right\} \right\} \\ & = \arg\inf_f \ \mathbb{E}_X \left\{ [X^2 - f(X)]^2 \right\}. \end{aligned} \tag{B.4}$$

Since $\mathbb{E}_X \left\{ [f(X) - X^2]^2 \right\} \geqslant 0$ and since if we select as function $f(x) = x^2$ we have that $\mathbb{E}_X \left\{ [f(X) - X^2]^2 \right\} = 0$ and we can state that:

$$f^{\text{Bayes}}(x) = x^2. \tag{B.5}$$

Let us remember a property of the Normal distribution in order to compute the generalization error of the Bayesian rule:

$$\mathbb{E}_\epsilon \left\{ (\mu - \epsilon)^p \right\} = \begin{cases} \sigma^p (p-1)!! & \text{if } p \text{ even} \\ 0 & \text{otherwise} \end{cases}, \quad p \in \mathbb{N}. \tag{B.6}$$

Consequently the generalization error of the Bayesian rule can we derived:

$$L(f^{\text{Bayes}}) = \mathbb{E}_X \mathbb{E}_\epsilon \left\{ [X^2 + \epsilon - f^{\text{Bayes}}(X)]^2 \right\} = \mathbb{E}_\epsilon \left\{ \epsilon^2 \right\} = \sigma^2. \tag{B.7}$$

If instead we search for the best approximation of the Bayesian rule in $\mathcal{F}_h$, we have that:

- if $h = 0$

$$\begin{aligned} f_0^*(x) = a_0^* :\ & \arg\inf_{a_0 \in \mathbb{R}} \ \mathbb{E}_X \mathbb{E}_\epsilon \left\{ [X^2 + \epsilon - a_0]^2 \right\} \\ & = \arg\inf_{a_0 \in \mathbb{R}} \ \mathbb{E}_X \left\{ [X^2 - a_0]^2 \right\} \\ & = \arg\inf_{a_0 \in \mathbb{R}} \ \int_0^1 [x^2 - a_0]^2 dx = \arg\inf_{a_0 \in \mathbb{R}} \ a_0^2 - \frac{2}{3} a_0 + \frac{1}{5}. \end{aligned} \tag{B.8}$$

Consequently we have that:

$$f_0^*(x) = \frac{1}{3}. \tag{B.9}$$

The generalization error of f_0^* is:

$$\begin{aligned}
L(f_0^*) &= \mathbb{E}_X\mathbb{E}_\epsilon\left\{\left[X^2+\epsilon-f_0^*(X)\right]^2\right\}\\
&= \mathbb{E}_X\mathbb{E}_\epsilon\left\{\left[X^2+\epsilon-\frac{1}{3}\right]^2\right\}\\
&= \mathbb{E}_X\mathbb{E}_\epsilon\left\{X^4+2\epsilon X^2+\epsilon^2-\frac{2}{3}(X^2+\epsilon)+\frac{1}{9}\right\}\\
&= \int_0^1 x^4-\frac{2}{3}x^2+\sigma^2+\frac{1}{9}dx\\
&= \sigma^2+\frac{4}{45}=\sigma^2+0.0889; \qquad \text{(B.10)}
\end{aligned}$$

- if $h = 1$

$$\begin{aligned}
f_1^*(x) = a_0^*+a_1^*x : \ &\arg\inf_{\{a_0,a_1\}\in\mathbb{R}^2} \ \mathbb{E}_X\mathbb{E}_\epsilon\left\{[X^2+\epsilon-a_0-a_1X]^2\right\} \qquad \text{(B.11)}\\
&= \arg\inf_{\{a_0,a_1\}\in\mathbb{R}^2} \ \mathbb{E}_X\left\{[X^2-a_0-a_1X]^2\right\}\\
&= \arg\inf_{\{a_0,a_1\}\in\mathbb{R}^2} \ \int_0^1[x^2-a_0-a_1x]^2dx\\
&= \arg\inf_{\{a_0,a_1\}\in\mathbb{R}^2} \ a_0^2+\frac{1}{3a_1^2}+a_0a_1-\frac{2}{3}a_0-\frac{1}{2}a_1+\frac{1}{5}.
\end{aligned}$$

By putting the derivative respect to a_0 and a_1 to zero we have that:

$$\begin{cases} 2a_0^*+a_1^* = \frac{2}{3} \\ a_0^*+\frac{2}{3}a_1^* = \frac{1}{2} \end{cases} . \qquad \text{(B.12)}$$

Consequently, by solving the above linear system, we have that:

$$f_1^*(x) = -\frac{1}{6}+x. \qquad \text{(B.13)}$$

The generalization error of f_1^* is:

$$\begin{aligned}
L(f_1^*) &= \mathbb{E}_X\mathbb{E}_\epsilon\left\{\left[X^2+\epsilon-f_1^*(X)\right]^2\right\}\\
&= \mathbb{E}_X\mathbb{E}_\epsilon\left\{\left[X^2+\epsilon+\frac{1}{6}-X\right]^2\right\}\\
&= \mathbb{E}_X\mathbb{E}_\epsilon\left\{X^4+2\epsilon X^2+\epsilon^2+2(X^2+\epsilon)\left(\frac{1}{6}-X\right)+\left(\frac{1}{6}-X\right)^2\right\}\\
&= \int_0^1 x^4+2x^2\left(\frac{1}{6}-x\right)+\left(\frac{1}{6}-x\right)^2+\sigma^2dx\\
&= \sigma^2+\frac{1}{180}=\sigma^2+0.0056; \qquad \text{(B.14)}
\end{aligned}$$

- if $h \geqslant 2$

  $$f_h^*(x) = x^2, \tag{B.15}$$

  since the space of function contains the Bayesian rule. The generalization error of f_h^* with $h \geqslant 2$ is exactly as the one of the Bayesian rule:

  $$L(f_h^*) = L(f^{\text{Bayes}}) = \sigma^2, \quad h \geqslant 2. \tag{B.16}$$

Unfortunately in a real applications $\mathfrak{S}$ is unknown so f^{Bayes} and f_h^* are unknown. For this reason let us suppose to have just some empirical data d_{n_l} available. We have to define an algorithm that tries to select in $\mathcal{F}_h$ a function that is as close as possible to the best approximation of the Bayesian rule in $\mathcal{F}_h$ which is f_h^*. The first idea about $\mathscr{A}_h$ is to search for the function according with the smallest error over the available data:

$$\begin{aligned}\widehat{f}_h^* = \sum_{i=0}^{h} \hat{a}_i^* x^i : \ \arg &\inf_{\{a_0,\cdots,a_h\}\in\mathbb{R}^{h+1}} \frac{1}{n_l}\sum_{i=1}^{n_l}[y_i - f_h(x_i)]^2 \\ = \arg &\inf_{\{a_0,\cdots,a_h\}\in\mathbb{R}^{h+1}} \sum_{i=1}^{n_l}\left[y_i - \sum_{i=0}^{h} a_i x^i\right]^2,\end{aligned} \tag{B.17}$$

and in order to find the best $\{a_0, \cdots, a_h\} \in \mathbb{R}^{h+1}$, we have the solve the following linear system:

$$\begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^h \\ 1 & x_2 & x_2^2 & \cdots & x_2^h \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n_l} & x_{n_l}^2 & \cdots & x_{n_l}^h \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_h \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}. \tag{B.18}$$

Since $\mathbb{E}_{(X_1,Y_1),\cdots,(X_{n_l},Y_{n_l})} \frac{1}{n_l}\sum_{i=1}^{n_l}[Y_i - f_h(X_i)]^2 = L(f_h)$ we have minimized the empirical unbiased estimator of $L(f)$. Note that $\widehat{f}_h^*$ can be seen as a random variable since it depends on d_{n_l} and $\mathscr{A}_h$ has the following characteristics:

- if $h \in \{0, 1\}$ the approximation error is different from zero and the estimation error, in general could be different from zero;
- if $h \geqslant 2$ the approximation error is zero and the estimation error could be different from zero;
- we consider the implementation error equal to zero even if the linear system will be solved with numerical routines.

Since in this case we know $\mathfrak{S}$ we can compute the generalization error of $\widehat{f}_h^*$:

$$\begin{aligned}
L(\widehat{f}_h^*) &= \mathbb{E}_X \mathbb{E}_\epsilon \left\{ \left[X^2 + \epsilon - \sum_{i=0}^{h} \hat{a}_i^* X^i \right]^2 \right\} \\
&= \mathbb{E}_X \mathbb{E}_\epsilon \left\{ X^4 + 2\epsilon X^2 + \epsilon^2 - 2(X^2+\epsilon) \sum_{i=0}^{h} \hat{a}_i^* X^i + \left(\sum_{i=0}^{h} \hat{a}_i^* X^i \right)^2 \right\} \\
&= \int_0^1 x^4 + \sigma^2 - 2x^2 \sum_{i=0}^{h} \hat{a}_i^* x^i + \left(\sum_{i=0}^{h} \hat{a}_i^* x^i \right)^2 dx \\
&= \sigma^2 + \frac{1}{5} + \int_0^1 -2x^2 \sum_{i=0}^{h} \hat{a}_i^* x^i + \left(\sum_{i=0}^{h} \hat{a}_i^* x^i \right)^2 dx, \quad f \in \mathcal{F}_h. \qquad \text{(B.19)}
\end{aligned}$$

Note that the last integral has no close form solution for a general h and $\{a_0, \cdots, a_h\}$, as soon as h becomes fixed the integral has a simple close form solution. In order to understand the quality of the algorithm it does not make sense to check the generalization error of $\widehat{f}_h^*$ for a single d_{n_l} but we have to compute its average behaviour over the whole possible dataset coming from $\mathfrak{S}$:

$$\mathbb{E}_{(X_1,Y_1),\cdots,(X_n,Y_n)} \inf_{\{a_0,\cdots,a_h\} \in \mathbb{R}^h} \sum_{i=1}^{n} \left[Y_i - \sum_{i=0}^{h} a_i X^i \right]^2 \qquad \text{(B.20)}$$

In Table B.1 we report, by varying n_l and h the estimation of the quantity of Eq. (B.20) with $\sigma = 1$ (the code behind Table B.1 is reported in Listing B.1 and it is just a mere application of Eq. (B.20)).

Listing B.1. Matlab code for reproducing Table B.1

```
%% Cleaning
clear
close all
clc

%% Parameters
seed  = 13;
sigma = 1;
nMC   = 10000;
rn    = [3 30 300 3000];
mh    = 4;

%% Definition of the integrals for the generalization error
i0 = @(f,s) s^2+1/5-((2*f(1))/3)+f(1)^2;
i1 = @(f,s) i0(f,s)-f(2)/2+f(1)*f(2)+f(2)^2/3;
i2 = @(f,s) i1(f,s)-(2*f(3))/5+2/3*f(1)*f(3)+1/2*f(2)*f(3)+f(3)^2/5;
i3 = @(f,s) i2(f,s)-f(4)/3+1/2*f(1)*f(4)+2/5*f(2)*f(4)+1/3*f(3)...
```

```
                     *f(4)+f(4)^2/7;
i4 = @(f,s) i3(f,s)-(2*f(5))/7+2/5*f(1)*f(5)+1/3*f(2)*f(5)+2/7*f(3)...
                     *f(5)+1/4*f(4)*f(5)+f(5)^2/9;

%% Set the random seed
s = RandStream('mcg16807','seed',seed);
RandStream.setGlobalStream(s);

%% Compute the average generalization error and its standard deviation
rism = zeros(mh+1,length(rn)); risv = rism;
in = 0;
for n = rn
    in = in + 1;
    for MC = 1:nMC
        x = rand(n,1);
        y = x.^2 + sigma*randn(size(x));
        m = [];
        for i = 0:4
            m = [m, x.^i]; %#ok<AGROW>
            f = m\y;
            tmp = eval(sprintf('i%d(f,sigma)',i));
            rism(i+1,in) = rism(i+1,in) + tmp;
            risv(i+1,in) = risv(i+1,in) + tmp^2;
        end
    end
end
rism = rism/nMC;
risv = sqrt(risv/nMC - rism.^2);

%% Student 95% confidence interval
eps = 2*risv/sqrt(nMC);

%% Make the table
ris = zeros(mh+1,2*length(rn));
ris(:,1:2:end) = rism;
ris(:,2:2:end) = eps;
fprintf(repmat([repmat('%.3e␣pm␣%.3e,␣',1,length(rn)), '\n']...
    ,1,mh+1),ris');
```

h \ n_l	3	30	300	3000
0	**1.4549 ± 0.0104**	1.1239 ± 0.0010	1.0925 ± 0.0001	1.0893 ± 0.0000
1	11.1917 ± 5.5205	**1.0737 ± 0.0014**	1.0124 ± 0.0001	1.0062 ± 0.0000
2	$5 \cdot 10^8 \pm 1 \cdot 10^9$	1.1095 ± 0.0020	**1.0102 ± 0.0002**	**1.0010 ± 0.0000**
3	$5 \cdot 10^8 \pm 1 \cdot 10^9$	1.1705 ± 0.0059	1.0136 ± 0.0002	1.0013 ± 0.0000
4	$5 \cdot 10^8 \pm 1 \cdot 10^9$	1.2728 ± 0.0110	1.0171 ± 0.0002	1.0017 ± 0.0000

Table B.1. No free lunch theorem: the right model is the wrong one.

From Table B.1 it is possible to retrieve many useful insight about the algorithm $\mathscr{A}_h$, in particular, intuitively, one expect $\mathscr{A}_h$ to have a smaller average generalization error for $h \geqslant 2$ since, in this case $\mathcal{F}_h$ contains the Bayesian classifier, so the approximation error is zero and the algorithm is affected just by the estimation error. Vice versa if $h \in \{0, 1\}$ the approximation error is different from zero (and is larger for $h = 0$ respect to when $h = 1$) moreover there is the estimation error which affects the procedure. Intuitively, based on these consideration these considerations, one should always select an h which is grater or equal to two since, thanks to this choice, the corresponding $\mathcal{F}_h$ has no approximation error. Moreover since we do not know a priori the right h (since in real applications just d_{n_l} is available) it is better to choose h as large as possible in order to be sure that the bayesian classifier is inside $\mathcal{F}_h$ and the approximation error is equal to zero. Table B.1 tell us a different story. In particular from Table B.1 it is possible to note a very important fact: the best choice of h does not depends just on $\mathfrak{S}$ but also on the number of observations n_l of $\mathfrak{S}$. Table B.1 tell us that:

- choosing $h > 2$ is never a good option especially if n_l is small;
- choosing $h = 2$ which is exactly the right $\mathcal{F}_h$ is the correct choice only if n_l is large enough;
- choosing $h \in \{0, 1\}$ is the best option if n_l is small. Moreover the smaller is n_l the smaller is the h that you should chose.

This simple example is at the basis of one of the biggest problem of learning based on empirical data. Sometimes and algorithm that choses from the wrong set of models based on d_{n_l} can show better generalization performances respect to the right one because the different sources of error (in this case approximation and estimation) one average can be larger or smaller based on $\mathfrak{S}$ and how many observation n_l of $\mathfrak{S}$ are available. This means that if, for example, we take $n_l = 3$ observations and $h = 0$ the approximation error is big but the estimation error is small enough to compensate the case when $h = 2$ where the approximation error is zero and we have just the estimation error. It is possible to show that also for the implementation error the same phenomena depicted in Table B.1 can be observed: basically an algorithm which is affected by the implementation error could work better then its implementation error-free counterpart. One can refer to [200] for some examples and theoretical analysis of the effect of the implementation error on learning. This effect can be explained with the Occam's Razor principle formulated by William of Occam in the late Middle Ages as a criticism to the science of that time (in particular philosophy) where the theories became more and more elaborate without any corresponding improvement in their predictive power. In its original form, it states that 'Nunquam ponenda est pluralitas sin necesitate', which approximately means 'Entities should not be multiplied beyond necessity'. Today this principle is used to explain the phenomena described above in the sense that there is no meaning in choosing a too complex set of rules $\mathcal{F}_h$, or in other words a too large h, if we do not have enough information available, or in other words if n_l is too small.

C

Properties and Inequalities

In this part of the appendix we recall some properties and inequalities that have been exploited in this monograph.

The Bonferroni Correction

In probability theory, the Bonferroni correction, also known as the Union Bound or Boole's inequality, states that for any finite or countable set of events, the probability that at least one of the events happens is no greater than the sum of the probabilities of the individual events [45]. Formally, for a countable set of events $\{E_1, \cdots, E_n\}$ we have that:

$$\mathbb{P}_{E_1,\cdots,E_n}\left\{\bigcup_{i=1}^{n} E_i\right\} \leqslant \sum_{i=1}^{n} \mathbb{P}_{E_i}\left\{E_i\right\}. \tag{C.1}$$

The inequality can be proved via induction method. In particular, let us first recall some properties [54]. For all $i, j \in \{1, \cdots, n\}$ it is possible to state that:

$$\bigcup_{i=1}^{n} E_i = \left(\bigcup_{i=1, i\neq j}^{n} E_i\right) \cup E_j, \tag{C.2}$$

$$\mathbb{P}_{E_i,E_j}\left\{E_i \cup E_j\right\} = \mathbb{P}_{E_i}\left\{E_i\right\} + \mathbb{P}_{E_j}\left\{E_j\right\} - \mathbb{P}_{E_i,E_j}\left\{E_i \cap E_j\right\}. \tag{C.3}$$

Consequently, we can derive the following bound:

$$
\begin{aligned}
&\mathbb{P}_{E_1,\cdots,E_n}\left\{\bigcup_{i=1}^{n} E_i\right\} = \mathbb{P}_{E_1,\cdots,E_n}\left\{E_1 \cup \left(\bigcup_{i=2}^{n} E_i\right)\right\} \\
&= \mathbb{P}_{E_1}\left\{E_1\right\} + \mathbb{P}_{E_2,\cdots,E_n}\left\{\bigcup_{i=2}^{n} E_i\right\} - \mathbb{P}_{E_1,\cdots,E_n}\left\{E_1 \cap \left(\bigcup_{i=2}^{n} E_i\right)\right\} \\
&\leqslant \mathbb{P}_{E_1}\left\{E_1\right\} + \mathbb{P}_{E_2,\cdots,E_n}\left\{\bigcup_{i=2}^{n} E_i\right\}, \qquad \text{(C.4)}
\end{aligned}
$$

since the probability of an event is always a nonnegative number [138]

$$
\mathbb{P}_{E_1,\cdots,E_n}\left\{E_1 \cap \left(\bigcup_{i=2}^{n} E_i\right)\right\} \geqslant 0. \qquad \text{(C.5)}
$$

By applying the induction hypotheses it is possible to retrieve the Bonferroni correction.

Jensen's Inequality

Jensen's inequality relates the value of a convex function of an integral to the integral of the convex function [126]. In the context of probability theory, it is stated as follows: if X is a random variable which takes values in $\mathcal{X} \subseteq \mathbb{R}$, and $\varphi : \mathbb{R} \to \mathbb{R}$ is a convex function, then

$$
\varphi(\mathbb{E}_X\{X\}) \leqslant \mathbb{E}_X\{\varphi(X)\}. \qquad \text{(C.6)}
$$

The proof in the continuous case is rather technical and can be retrieved from [126]. Here we report the proof for the finite case: $\mathcal{X} = \{x_1, \cdots, x_n\}$. Since φ is convex, if we take $t \in [0, 1]$ we obtain:

$$
\varphi(tx_i + (1-t)x_j) \leqslant t\varphi(x_i) + (1-t)\varphi(x_j), \quad \forall i, j \in \{1, \cdots, n\}. \qquad \text{(C.7)}
$$

Note also that the expected value of the random variable X can be written as:

$$
\mathbb{E}_X\{X\} = \sum_{i=1}^{n} x_i \mathbb{P}_X\{X = x_i\}, \qquad \text{(C.8)}
$$

where

$$
\sum_{i=1}^{n} \mathbb{P}_X\{X = x_i\} = 1. \qquad \text{(C.9)}
$$

Thanks to these definitions we can prove the inequality:

$$\begin{aligned}\varphi(\mathbb{E}_X\{X\}) &= \varphi\left(\sum_{i=1}^{n} x_i \mathbb{P}_X\{X = x_i\}\right) \qquad (C.10)\\ &= \varphi\left(x_1\mathbb{P}_X\{X = x_1\} + (1 - \mathbb{P}_X\{X = x_1\})\sum_{i=2}^{n} x_i \frac{\mathbb{P}_X\{X = x_i\}}{1 - \mathbb{P}_X\{X = x_1\}}\right)\\ &\leqslant \mathbb{P}_X\{X = x_1\}\varphi(x_1) + (1 - \mathbb{P}_X\{X = x_1\})\varphi\left(\sum_{i=2}^{n} x_i \frac{\mathbb{P}_X\{X = x_i\}}{1 - \mathbb{P}_X\{X = x_1\}}\right).\end{aligned}$$

Since

$$\sum_{i=2}^{n} \frac{\mathbb{P}_X\{X = x_i\}}{1 - \mathbb{P}_X\{X = x_1\}} = 1, \qquad (C.11)$$

one can apply the induction hypotheses to the last term in the previous formula to obtain the result:

$$\begin{aligned}&\varphi(\mathbb{E}_X\{X\}) \qquad (C.12)\\ &= \varphi\left(\sum_{i=1}^{n} x_i \mathbb{P}_X\{X = x_i\}\right) \leqslant \sum_{i=1}^{n} \varphi(x_i)\mathbb{P}_X\{X = x_i\} = \mathbb{E}_X\{\varphi(X)\}.\end{aligned}$$

Note also that, if we multiply both sides of Jensen's inequality by -1, we can obtain:

$$-\varphi(\mathbb{E}_X\{X\}) \geqslant -\mathbb{E}_X\{\varphi(X)\} = \mathbb{E}_X\{-\varphi(X)\}. \qquad (C.13)$$

Since φ is convex $-\varphi$ is concave. Consequently, if $\varphi : \mathbb{R}^d \to \mathbb{R}$ is a concave function we obtain:

$$\varphi(\mathbb{E}_X\{X\}) \geqslant \mathbb{E}_X\{\varphi(X)\}. \qquad (C.14)$$

The Jensen's inequality is one of the most exploited properties in statistics because of its generality and simplicity.

Markov's Inequality

Markov's inequality gives an upper bound for the probability that a non-negative random variable X, which assumes values in $\mathcal{X} \subseteq [0, \infty)$, is greater than or equal to some positive constant [54]. In other words, for any $t \in (0, \infty)$:

$$\mathbb{P}_X\{X \geqslant t\} \leqslant \frac{\mathbb{E}_X\{X\}}{t}. \qquad (C.15)$$

In order to prove the property let us suppose that X is distributed according to a probability distribution P where its PDF is $\mathsf{p} : \mathcal{X} \to [0, \infty)$. Based on these definitions we can state that:

$$\begin{aligned}\mathbb{E}_X\{X\} = \int_0^\infty x\mathsf{p}(x)dx = \int_0^t x\mathsf{p}(x)dx + \int_t^\infty x\mathsf{p}(x)dx \geqslant \int_t^\infty x\mathsf{p}(x)dx \\ \geqslant \int_t^\infty t\mathsf{p}(x)dx \geqslant t\int_t^\infty \mathsf{p}(x)dx = t\mathbb{P}_X\{X \geqslant t\}. \qquad \text{(C.16)}\end{aligned}$$

Moreover, if $\varphi : \mathbb{R} \to [0, \infty)$ is a strictly monotonically increasing nonnegative-valued function, for any $t \in (0, \infty)$ we obtain:

$$\mathbb{P}_X\{\varphi(X) \geqslant \varphi(t)\} \leqslant \frac{\mathbb{E}_X\{\varphi(X)\}}{\varphi(t)}. \qquad \text{(C.17)}$$

The proof is analogous to the one of Eq. (C.16). Markov's inequality is extremely useful and it is the basis of each result presented in this monograph. In fact, it is used to prove all the CIQs for sum of i.i.d. variables [118] and for functions of i.i.d. random variables [47].

Chebyshev's Inequality

The Chebyshev's inequality measures how large is the probability of a random variable X, which assumes values in $\mathcal{X} \subseteq [0, \infty)$, to be far from its expected value in terms of its variance [258]. Note that Chebyshev's inequality can be seen and as a simple application of Markov's inequality, for any $t \in (0, \infty)$:

$$\begin{aligned}\mathbb{P}_X\{X - \mathbb{E}_X\{X\} \geqslant t\} \leqslant \mathbb{P}_X\{|X - \mathbb{E}_X\{X\}| \geqslant t\} = \mathbb{P}_X\{(X - \mathbb{E}_X\{X\})^2 \geqslant t^2\} \\ \leqslant \frac{\mathbb{E}_X\{(X - \mathbb{E}_X\{X\})^2\}}{t^2} = \frac{\mathbb{V}_X\{X\}}{t^2}. \qquad \text{(C.18)}\end{aligned}$$

Analogously we can state that for any $t \in (0, \infty)$:

$$\mathbb{P}_X\{\mathbb{E}_X\{X\} - X \geqslant t\} \leqslant \frac{\mathbb{V}_X\{X\}}{t^2}. \qquad \text{(C.19)}$$

Analogously to Markov's inequality, Chebyshev's inequality possesses great utility because it can be applied to completely arbitrary distributions.

Hoeffding's Lemma

Hoeffding's lemma is an inequality which bounds the moment-generating function of any bounded random variable [118]. Let X be a random variable which assumes values in $\mathcal{X} \subseteq [a, b]$, with $a, b \in \mathbb{R}$, $a \leqslant b$ and $\mathbb{E}_X\{X\} = 0$. Because of the convexity of the exponential function we can state that $\forall h \in \mathbb{R}$:

$$\mathbb{E}_X\left\{e^{hX}\right\} \leqslant \mathbb{E}_X\left\{\frac{b-X}{b-a}e^{ha} + \frac{X-a}{b-a}e^{hb}\right\} \tag{C.20}$$

$$= \frac{b-\mathbb{E}_X\{X\}}{b-a}e^{ha} + \frac{\mathbb{E}_X\{X\}-a}{b-a}e^{hb} \tag{C.21}$$

$$= \frac{b}{b-a}e^{ha} - \frac{a}{b-a}e^{hb}. \tag{C.22}$$

By setting $\lambda = h(b-a)$, $p = -a/(b-a)$ and $f(\lambda) = \ln(1-p+pe^h) - hp$ we obtain:

$$\mathbb{E}_X\left\{e^{hX}\right\} \leqslant e^{f(\lambda)}. \tag{C.23}$$

Taking the derivative with respect to λ we obtain:

$$f(0) = 0, \quad \frac{df(\lambda)}{d\lambda}\Big|_{\lambda=0} = 0, \quad \frac{d^2 f(\lambda)}{d\lambda^2}\Big|_{\lambda=0} \leqslant \frac{1}{4}, \tag{C.24}$$

Consequently by Taylor expansion, we obtain:

$$\mathbb{E}_X\left\{e^{hX}\right\} \leqslant e^{f(\lambda)} \leqslant e^{\frac{1}{8}\lambda^2} = e^{\frac{1}{8}h^2(b-a)^2}. \tag{C.25}$$

This is a property that is widely used for deriving CIQs as the HIQs [118] and the BDFIQs [181].

Kullback-Leibler Divergence and Pinsker's Inequality

The Kullback-Leibler divergence [142] is a non-symmetric measure of the difference between two probability distributions Q and P defined over a set $\mathcal{X}$. If Q and P are continuous distributions their PDFs are respectively $\mathsf{p} : \mathcal{X} \to [0, \infty)$ and $\mathsf{q} : \mathcal{X} \to [0, \infty)$ and with x we indicate a point in the set $\mathcal{X}$. The divergence is defined as follows:

$$\mathtt{KL}(\mathsf{Q}||\mathsf{P}) = \int_{\mathcal{X}} \mathsf{q}(x) \ln\left[\frac{\mathsf{q}(x)}{\mathsf{p}(x)}\right] dx. \tag{C.26}$$

If Q and P are discrete distributions and Q and P are random variables distributed according to respectively Q and P, we have that:

$$\mathtt{KL}(\mathsf{Q}||\mathsf{P}) = \sum_{x\in\mathcal{X}} \mathbb{P}\{Q = x\} \ln\left[\frac{\mathbb{P}\{Q = x\}}{\mathbb{P}\{P = x\}}\right] dx. \tag{C.27}$$

Whenever $\mathsf{q}(x)$ (or $\mathbb{P}\{Q = x\}$) is zero the contribution of that term is interpreted as zero because:

$$\lim_{a\to 0} a \ln\left[\frac{a}{b}\right] = 0, \quad a, b \in [0, \infty). \tag{C.28}$$

The Kullback-Leibler divergence is defined only if $\mathsf{p}(x) = 0$ (or $\mathbb{P}\{P = x\} = 0$) implies $\mathsf{q}(x) = 0$ (or $\mathbb{P}\{Q = x\} = 0$). Nevertheless, in [31] it is noted that the divergence becomes infinite whenever $\mathsf{p}(x) = 0$ (or $\mathbb{P}\{P = x\} = 0$) and $\mathsf{q}(x) \neq 0$ (or $\mathbb{P}\{Q = x\} \neq 0$) (no matter how small):

$$\lim_{b\to 0} a \ln\left[\frac{a}{b}\right] = \infty, \quad a, b \in [0, \infty). \tag{C.29}$$

A particular case is when Q and P are Bernoulli distributions [127] with probability of success respectively equal to $q \in [0, 1]$ and $p \in [0, 1]$. In this case we obtain:

$$\mathtt{KL}(\mathsf{Q}||\mathsf{P}) = q \ln\left[\frac{q}{p}\right] + (1-q)\ln\left[\frac{(1-q)}{(1-p)}\right] \triangleq \mathtt{kl}(q||p). \tag{C.30}$$

Pinsker's inequality, instead, is an inequality that bounds the total variation distance (or statistical distance) in terms of the Kullback-Leibler divergence [211, 262]. The inequality is tight up to constant factors [73].

$$\sup_{\mathcal{A}\subseteq\mathcal{X}} \left|\int_{\mathcal{A}} [\mathsf{q}(x) - \mathsf{p}(x)]\, dx\right| \leqslant \sqrt{\frac{1}{2}\mathtt{KL}(\mathsf{Q}||\mathsf{P})}, \quad \begin{matrix}\text{continuous}\\ \text{distributions}\end{matrix}, \tag{C.31}$$

$$\sup_{\mathcal{A}\subseteq\mathcal{X}} \left|\sum_{x\in\mathcal{A}} [\mathbb{P}\{Q{=}x\} - \mathbb{P}\{P{=}x\}]\right| \leqslant \sqrt{\frac{1}{2}\mathtt{KL}(\mathsf{Q}||\mathsf{P})}, \quad \begin{matrix}\text{discrete}\\ \text{distributions}\end{matrix}. \tag{C.32}$$

The proof is rather technical and it will not be reported here but it can be retrieved from [262]. Again, when Q and P are Bernoulli distributions, we obtain:

$$|q - p| \leqslant \sqrt{\frac{1}{2}\mathtt{kl}(q||p)}. \tag{C.33}$$

The Kullback-Leibler divergence and Pinsker's inequality are widely used both in deriving CIQs (e.g. the HIQs [118]) and in the SLT [267] (e.g. the PAC Bayes Theory [180]).

Inequalities for Sum of Random Variables

The law of large numbers of classical probability theory states that sums of independent random variables are, under very mild conditions, close to their expectation with large probability [54]. A series of inequalities can be derived for specifying how close they are. Let $\{X, X_1, \cdots, X_n\}$ be $n+1$ bounded i.i.d. random variables which take values in $\mathcal{X} \subseteq [0,1]$ such that:

$$\mathbb{E}_X\{X\} = \mu, \tag{C.34}$$

$$\mathbb{V}_X\{X\} = \sigma^2. \tag{C.35}$$

Moreover let us define their empirical unbiased counterparts:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i, \tag{C.36}$$

$$\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2 = \frac{1}{2n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)^2. \tag{C.37}$$

Note that [54]:

$$\mathbb{E}_{X_1,\cdots,X_n}\{\widehat{\mu}\} = \mu, \tag{C.38}$$

$$\mathbb{E}_{X_1,\cdots,X_n}\{\widehat{\sigma}^2\} = \sigma^2. \tag{C.39}$$

Based on this preliminary definitions we can present the different inequalities.

Clopper-Pearson's Inequalities

The CPIQ state the exact confidence intervals for Binomial tails [67]. In particular let $\mathcal{X} = \{0,1\}$, so the random variables under analysis are Bernoulli random variables [127] with probability of success μ. Then, for each $t \in \{0, 1/n, 2/n, \cdots, 1\}$, we obtain:

$$\mathbb{P}_{X_1,\cdots,X_n}\{\widehat{\mu} = t\} = \binom{n}{nt}\mu^{nt}(1-\mu)^{n-nt}, \tag{C.40}$$

$$\mathbb{P}_{X_1,\cdots,X_n}\{\widehat{\mu} \leqslant t\} = \sum_{i=0}^{nt}\binom{n}{i}\mu^{i}(1-\mu)^{n-i}, \tag{C.41}$$

$$\mathbb{P}_{X_1,\cdots,X_n}\{\widehat{\mu} \geqslant t\} = \sum_{i=nt}^{n}\binom{n}{i}\mu^{i}(1-\mu)^{n-i}. \tag{C.42}$$

Based on these definitions we can bound, with probability $(1-\delta)$, the probability of success μ given the fact that we observed $n\widehat{\mu}$ successes:

$$\mu \leqslant \max_{p\in[0,1]}\left[p : \sum_{i=0}^{n\widehat{\mu}}\binom{n}{i}p^{i}(1-p)^{n-i} \geqslant \delta\right] \triangleq \mathtt{cp}_{\leqslant}(\widehat{\mu}, n, \delta), \tag{C.43}$$

$$\mu \geqslant \min_{p\in[0,1]}\left[p : \sum_{i=n\widehat{\mu}}^{n}\binom{n}{i}p^{i}(1-p)^{n-i} \geqslant \delta\right] \triangleq \mathtt{cp}_{\geqslant}(\widehat{\mu}, n, \delta). \tag{C.44}$$

We can also bound, with probability $(1-\delta)$, $\widehat{\mu}$ given that the probability of success is μ:

$$\widehat{\mu} \leqslant \max_{c \in \{1,\cdots,n\}} \left[\frac{c}{n} : \sum_{i=c}^{n} \binom{n}{i} \mu^i (1-\mu)^{n-i} \geqslant \delta \right] \triangleq \widehat{\mathtt{cp}}_{\leqslant}(\mu, n, \delta), \tag{C.45}$$

$$\widehat{\mu} \geqslant \min_{c \in \{1,\cdots,n\}} \left[\frac{c}{n} : \sum_{i=0}^{c} \binom{n}{i} \mu^i (1-\mu)^{n-i} \geqslant \delta \right] \triangleq \widehat{\mathtt{cp}}_{\geqslant}(\mu, n, \delta). \tag{C.46}$$

Unfortunately these bounds are all in closed form but they can be easily computed with simple numerical routines [7]. For example, because of a relationship between the cumulative Binomial distribution and the Beta distribution [127], $\mathtt{cp}_{\leqslant}$ and $\mathtt{cp}_{\geqslant}$ are sometimes presented in an alternative format that uses quantiles from the beta distribution:

$$\mathtt{cp}_{\leqslant}(\widehat{\mu}, n, \delta) = \mathsf{Q}(1-\delta; n\widehat{\mu}+1, n-n\widehat{\mu}), \tag{C.47}$$

$$\mathtt{cp}_{\geqslant}(\widehat{\mu}, n, \delta) = \mathsf{Q}(\delta; n\widehat{\mu}, n-n\widehat{\mu}+1), \tag{C.48}$$

where $\mathsf{Q}(p; v, w)$ is the p-th quantile from a Beta distribution with shape parameters v and w.

Even if the bounds are in implicit form, there are some properties which may be useful in order to better handle and understand them. In particular, in [147], it has been proved that:

$$\mathtt{cp}_{\leqslant}(\widehat{\mu}, n, e^{-x}) \leqslant \widehat{\mu} + \sqrt{\frac{x}{2n}}, \tag{C.49}$$

$$\mathtt{cp}_{\leqslant}(\widehat{\mu}, n, e^{-x}) \leqslant \widehat{\mu} + \sqrt{\widehat{\mu}\frac{2x}{n}} + \frac{2x}{n}, \tag{C.50}$$

$$\mathtt{cp}_{\leqslant}(0, n, e^{-x}) \leqslant \frac{x}{n}, \tag{C.51}$$

$$\mathtt{cp}_{\leqslant}(\widehat{\mu}, n, \delta) - \widehat{\mu} \leqslant \mathtt{cp}_{\leqslant}(0.5, n, \delta) - 0.5. \tag{C.52}$$

A further result can be retrieved by exploiting the CPIQ. In particular, it is possible to establish a fundamental connection between binomial parameters and means of bounded random variables [59, 60]. Let $\mathcal{X} = [0,1]$ distributed according to P with its PDF $\mathsf{p} : \mathcal{X} \to \mathbb{R}$. Let U be a random variable uniformly distributed over $\mathcal{U} = [0,1]$, and consequently its PDF is $\mathsf{u}(u) = 1$ if $u \in [0,1]$ and $\mathsf{u}(u) = 0$ elsewhere. Let us suppose that X and U are independent. Then:

$$\mathbb{P}_{X,U}\{X \geqslant U\} = \int_0^1 \left(\int_0^x \mathsf{u}(u) du \right) \mathsf{p}(x) dx = \int_0^1 x\mathsf{p}(x) dx = \mathbb{E}_X\{X\}. \tag{C.53}$$

We can exploit this property in order to prove a general result for sum of bounded random variables. Let $\{U, U_1, \cdots, U_n\}$ be $n+1$ i.i.d. random variables. Let also $\{Z, Z_1, \cdots, Z_n\}$ be other $n+1$ i.i.d. random variables such that $Z = [X \geqslant U]$. Note that Z is a Bernoulli random variable. By exploiting the previous results we can state that the following bound holds with probability $(1-\delta)$:

$$\begin{aligned}\mu &= \mathbb{E}_X\{X\} = \mathbb{P}_{X,U}\{X \geqslant U\} \\ &= \mathbb{E}_{X,U}\{[X \geqslant U]\} \leqslant \mathtt{cp}_{\leqslant}\left(\frac{1}{n}\sum_{i=1}^{n}[X_i \geqslant U_i], n, \delta\right). \end{aligned} \tag{C.54}$$

Analogously, we can state that:

$$\mu \geqslant \mathtt{cp}_{\geqslant}\left(\frac{1}{n}\sum_{i=1}^{n}[X_i \geqslant U_i], n, \delta\right). \tag{C.55}$$

In order to avoid an unlucky realization, and since we know the distribution of U, one can generate much more realizations of U: $\{U_1, \cdots, U_{mn}\}$. By taking the average of the different realizations we can state that the following bounds hold with probability $(1-e^{-x})$:

$$\mu \leqslant \frac{1}{m}\sum_{j=1}^{n}\mathtt{cp}_{\leqslant}\left(\frac{1}{n}\sum_{i=1}^{n}[X_i \geqslant U_{(j-1)n+i}], n, \delta\right), \tag{C.56}$$

$$\mu \geqslant \frac{1}{m}\sum_{j=1}^{n}\mathtt{cp}_{\geqslant}\left(\frac{1}{n}\sum_{i=1}^{n}[X_i \geqslant U_{(j-1)n+i}], n, \delta\right). \tag{C.57}$$

Note that this result is quite important since it represents the best known bound for sum of $[0,1]$-bounded random variables which does not exploits moments of higher order (e.g. the variance).

Hoeffding's Inequalities

The HIQ are very general results for sum of bounded random variables [118]. Let $\mathcal{X} = [0,1]$. The proof of the HIQ is quite simple:

$$
\begin{aligned}
&\mathbb{P}_{X_1,\cdots,X_n}\left\{\widehat{\mu}-\mu \geqslant t\right\} && \text{(C.58)}\\
&= \mathbb{P}_{X_1,\cdots,X_n}\left\{\sum_{i=1}^{n} X_i - n\mu \geqslant nt\right\}, \quad t \in (0, 1-\mu) && \text{(C.59)}\\
&\leqslant \mathbb{E}_{X_1,\cdots,X_n} e^{h\left(\sum_{i=1}^{n} X_i - n\mu - nt\right)}, \quad h \in \mathbb{R} && \text{(C.60)}\\
&= e^{-hnt}\mathbb{E}_{X_1,\cdots,X_n}\left\{e^{h\sum_{i=1}^{n}(X_i-\mu)}\right\} && \text{(C.61)}\\
&= e^{-hnt}\prod_{i=1}^{n}\mathbb{E}_{X_i}\left\{e^{h(X_i-\mu)}\right\} && \text{(C.62)}\\
&= e^{-hnt}e^{-nh\mu}\prod_{i=1}^{n}\mathbb{E}_{X_i}\left\{e^{hX_i}\right\} && \text{(C.63)}\\
&\leqslant e^{-hnt}e^{-nh\mu}\prod_{i=1}^{n}(1-\mu+\mu e^{h}) && \text{(C.64)}\\
&= e^{-hnt}e^{-nh\mu}(1-\mu+\mu e^{h})^{n} && \text{(C.65)}\\
&= \left[e^{-ht}e^{-h\mu}(1-\mu+\mu e^{h})\right]^{n} && \text{(C.66)}\\
&= \left[\left(\frac{1-\mu}{-\mu-t+1}\right)^{-\mu-t+1}\left(\frac{\mu}{\mu+t}\right)^{\mu+t}\right]^{n} && \text{(C.67)}\\
&= e^{-n\mathtt{kl}(\mu+t,\mu)} && \text{(C.68)}\\
&\leqslant e^{-2nt^2}, \quad t \in [0,\infty). && \text{(C.69)}
\end{aligned}
$$

Note that:

- From Eq. (C.59) to Eq. (C.60) we have exploited Markov's inequality;
- From Eq. (C.61) to Eq. (C.62) we have exploited the i.i.d. hypothesis;
- From Eq. (C.63) to Eq. (C.64) we have exploited the convexity of the exponential function;
- From Eq. (C.66) to Eq. (C.67) we have optimized respect to h:

$$
\begin{aligned}
&\frac{\partial\left[e^{-ht}e^{-h\mu}(1-\mu+\mu e^{h})\right]}{\partial h} = 0, && \text{(C.70)}\\
&\mu e^{h(-\mu)-ht+h} + (e^{h}\mu-\mu+1)(-\mu-t)e^{-h\mu-ht} = 0, && \text{(C.71)}\\
&h \to \ln\left[\frac{(\mu-1)(\mu+t)}{\mu(\mu+t-1)}\right]; && \text{(C.72)}
\end{aligned}
$$

- From Eq. (C.68) to Eq. (C.69) we have exploited Pinsker's inequality.

By following the same argument it is possible to prove that:

$$
\mathbb{P}_{X_1,\cdots,X_n}\left\{\mu-\widehat{\mu} \geqslant t\right\} \leqslant e^{-2nt^2}, \quad t \in [0,\infty). \tag{C.73}
$$

The two bounds can be presented in their explicit form and, with probability $(1 - e^{-x})$, we obtain:

$$\mu - \widehat{\mu} \leqslant \sqrt{\frac{x}{2n}}, \tag{C.74}$$

$$\widehat{\mu} - \mu \leqslant \sqrt{\frac{x}{2n}}. \tag{C.75}$$

Note that HIQ imply also some tighter results (see Eq. (C.68)) which are often less known because of their implicit form:

$$\mathbb{P}_{X_1,\cdots,X_n}\left\{\widehat{\mu} - \mu \geqslant t\right\} \leqslant e^{-nkl(\mu+t,\mu)}, \quad t \in (0, 1-\mu), \tag{C.76}$$

$$\mathbb{P}_{X_1,\cdots,X_n}\left\{\mu - \widehat{\mu} \geqslant t\right\} \leqslant e^{-nkl(1-\mu+t,1-\mu)}, \quad t \in (0, \mu-1). \tag{C.77}$$

The latter bounds are much tighter with respect to the conventional ones of Eqns. (C.74) and (C.75) but a numerical routine is needed in order to compute the upper or lower bounds of μ (by solving the problem with respect to μ [7]) or $\widehat{\mu}$. Note that recently in [36, 255] it has been shown that the bounds of Eqns. (C.76) and (C.77) can be further improved by a constant factor. We do not include this result since the results reported in Appendix C are tighter with respect to the ones of [36, 255].

Taking Into Account Moments of Higher Order

The inequalities presented in the previous Appendixes (CPIQ and HIQ) are the best known bounds that exploit information about the first moment of the distribution (the mean value). Other results, which exploit moment of higher order like the variance, have been proved in the past. In particular, by following a slightly different proof with respect to the one presented for the HIQ, it is possible to prove that:

$$\mathbb{P}_{X_1,\cdots,X_n}\left\{\widehat{\mu} - \mu \geqslant t\right\} \leqslant e^{-nkl\left(\frac{t+\sigma^2}{1+\sigma^2},\frac{\sigma^2}{1+\sigma^2}\right)}, \quad t \in (0, 1-\mu), \tag{C.78}$$

$$\mathbb{P}_{X_1,\cdots,X_n}\left\{\mu - \widehat{\mu} \geqslant t\right\} \leqslant e^{-nkl\left(\frac{t+\sigma^2}{1+\sigma^2},\frac{\sigma^2}{1+\sigma^2}\right)}, \quad t \in (0, \mu-1). \tag{C.79}$$

The proof can be retrieved from [36, 118]. By upper bounding the previous results one can derive the Bernstein's [38], Bennett's [35] and Prohorov's [215] inequalities (refer to [118] for details). The bounds of Eqns. (C.78) and (C.79) can be much tighter respect to their counterparts (Eqns. (C.76) and (C.77)). In particular, by using rescaling and choosing the maximal possible variance $\sigma^2 = \mu(1-\mu)$, the bounds of Eqns. (C.78) and (C.79) degenerate into the ones of Eqns. (C.76) and (C.77). Moreover, the bounds of Eqns. (C.78) and (C.79) can be further improved by a constant factor (see [36, 118]), but this is out of our scope. The problem here it that the bounds of Eqns. (C.78) and (C.79) cannot be empirically computed since σ^2 is unknown.

In order to estimate σ^2 we have to exploit a result from the literature [173, 175]. In particular, it is possible to prove that $\widehat{\sigma}^2$ is a SBF (see Appendix C); therefore it is possible to state that:

$$\mathbb{P}_{X_1,\cdots,X_n}\left\{\sigma^2 - \widehat{\sigma}^2 \geqslant t\right\} \leqslant e^{-\frac{(n-1)t^2}{2\sigma^2}}, \quad t \in [0,\infty). \tag{C.80}$$

Note that the above mentioned bound is not the sharpest one (see Appendix C and [173, 175]) but for the purpose of our presentation the reported bound is good enough. Moreover, we do not report the proof here because it is rather technical and it is out of the scope of this Appendix, in any case, it can be retrieved from [173, 175]. The bound of Eq. (C.80) allows to upper bound σ^2 based on its empirical estimator $\widehat{\sigma}^2$. In particular, by exploiting Eq. (C.80) we can state that with probability $(1-e^{-x})$:

$$\sigma^2 - \widehat{\sigma}^2 \leqslant \sqrt{\frac{2\sigma^2 x}{n-1}}. \tag{C.81}$$

By solving it with respect to σ^2, and by taking the largest solution, with probability $(1-e^{-x})$:

$$\sigma^2 \leqslant \widehat{\sigma}^2 + \sqrt{\frac{2\widehat{\sigma}^2 x}{n-1}} + \frac{2x}{n-1}. \tag{C.82}$$

By plugging this last result into the bounds of Eqns. (C.78) and (C.79) we get their empirical counterparts. When n is small the advantage of using the bounds of Eqns. (C.78) and (C.79) with respect to the ones of Eqns. (C.76) and (C.77) is wiped out by the estimation of σ^2 through the bound of Eq. (C.82). Instead, as n increases, Eq. (C.82) becomes tighter and the bounds of Eqns. (C.78) and (C.79) still improve over the ones of Eqns. (C.76) and (C.77). Unfortunately this effect becomes evident for really large $n \approx 10^3 \div 10^4$, therefore the bounds of Eqns. (C.78) and (C.79) are often not taken into account in practical applications [7].

Inequalities for functions of random variables

As described in Appendix C, sums of i.i.d. random variables can be close to their expectation with a large probability. This appendix, instead, gives some results about how close general functions of i.i.d. random variables are from their expected value. For this purpose we assume that $\{X_1, \cdots, X_n\}$ are n i.i.d. random variables taking values in $\mathcal{X}$ and we report some preliminary definitions. In particular we will denote with Z a function on n random variables

$$Z = f(X_1, \cdots, X_n), \quad f : \mathcal{X}^n \to \mathbb{R}. \tag{C.83}$$

We will denote Z^k the same function Z but where the k-th element X_k is replaced with another one X'_k i.i.d. to X_k

$$\begin{aligned} Z^k =& f(X_1, \cdots, X_{k-1}, X'_k, X_{k+1}, \cdots, X_n), \\ & X_k, X'_k \ \text{ i.i.d.}, \quad k \in \{1, \cdots, n\}. \end{aligned} \tag{C.84}$$

Finally we will denote with $G^{\backslash k}$ another function of $n-1$ variables defined as $\{X_1, \cdots, X_n\} \backslash X_k$

$$\begin{aligned} G^{\backslash k} =& g(X_1, \cdots, X_{k-1}, X_{k+1}, \cdots, X_n), \\ & g : \mathcal{X}^{n-1} \to \mathbb{R}, \quad k \in \{1, \cdots, n\} \end{aligned} \tag{C.85}$$

We will make also use of some additional auxiliary mathematical functions [196]

$$\phi(a) = (1+a)\ln(1+a) - a, \quad a > -1, \tag{C.86}$$

$$\widehat{\phi}(a) = 1 - \exp\left[1 + W_{-1}\left(\frac{a-1}{e}\right)\right], \quad \phi\left[-\widehat{\phi}(a)\right] = a, \quad a \in [0,1], \tag{C.87}$$

$$\check{\phi}(a) = \exp\left[1 + W_0\left(\frac{a-1}{e}\right)\right] - 1, \quad \phi\left[\check{\phi}(a)\right] = a, \quad a \in [0,\infty), \tag{C.88}$$

where W_{-1} and W_0 are, respectively, two solutions of the Lambert W function [69]. Note also that [46]:

$$\phi(a) \geqslant \frac{a^2}{2 + \frac{2a}{3}}, \quad a \in [0, \infty], \tag{C.89}$$

$$\phi(-a) \geqslant \frac{a^2}{2}, \quad a \in [0,1]. \tag{C.90}$$

In the next sections of this appendix we will present some inequalities which bound, with high probability, the difference between the random variable Z and its expected value. The first inequalities, discovered by [181], are the ones for Bounded Difference Functions reported in Section C. Then the ones for Self Bounding Function have been proposed in [46] and then further refined

for Generalized Self Bounding Functions in [48]. The latter are reported in SectionsC and C respectively.
Generally speaking, the loosest inequalities are the ones of Appendix C, while the sharpest one are those of Appendix C. The tightness of the inequalities of Appendix C, instead, remarkably depends on the variance of Z: this is the typical behaviour of the Bennet-type inequalities. Obviously, in order to apply the inequalities of Appendix C, f must satisfy less restrictive conditions respect when we want to apply the inequalities of Appendix C. The inequalities of Appendix C require instead less restrictive conditions with respect to the ones required by the inequalities of Appendix C but more restrictive ones respect to the ones required by the inequalities of Appendix C. All the presented results are usually called CIQs. Several methods have been proposed to prove such inequalities, including martingale methods [181, 182], information-theoretic methods [2, 74, 167–170, 222], induction method [254, 256, 257], entropy method based on logarithmic Sobolev inequalities [44, 46, 48, 154, 155, 171, 222], and various problem-specific methods [125]. All these results have been recently surveyed in [47].

Bounded Difference Function Inequalities

The function Z is a BDF if it satisfies the following property [181]:

$$\left|Z - Z^k\right| \leqslant c, \quad c \in (0, \infty), \quad \forall k \in \{1, \cdots, n\}\,. \tag{C.91}$$

If Z is a BDF we can state that [181]:

$$\mathbb{P}_{X_1,\cdots,X_n}\left\{Z - \mathbb{E}_{X_1,\cdots,X_n}\{Z\} > t\right\} \leqslant e^{\frac{-2t^2}{nc^2}}, \quad t \in [0, \infty), \tag{C.92}$$

$$\mathbb{P}_{X_1,\cdots,X_n}\left\{\mathbb{E}_{X_1,\cdots,X_n}\{Z\} - Z > t\right\} \leqslant e^{\frac{-2t^2}{nc^2}}, \quad t \in [0, \infty). \tag{C.93}$$

The two above mentioned properties can be rewritten and we can state that with probability $(1 - e^{-x})$:

$$Z - \mathbb{E}_{X_1,\cdots,X_n}\{Z\} \leqslant \sqrt{\frac{nc^2x}{2}}, \tag{C.94}$$

$$\mathbb{E}_{X_1,\cdots,X_n}\{Z\} - Z \leqslant \sqrt{\frac{nc^2x}{2}}. \tag{C.95}$$

$$\tag{C.96}$$

In order to proceed with the proof, let us define the following random variable:

$$Z_i = \mathbb{E}_{X_1,\cdots,X_i}\{f(X_1,\cdots,X_n)\}: \quad Z_0 = Z, \quad Z_n = \mathbb{E}_{X_1,\cdots,X_n}\{Z\}, \tag{C.97}$$

and note that

$$Z - \mathbb{E}_{X_1,\cdots,X_n}\{Z\} = \sum_{i=1}^{n} (Z_{i-1} - Z_i). \tag{C.98}$$

By construction we obtain:

$$\mathbb{E}_{X_i}\{Z_{i-1} - Z_i\} = 0, \tag{C.99}$$

and thanks to the fact that Z is a self bounding function we obtain:

$$Z_{i-1} - Z_i \leqslant c. \tag{C.100}$$

Consequently by exploiting Hoeffding's Lemma we can state that:

$$\mathbb{E}_{X_1,\cdots,X_i}\left\{e^{h(Z_{i-1}-Z_i)}\right\} \leqslant e^{\frac{1}{8}h^2c^2}. \tag{C.101}$$

Based on these preliminary definitions we can state that:

$$\begin{aligned}
&\mathbb{P}_{X_1,\cdots,X_n}\{Z - \mathbb{E}_{X_1,\cdots,X_n}\{Z\} > t\} && \text{(C.102)}\\
&\leqslant e^{-ht}\mathbb{E}_{X_1,\cdots,X_n}\left\{e^{h(Z-\mathbb{E}_{X_1,\cdots,X_n}\{Z\})}\right\}, \quad t \in [0,\infty), \quad h \in \mathbb{R} && \text{(C.103)}\\
&= e^{-ht}\mathbb{E}_{X_1,\cdots,X_n}\left\{e^{h\sum_{i=1}^{n}(Z_{i-1}-Z_i)}\right\} && \text{(C.104)}\\
&= e^{-ht}\mathbb{E}_{X_1,\cdots,X_n}\left\{e^{h\sum_{i=2}^{n}(Z_{i-1}-Z_i)}\mathbb{E}_{X_1}\left\{e^{h(Z_0-Z_1)}\right\}\right\} && \text{(C.105)}\\
&\leqslant e^{-ht}e^{\frac{1}{8}h^2c^2}\mathbb{E}_{X_1,\cdots,X_n}\left\{e^{h\sum_{i=2}^{n}(Z_{i-1}-Z_i)}\right\} && \text{(C.106)}\\
&\leqslant e^{-ht}e^{\frac{1}{8}nh^2c^2} && \text{(C.107)}\\
&\leqslant e^{\frac{-2t^2}{nc^2}} && \text{(C.108)}
\end{aligned}$$

where:

- In Eq. (C.103) we have exploited Markov's inequality;
- From Eq. (C.103) to Eq. (C.104) we have exploited the property of Eq. (C.98);
- From Eq. (C.105) to Eq. (C.106) we have exploited the property of Eq. (C.101);
- From Eq. (C.106) to Eq. (C.107) we applied the induction hypotheses;
- From Eq. (C.107) to Eq. (C.108) we have optimized respect to h.

The proof for the other BDFIQ are analogous.

Self Bounding Function Inequalities

The function Z is a SBF if it satisfies the following properties [46]:

$$Z \geqslant 0, \tag{C.109}$$

$$0 \leqslant Z - G^{\backslash k} \leqslant c, \quad c \in (0, \infty), \quad \forall k \in \{1, \cdots, n\}, \tag{C.110}$$

$$\sum_{k=1}^{n} \left[Z - G^{\backslash k} \right] \leqslant Z. \tag{C.111}$$

If Z is a SBF we can state that [46]:

$$\begin{aligned} &\mathbb{P}_{X_1, \cdots, X_n} \left\{ Z - \mathbb{E}_{X_1, \cdots, X_n}\{Z\} \geqslant t \right\} \\ &\qquad \leqslant e^{-\frac{\mathbb{E}_{X_1, \cdots, X_n}\{Z\}}{c} \phi\left(\frac{t}{\mathbb{E}_{X_1, \cdots, X_n}\{Z\}} \right)} \\ &\qquad \leqslant e^{-\frac{t^2}{2c\mathbb{E}_{X_1, \cdots, X_n}\{Z\} + \frac{2}{3}ct}}, \quad t \in [0, \infty). \end{aligned} \tag{C.112}$$

The above mentioned result can be rewritten and we can state that with probability $(1 - e^{-x})$:

$$Z - \mathbb{E}_{X_1, \cdots, X_n}\{Z\} \leqslant \mathbb{E}_{X_1, \cdots, X_n}\{Z\} \check{\phi} \left(\frac{cx}{\mathbb{E}_{X_1, \cdots, X_n}\{Z\}} \right), \tag{C.113}$$

$$Z - \mathbb{E}_{X_1, \cdots, X_n}\{Z\} \leqslant \sqrt{2cx\mathbb{E}_{X_1, \cdots, X_n}\{Z\}} + \frac{cx}{3}. \tag{C.114}$$

Moreover [46]:

$$\begin{aligned} &\mathbb{P}_{X_1, \cdots, X_n} \left\{ \mathbb{E}_{X_1, \cdots, X_n}\{Z\} - Z \geqslant t \right\} \\ &\qquad \leqslant e^{-\frac{\mathbb{E}_{X_1, \cdots, X_n}\{Z\}}{c} \phi\left(-\frac{t}{\mathbb{E}_{X_1, \cdots, X_n}\{Z\}} \right)} \\ &\qquad \leqslant e^{-\frac{t^2}{2c\mathbb{E}_{X_1, \cdots, X_n}\{Z\}}}, \quad t \in [0, \mathbb{E}_{X_1, \cdots, X_n}\{Z\}], \end{aligned} \tag{C.115}$$

which can be rewritten and we can state that with probability $(1 - e^{-x})$:

$$\mathbb{E}_{X_1, \cdots, X_n}\{Z\} - Z \leqslant \mathbb{E}_{X_1, \cdots, X_n}\{Z\} \widehat{\phi} \left(\frac{cx}{\mathbb{E}_{X_1, \cdots, X_n}\{Z\}} \right), \tag{C.116}$$

$$\mathbb{E}_{X_1, \cdots, X_n}\{Z\} - Z \leqslant \sqrt{2cx\mathbb{E}_{X_1, \cdots, X_n}\{Z\}}. \tag{C.117}$$

In this case we do not report the proofs which is rather technical, but they can be retrieved from [46].

Generalized Self Bounding Function Inequalities

The function Z is a GSBF if it satisfies the following properties [48]:

$$Z'_k \leqslant Z - G^{\setminus k} \leqslant c, \quad \forall k \in \{1, \cdots, n\}, \tag{C.118}$$

$$Z'_k \leqslant c', \quad \forall k \in \{1, \cdots, n\}, \tag{C.119}$$

$$\mathbb{E}\left[Z'_k\right] \geqslant 0, \quad \forall k \in \{1, \cdots, n\}, \tag{C.120}$$

$$\sum_{k=1}^{n} \left[Z - G^{\setminus k}\right] \leqslant Z. \tag{C.121}$$

Let us define the following quantities:

$$u^2 \geqslant \frac{1}{c^2 n} \sum_{k=1}^{n} \mathbb{E}_{X_1, \cdots, X_n} \left[Z'_k\right]^2, \tag{C.122}$$

$$v = \frac{1 + nc'}{c} \mathbb{E}_{X_1, \cdots, X_n}\{Z\} + nu^2. \tag{C.123}$$

If Z is a GSBF we can state that [48]:

$$\begin{aligned} &\mathbb{P}_{X_1, \cdots, X_n} \left\{Z - \mathbb{E}_{X_1, \cdots, X_n}\{Z\} \geqslant t\right\} \\ &\qquad \leqslant e^{-v\phi\left(\frac{t}{cv}\right)} \\ &\qquad \leqslant e^{-\frac{t^2}{2c^2 v + \frac{2}{3}ct}}, \quad t \in [0, \infty). \end{aligned} \tag{C.124}$$

The above mentioned result can be easily reformulated. In particular with probability $(1 - e^{-x})$ we have that:

$$Z - \mathbb{E}_{X_1, \cdots, X_n}\{Z\} \leqslant cv\check{\phi}\left(\frac{x}{v}\right), \tag{C.125}$$

$$Z - \mathbb{E}_{X_1, \cdots, X_n}\{Z\} \leqslant \sqrt{2c^2 xv} + \frac{cx}{3}. \tag{C.126}$$

Moreover [133, 134]:

$$\begin{aligned} &\mathbb{P}_{X_1, \cdots, X_n} \left\{\mathbb{E}_{X_1, \cdots, X_n}\{Z\} - Z \geqslant t\right\} \\ &\qquad \leqslant e^{-v\phi\left(-\frac{t}{cv}\right)} \\ &\qquad \leqslant e^{-\frac{t^2}{2c^2 v}}, \quad t \in [0, \mathbb{E}_{X_1, \cdots, X_n}\{Z\}]. \end{aligned} \tag{C.127}$$

The above mentioned result can be rewritten and we can state that with probability $(1 - e^{-x})$:

$$\mathbb{E}_{X_1, \cdots, X_n}\{Z\} - Z \leqslant cv\widehat{\phi}\left(\frac{x}{v}\right), \tag{C.128}$$

$$\mathbb{E}_{X_1, \cdots, X_n}\{Z\} - Z \leqslant \sqrt{2c^2 xv}. \tag{C.129}$$

Here the proofs are not reported but they can be retrieved from [48, 133, 134].

References

[1] Y. S. Abu-Mostafa. The vapnik-chervonenkis dimension: Information versus complexity in learning. *Neural Computation*, 1(3):312–317, 1989.

[2] R. Ahlswede, P. Gács, and J. Körner. Bounds on conditional probabilities with applications in multi-user communication. *Probability Theory and Related Fields*, 34(2):157–177, 1976.

[3] A. Ambroladze, E. Parrado-Hernández, and J. Shawe-Taylor. Tighter pac-bayes bounds. *Advances in neural information processing systems*, 2006.

[4] O. Anava, E. Hazan, S. Mannor, and O. Shamir. Online learning for time series prediction. In *Computational Learning Theory*, 2013.

[5] D. Anguita, A. Boni, and S. Ridella. Evaluating the generalization ability of support vector machines through the bootstrap. *Neural Processing Letters*, 11(1):51–58, 2000.

[6] D. Anguita, L. Ghelardoni, A. Ghio, L. Oneto, and S. Ridella. The 'k' in k-fold cross validation. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2012.

[7] D. Anguita, L. Ghelardoni, A. Ghio, and S. Ridella. A survey of old and new results for the test error estimation of a classifier. *Journal of Artificial Intelligence and Soft Computing Research*, 3(4):229–242, 2013.

[8] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz. Energy efficient smartphone-based activity recognition using fixed-point arithmetic. *Journal of Universal Computer Science*, 19(9):1295–1314, 2013.

[9] D. Anguita, A. Ghio, L. Oneto, and S. Ridella. The impact of unlabeled patterns in rademacher complexity theory for kernel classifiers. In *Advances in Neural Information Processing Systems*, 2011.

[10] D. Anguita, A. Ghio, L. Oneto, and S. Ridella. In-sample and out-of-sample model selection and error estimation for support vector machines. *IEEE Transactions on Neural Networks and Learning Systems*, 23(9):1390–1406, 2012.

[11] D. Anguita, A. Ghio, L. Oneto, and S. Ridella. In-sample model selection for trimmed hinge loss support vector machine. *Neural processing letters*, 36(3):275–283, 2012.

[12] D. Anguita, A. Ghio, L. Oneto, and S. Ridella. A support vector machine classifier from a bit-constrained, sparse and localized hypothesis space. In *IEEE International Joint Conference on Neural Networks*, 2013.

[13] D. Anguita, A. Ghio, L. Oneto, and S. Ridella. A deep connection between the vapnik-chervonenkis entropy and the rademacher complexity. *IEEE Transactions on Neural Networks and Learning Systems*, 25(12):2202–2211, 2014.

[14] D. Anguita, A. Ghio, L. Oneto, and S. Ridella. Unlabeled patterns to tighten rademacher complexity error bounds for kernel classifiers. *Pattern Recognition Letters*, 37:210–219, 2014.

[15] D. Anguita, A. Ghio, and S. Ridella. Maximal discrepancy for support vector machines. *Neurocomputing*, 74(9):1436–1443, 2011.

[16] Davide Anguita, Alessandro Ghio, Sandro Ridella, and Dario Sterpi. K-fold cross validation for error rate estimate in support vector machines. In *International Conference on Data Mining*, 2009.

[17] S. Anguita, A. Ghio, L. Oneto, and S. Ridella. Maximal discrepancy vs. rademacher complexity for error estimation. In *European Symposium on Artificial Neural Networks*, 2011.

[18] S. Arlot. V-fold cross-validation improved: V-fold penalization. *arXiv preprint arXiv:0802.0566*, 2008.

[19] S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

[20] Sylvain Arlot and Matthieu Lerasle. Why v= 5 is enough in v-fold cross-validation. *ArXiv e-prints*, 2012.

[21] J. Y. Audibert. Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37(4):1591–1646, 2009.

[22] J. Y. Audibert. Pac-bayesian aggregation and multi-armed bandits. *arXiv preprint arXiv:1011.3396*, 2010.

[23] J. Y. Audibert and O. Bousquet. Pac-bayesian generic chaining. In *Neural Information Processing Systems*, 2003.

[24] J. Y. Audibert and A. B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

[25] P. L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48(1-3):85–113, 2002.

[26] P. L. Bartlett, O. Bousquet, and S. Mendelson. Localized rademacher complexities. In *Computational Learning Theory*, 2002.

[27] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[28] P. L. Bartlett, S. R. Kulkarni, and S. E. Posner. Covering numbers for real-valued function classes. *IEEE transactions on information theory*, 43(5):1721–1724, 1997.

[29] P. L. Bartlett, P. M. Long, and R. C. Williamson. Fat-shattering and the learnability of real-valued functions. In *Computational Learning Theory*, 1994.

[30] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2003.

[31] M. Basu and T. K. Ho. *Data complexity in pattern recognition.* Springer Science & Business Media, 2006.

[32] L. Bégin, P. Germain, F. Laviolette, and J. F. Roy. Pac-bayesian theory for transductive learning. In *International Conference on Artificial Intelligence and Statistics*, 2014.

[33] L. Bégin, P. Germain, F. Laviolette, and J. F. Roy. Pac-bayesian bounds based on the rényi divergence. In *International Conference on Artificial Intelligence and Statistics*, 2016.

[34] S. Ben-David, T. Lu, and D. Pál. Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In *computational learning theory*, 2008.

[35] G. Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

[36] V. Bentkus. On hoeffding's inequalities. *The Annals of Probability*, 32(2):1650–1673, 2004.

[37] D. Berend and A. Kontorovitch. Consistency of weighted majority votes. In *Neural Information Processing Systems*, 2014.

[38] S. Bernstein. On a modification of chebyshev's inequality and of the error formula of laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math*, 1(4):38–49, 1924.

[39] C. M. Bishop. *Neural networks for pattern recognition.* Oxford university press, 1995.

[40] G. Blanchard and P. Massart. Discussion: Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2664–2671, 2006.

[41] A. Blum and M. Hardt. The ladder: A reliable leaderboard for machine learning competitions. In *International Conference on Machine Learning*, 2015.

[42] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Occam's razor. *Information processing letters*, 24(6):377–380, 1987.

[43] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

[44] S. G. Bobkov and M. Ledoux. On modified logarithmic sobolev inequalities for bernoulli and poisson measures. *journal of functional analysis*, 156(2):347–365, 1998.

[45] C. E. Bonferroni. *Teoria statistica delle classi e calcolo delle probabilita.* Libreria internazionale Seeber, 1936.

[46] S. Boucheron, G. Lugosi, and P. Massart. A sharp concentration inequality with applications. *Random Structures & Algorithms*, 16(3):277–292, 2000.

[47] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence.* Oxford university press, 2013.

[48] O. Bousquet. A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500, 2002.

[49] O. Bousquet and A. Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

[50] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[51] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[52] A. Cannon, J. M. Ettinger, D. Hush, and C. Scovel. Machine learning with data dependent hypothesis classes. *The Journal of Machine Learning Research*, 2:335–358, 2002.

[53] B. P. Carlin and T. A. Louis. *Bayesian methods for data analysis*. CRC Press, 2008.

[54] G. Casella and R. L. Berger. *Statistical inference*. Duxbury Pacific Grove, CA, 2002.

[55] O. Catoni. *Pac-Bayesian Supervised Classification*. Institute of Mathematical Statistics, 2007.

[56] K. Chaudhuri and D. Hsu. Sample complexity bounds for differentially private learning. In *Conference on Learning Theory*, 2011.

[57] K. Chaudhuri, D. J. Hsu, and S. Song. The large margin mechanism for differentially private maximization. In *Neural Information Processing Systems*, 2014.

[58] K. Chaudhuri and S. A. Vinterbo. A stability-based validation procedure for differentially private machine learning. In *Neural Information Processing Systems*, 2013.

[59] X. Chen. A link between binomial parameters and means of bounded random variables. *arXiv preprint arXiv:0802.3946*, 2008.

[60] X. Chen. Multistage estimation of bounded-variable means. *arXiv preprint arXiv:0809.4679*, 2008.

[61] V. Cherkassky. Model complexity control and statistical learning theory. *Natural computing*, 1(1):109–133, 2002.

[62] V. Cherkassky and F. Mulier. Vapnik-chervonenkis (vc) learning theory and its applications. *IEEE Transactions on Neural Networks*, 10(5):985–987, 1999.

[63] V. Cherkassky and F. M. Mulier. *Learning from data: concepts, theory, and methods*. John Wiley & Sons, 2007.

[64] V. Cherkassky, X. Shao, F. M. Mulier, and V. N. Vapnik. Model complexity control for regression using vc generalization bounds. *IEEE Transactions on Neural Networks*, 10(5):1075–1089, 1999.

[65] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 1952.

[66] R. Christensen, W. Johnson, A. Branscum, and T. E. Hanson. *Bayesian ideas and data analysis: an introduction for scientists and statisticians*. CRC Press, 2011.

[67] C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, pages 404–413, 1934.

[68] D. Corfield, B. Schölkopf, and V. N. Vapnik. Falsificationism and statistical learning theory: Comparing the popper and vapnik-chervonenkis dimensions. *Journal for General Philosophy of Science*, 40(1):51–58, 2009.

[69] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth. On the lambert w function. *Advances in Computational Mathematics*, 5(1):329–359, 1996.

[70] C. Cortes, M. Kloft, and M. Mohri. Learning kernels using local rademacher complexity. In *Advances in Neural Information Processing Systems*, 2013.

[71] C. Cortes and V. N. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[72] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[73] I. Csiszar and J. Körner. *Information theory: coding theorems for discrete memoryless systems.* Cambridge University Press, 2011.

[74] A. Dembo. Information inequalities and concentration of measure. *The Annals of Probability*, 25(2):927–939, 1997.

[75] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition.* Springer, 1996.

[76] L. Devroye and T. Wagner. Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Transactions on Information Theory*, 25(2):202–207, 1979.

[77] V. Dhar. Data science and prediction. *Communications of the ACM*, 56(12):64–73, 2013.

[78] R. Dietrich, M. Opper, and H. Sompolinsky. Statistical mechanics of support vector networks. *Physical review letters*, 82(14):2975, 1999.

[79] M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

[80] N. R. Draper, H. Smith, and E. Pownell. *Applied regression analysis.* Wiley New York, 1966.

[81] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. N. Vapnik. Support vector regression machines. In *Advances in neural information processing systems*, 1997.

[82] C. Dwork. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, 2008.

[83] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. Generalization in adaptive data analysis and holdout reuse. In *Neural Information Processing Systems*, 2015.

[84] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. Preserving statistical validity in adaptive data analysis. In *Annual ACM Symposium on Theory of Computing*, 2015.

[85] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015.

[86] C. Dwork and J. Lei. Differential privacy and robust statistics. In *Annual ACM Symposium on Theory of computing*, pages 371–380, 2009.

[87] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):1–277, 2014.

[88] C. Dwork, G. N. Rothblum, and S. Vadhan. Boosting and differential privacy. In *IEEE Annual Symposium on Foundations of Computer Science*, 2010.

[89] B. Efron. *The jackknife, the bootstrap and other resampling plans.* SIAM, 1982.

[90] B. Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, 1992.

[91] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap.* CRC press, 1994.

[92] R. El-Yaniv and D. Pechyony. Transductive rademacher complexity and its applications. *Journal of Artificial Intelligence Research*, 35(1):193, 2009.

[93] A. Elisseeff, T. Evgeniou, and M. Pontil. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6:55–79, 2005.

[94] V. Feldman and D. Xiao. Sample complexity bounds on differentially private learning via communication complexity. In *Conference on Learning Theory*, 2014.

[95] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we need hundreds of classifiers to solve real world classification problems. *Journal of machine learning research*, 15(1):3133–3181, 2014.

[96] S. Floyd and M. Warmuth. Sample compression, learnability, and the vapnik-chervonenkis dimension. *Machine learning*, 21(3):269–304, 1995.

[97] M. R. Forster. *Notice: No Free Lunches for Anyone, Bayesians Included.* University of Wisconsin-Madison Madison Department of Philosophy, 2005.

[98] L. Friedland, D. Jensen, and M. Lavine. Copy or coincidence? a model for detecting social influence and duplication events. In *International Conference on Machine Learning*, 2013.

[99] A. Friedman and A. Schuster. Data mining with differential privacy. In *ACM international conference on Knowledge discovery and data mining*, 2010.

[100] K. Fukunaga and D. M. Hummels. Leave-one-out procedures for non-parametric error estimates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(4):421–423, 1989.

[101] S. Geisser and W. O. Johnson. *Modes of parametric statistical inference*, volume 529. John Wiley & Sons, 2006.

[102] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*, volume 2. Taylor & Francis, 2014.

[103] P. Germain, A. Lacasse, F. Laviolette, and M. Marchand. Pac-bayesian learning of linear classifiers. In *International Conference on Machine Learning*, 2009.

[104] P. Germain, A. Lacasse, F. Laviolette, M. Marchand, and J. F. Roy. Risk bounds for the majority vote: From a pac-bayesian analysis to a learning algorithm. *The Journal of Machine Learning Research*, 16(4):787–860, 2015.

[105] P. Germain, A. Lacoste, M. Marchand, S. Shanian, and F. Laviolette. A pac-bayes sample-compression approach to kernel methods. In *International Conference on Machine Learning*, 2011.

[106] M. Gönen and E. Alpaydın. Multiple kernel learning algorithms. *Journal of machine learning research*, 12:2211–2268, 2011.

[107] S. Gopal, B. Bai, Y. Yang, and A. Niculescu-Mizil. Bayesian models for large-scale hierarchical classification. In *Neural Information Processing Systems*, 2013.

[108] S. Greengard. Privacy matters. *Communications of ACM*, 51(9):17–18, 2008.

[109] P. D. Grünwald. *The minimum description length principle.* MIT press, 2007.

[110] I. Guyon, A. Saffari, G. Dror, and G. Cawley. Model selection: Beyond the bayesian/frequentist divide. *Journal of Machine Learning research*, 11:61–87, 2010.

[111] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression.* Springer Science & Business Media, 2006.

[112] J. B. S. Haldane. A note on inverse probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 28(1):55–61, 1932.

[113] J. Han, J. Pei, and M. Kamber. *Data mining: concepts and techniques.* Elsevier, 2011.

[114] M. Hardt, K. Ligett, and F. McSherry. A simple and practical algorithm for differentially private data release. In *Neural Information Processing Systems*, 2012.

[115] M. Hardt and J. Ullman. Preventing false discovery in interactive data analysis is hard. In *IEEE Annual Symposium on Foundations of Computer Science*, 2014.

[116] J. Harold. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

[117] R. Herbrich and T. Graepel. A pac-bayesian margin bound for linear classifiers. *IEEE Transactions on Information Theory*, 48(12):3140–3150, 2002.

[118] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[119] A. Inoue and L. Kilian. In-sample or out-of-sample tests of predictability: Which one should we use? *Econometric Reviews*, 23(4):371–402, 2005.

[120] J. P. A. Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

[121] K. E. Iverson. A programming language. In *ACM spring joint computer conference*, 1962.

[122] P. Jain, P. Kothari, and A. Thakurta. Differentially private online learning. In *Conference on Learning Theory*, 2012.

[123] P. Jain and A. Thakurta. Differentially private learning with kernels. In *International Conference on Machine Learning*, 2013.

[124] P. Jain and A. G. Thakurta. (near) dimension independent risk bounds for differentially private learning. In *International Conference on Machine Learning*, 2014.

[125] S. Janson, T. Luczak, and A. Rucinski. *Random graphs.* John Wiley & Sons, 2011.

[126] J. L. W. V. Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta mathematica*, 30(1):175–193, 1906.

[127] N. L. Johnson, A. W. Kemp, and S. Kotz. *Univariate discrete distributions.* John Wiley & Sons, 2005.

[128] M. Kääriäinen. Generalization error bounds using unlabeled data. In *Learning Theory*, 2005.

[129] P. Kairouz, S. Oh, and P. Viswanath. Secure multi-party differential privacy. In *Neural Information Processing Systems*, 2015.

[130] S. Katrenko and M. Van Zaanen. Rademacher complexity and grammar induction algorithms: what it may (not) tell us. In *Grammatical Inference: Theoretical Results and Applications*, 2010.

[131] M. Kearns and D. Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11(6):1427–1453, 1999.

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

[133] T. Klein. Une inégalité de concentration à gauche pour les processus empiriques. *Comptes Rendus Mathematique*, 334(6):501–504, 2002.

[134] T. Klein and E. Rio. Concentration around the mean for maxima of empirical processes. *The Annals of Probability*, 33(3):1060–1077, 2005.

[135] P. Klesk and M. Korzen. Sets of approximating functions with finite vapnik-chervonenkis dimension for nearest-neighbors algorithms. *Pattern Recognition Letters*, 32(14):1882–1893, 2011.

[136] M. Kloft and G. Blanchard. The local rademacher complexity of lp-norm multiple kernel learning. In *Neural Information processing systems*, 2011.

[137] R. Koavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence*, 1995.

[138] A. N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Berlin: Julius Springer, 1933.

[139] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

[140] V. Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

[141] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems.* Springer, 2011.

[142] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[143] A. Kumar, A. Saha, and H. Daume. Co-regularization based semi-supervised domain adaptation. In *Advances in Neural Information Processing Systems*, 2010.

[144] M. J. Kusner, J. Gardner, R. Garnett, and K. Weinberger. Differentially private bayesian optimization. In *International Conference on Machine Learning*, 2015.

[145] John L. Clever methods of overfitting. In *http://hunch.net/?p=22*, 2005.

[146] A. Lacasse, F. Laviolette, M. Marchand, P. Germain, and N. Usunier. Pac-bayes bounds for the risk of the majority vote and the variance of the gibbs classifier. In *Neural Information Processing Systems*, 2006.

[147] J. Langford. Tutorial on practical prediction theory for classification. *Journal of machine learning research*, 6:273–306, 2005.

[148] J. Langford and D. McAllester. Computable shell decomposition bounds. *Journal of Machine Learning Research*, 5:529–547, 2004.

[149] J. Langford and D. A. McAllester. Computable shell decomposition bounds. In *Computational Learning Theory*, 2000.

[150] J. Langford and M. Seeger. *Bounds for averaging classifiers.* Technical report, Carnegie Mellon, Departement of Computer Science,, 2001.

[151] P. S. Laplace. *Essai philosophique sur les probabilités.* Bachelier, 1825.

[152] F. Laviolette and M. Marchand. Pac-bayes risk bounds for sample-compressed gibbs classifiers. In *International conference on Machine learning*, 2005.

[153] F. Laviolette and M. Marchand. Pac-bayes risk bounds for stochastic averages and majority votes of sample-compressed classifiers. *Journal of Machine Learning Research*, 8(7):1461–1487, 2007.

[154] M. Ledoux. Isoperimetry and gaussian analysis. In *Lectures on probability theory and statistics*, pages 165–294, 1996.

[155] M. Ledoux. On talagrand's deviation inequalities for product measures. *ESAIM: Probability and statistics*, 1:63–87, 1997.

[156] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer, 2011.

[157] M. M. S. Lee, S. S. Keerthi, C. J. Ong, and D. DeCoste. An efficient method for computing leave-one-out error in support vector machines with gaussian kernels. *IEEE Transactions on Neural Networks*, 15(3):750–757, 2004.

[158] J. Lei. Differentially private m-estimators. In *Neural Information Processing Systems*, 2011.

[159] G. Lever, F. Laviolette, and J. Shawe-Taylor. Distribution-dependent pac-bayes priors. In *Algorithmic Learning Theory*, 2010.

[160] G. Lever, F. Laviolette, and J. Shawe-Taylor. Tighter pac-bayes bounds through distribution-dependent priors. *Theoretical Computer Science*, 473:4–28, 2013.

[161] M. Li, J. Tromp, and P. Vitányi. *Sharpening Occam razor*. Springer, 2002.

[162] N. Littlestone and M. Warmuth. Relating data compression and learnability. In *Technical report, University of California, Santa Cruz*, 1986.

[163] B. London, B. Huang, B. Taskar, L. Getoor, and S. Cruz. Pac-bayesian collective stability. In *Artificial Intelligence and Statistics*, 2014.

[164] G. Lugosi and M. Wegkamp. Complexity regularization via localized random penalties. *The Annals of Statistics*, 32(4):1679–1697, 2004.

[165] M. Magdon-Ismail. No free lunch for noise prediction. *Neural computation*, 12(3):547–564, 2000.

[166] E. Mammen and A. B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

[167] K. Marton. A simple proof of the blowing-up lemma. *IEEE Transaction on Information Theory*, 24:857–866, 1966.

[168] K. Marton. Bounding d-distance by informational divergence: a method to prove measure concentration. *The Annals of Probability*, 24(2):857–866, 1996.

[169] K. Marton. A measure concentration inequality for contracting markov chains. *Geometric & Functional Analysis GAFA*, 6(3):556–571, 1996.

[170] P. Massart. *Optimal constants for Hoeffding type inequalities*. Université de Paris-Sud. Département de Mathématique, 1998.

[171] P. Massart. About the constants in talagrand's concentration inequalities for empirical processes.(english summary). *Annals of Probability*, 28(2):863–884, 2000.

[172] A. Maurer. A note on the pac bayesian theorem. *arXiv preprint cs/0411099*, 2004.

[173] A. Maurer. Concentration inequalities for functions of independent variables. *Random Structures & Algorithms*, 29(2):121–138, 2006.

[174] A. Maurer. A second-order look at stability and generalization. In *Conference on Learning Theory*, pages 1461–1475, 2017.

[175] A. Maurer and M. Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

[176] D. A. McAllester. Some pac-bayesian theorems. In *Computational learning theory*, 1998.

[177] D. A. McAllester. Pac-bayesian model averaging. In *Computational Learning Theory*, 1999.

[178] D. A. McAllester. Pac-bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, 2003.

[179] D. A. McAllester. Simplified pac-bayesian margin bounds. In *Learning Theory and Kernel Machines*, 2003.

[180] D. A. McAllester and T. Akinbiyi. Pac-bayesian theory. In *Empirical Inference*, 2013.

[181] C. McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

[182] C. McDiarmid. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, pages 195–248, 1998.

[183] S. Mendelson. Learning without concentration. *Journal of the ACM (JACM)*, 62(3):21, 2015.

[184] L. Ming and P. Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer Heidelberg, 1997.

[185] E. Morvant. *Apprentissage de vote de majorité pour la classification supervisée et l'adaptation de domaine: approches PAC-Bayésiennes et combinaison de similarités*. Aix-Marseille Université, 2013.

[186] S. Mukherjee, P. Niyogi, T. Poggio, and R. Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1):161–193, 2006.

[187] S. Mukherjee, P. Tamayo, S. Rogers, R. Rifkin, A. Engle, C. Campbell, T. R. Golub, and J. P. Mesirov. Estimating dataset size requirements for classifying dna microarray data. *Journal of Computational Biology*, 10(2):119–142, 2003.

[188] K.R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2):181–201, 2001.

[189] K. Nissim and U. Stemmer. On the generalization properties of differential privacy. *arXiv preprint arXiv:1504.05800*, 2015.

[190] S. Nitzan and J. Paroush. Optimal decision rules in uncertain dichotomous choice situations. *International Economic Review*, 23(2):289–97, 1982.

[191] S. Oh and P. Viswanath. The composition theorem for differential privacy. In *International Conference on Machine Learning*, 2015.

[192] L. Oneto, D. Anguita, and S. Ridella. A local vapnik-chervonenkis complexity. *Neural Networks*, 82:62–75, 2016.

[193] L. Oneto, D. Anguita, and S. Ridella. Pac-bayesian analysis of distribution dependent priors: Tighter risk bounds and stability analysis. *Pattern Recognition Letters*, 80:200–207, 2016.

[194] L. Oneto, A. Ghio, D. Anguita, and S. Ridella. An improved analysis of the rademacher data-dependent bound using its self bounding property. *Neural Networks*, 44:107–111, 2013.

[195] L. Oneto, A. Ghio, S. Ridella, and D. Anguita. Fully empirical and data-dependent stability-based bounds. *IEEE Transactions on Cybernetics*, 45(9):1913–1926, 2015.

[196] L. Oneto, A. Ghio, S. Ridella, and D. Anguita. Global rademacher complexity bounds: From slow to fast convergence rates. *Neural Processing Letters*, 43(2):567–602, 2015.

[197] L. Oneto, A. Ghio, S. Ridella, and D. Anguita. Learning resource-aware models for mobile devices: from regularization to energy efficiency. *Neurocomputing*, 169(-):225–235, 2015.

[198] L. Oneto, A. Ghio, S. Ridella, and D. Anguita. Local rademacher complexity: Sharper risk bounds with and without unlabeled samples. *Neural Networks*, 65:115–125, 2015.

[199] L. Oneto, B. Pilarz, A. Ghio, and D. Anguita. Model selection for big data: Algorithmic stability and bag of little bootstraps on gpus.

In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2015.

[200] L. Oneto, S. Ridella, and D. Anguita. Learning hardware-friendly classifiers through algorithmic stability. *ACM Transaction on Embedded Computing*, 15(2):23:1–23:29, 2016.

[201] L. Oneto, S. Ridella, and D. Anguita. Tuning the distribution dependent prior in the pac-bayes framework based on empirical data. In *ESANN*, 2016.

[202] L. Oneto, S. Ridella, and D. Anguita. Differential privacy and generalization: Sharper bounds with applications. *Pattern Recognition Letters*, 89:31–38, 2017.

[203] L. Oneto, S. Ridella, and D. Anguita. Generalization performances of randomized classifiers and algorithms built on data dependent distributions. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2017.

[204] M. Opper. Statistical mechanics of learning: Generalization. In *The Handbook of Brain Theory and Neural Networks*, 1995.

[205] M. Opper, W. Kinzel, J. Kleinz, and R. Nehl. On the ability of the optimal perceptron to generalise. *Journal of Physics A: Mathematical and General*, 23(11):L581, 1990.

[206] D. Panchenko. Some extensions of an inequality of vapnik and chervonenkis. *Electronic Communications in Probability*, 7:55–65, 2002.

[207] E. Parrado-Hernández, A. Ambroladze, J. Shawe-Taylor, and S. Sun. Pac-bayes bounds with data dependent priors. *The Journal of Machine Learning Research*, 13(1):3507–3531, 2012.

[208] J. M. R. Parrondo and C. Van den Broeck. Vapnik-chervonenkis bounds for generalization. *Journal of Physics A: Mathematical and General*, 26(9):2211, 1993.

[209] J. L. Paulo and F. G. T. Azzam. The use of artificial neural networks in decision support in cancer: A systematic review. *Neural Networks*, 19(4):408–415, 2006.

[210] C. S. Peirce. *Collected papers of charles sanders peirce.* Harvard University Press, 1974.

[211] M. S. Pinsker. *Information and information stability of random variables and processes.* Izv. Akad. Nauk, 1960.

[212] T. Poggio, R. Rifkin, S. Mukherjee, and P. Niyogi. General conditions for predictivity in learning theory. *Nature*, 428(6981):419–422, 2004.

[213] R. H. Popkin and A. Stroll. *Philosophy made simple.* Made Simple Books, 1993.

[214] K. R. Popper. *The logic of scientific discovery.* London Hutchinson, 1959.

[215] Y. V. Prokhorov. An extremal problem in probability theory. *Theory of Probability & Its Applications*, 4(2):201–203, 1959.

[216] J. R. Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[217] L. Ralaivola, M. Szafranski, and G. Stempfel. Chromatic pac-bayes bounds for non-iid data: Applications to ranking and stationary β-mixing processes. *The Journal of Machine Learning Research*, 11:1927–1956, 2010.

[218] R. H. Randles, T. P. Hettmansperger, and G. Casella. Introduction to the special issue: Nonparametric statistics. *Statistical Science*, 19(4):561–561, 2004.

[219] R. B. Rao, G. Fung, and R. Rosales. On the dangers of cross-validation: An experimental evaluation. In *International Conference on Data Mining*, 2008.

[220] C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning.* MIT press Cambridge, 2006.

[221] N. Reid and D. R. Cox. On some principles of statistical inference. *International Statistical Review*, 83(2):293–308, 2015.

[222] E. Rio. Inégalités de concentration pour les processus empiriques de classes de parties. *Probability Theory and Related Fields*, 119(2):163–175, 2001.

[223] R. Rogers, S. Vadhan, H. Lim, and M. Gaboardi. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *International Conference on Machine Learning*, 2016.

[224] W. H. Rogers and T. J. Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, pages 506–514, 1978.

[225] L. Rosasco, E. De Vito, A. Caponnetto, M. Piana, and A. Verri. Are loss functions all the same? *Neural Computation*, 16(5):1063–1076, 2004.

[226] D. S. Rosenberg and P. L. Bartlett. The rademacher complexity of co-regularized kernel classes. In *International Conference on Artificial Intelligence and Statistics*, 2007.

[227] J. F. Roy, M. Marchand, and F. Laviolette. From pac-bayes bounds to quadratic programs for majority votes. In *International Conference on Machine Learning*, 2011.

[228] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

[229] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.

[230] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine learning*, 37(3):297–336, 1999.

[231] M. Seeger. Pac-bayesian generalisation error bounds for gaussian process classification. *The Journal of Machine Learning Research*, 3:233–269, 2002.

[232] M. Seeger. *Bayesian Gaussian process models: PAC-Bayesian generalisation error bounds and sparse approximations.* PhD thesis, University of Edinburgh, 2003.

[233] Y. Seldin, P. Auer, J. S. Shawe-Taylor, R. Ortner, and F. Laviolette. Pac-bayesian analysis of contextual bandits. In *Neural Information Processing Systems*, 2011.

[234] Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer. Pac-bayesian inequalities for martingales. *Information Theory, IEEE Transactions on*, 58(12):7086–7093, 2012.

[235] Y. Seldin and N. Tishby. Pac-bayesian generalization bound for density estimation with application to co-clustering. In *International Conference on Artificial Intelligence and Statistics*, 2009.

[236] Y. Seldin and N. Tishby. Pac-bayesian analysis of co-clustering and beyond. *The Journal of Machine Learning Research*, 11:3595–3646, 2010.

[237] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms.* Cambridge university press, 2014.

[238] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

[239] X. Shao, V. Cherkassky, and W. Li. Measuring the vc-dimension using optimized experimental design. *Neural computation*, 12(8):1969–1986, 2000.

[240] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE transactions on Information Theory*, 44(5):1926–1940, 1998.

[241] J. Shawe-Taylor and J. Langford. Pac-bayes & margins. *Neural information processing systems*, 2002.

[242] J. Shawe-Taylor and R. C. Williamson. A pac analysis of a bayesian estimator. In *Computational learning theory*, 1997.

[243] S. Shelah. A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41(1):247–261, 1972.

[244] J. R. Shewchuk. An introduction to the conjugate gradient method without the agonizing pain, 1994.

[245] A. Smith and A. Thakurta. Differentially private feature selection via stability arguments, and the robustness of the lasso. In *Conference on Learning Theory*, 2013.

[246] R. J. Solomonoff. Complexity based induction systems: Comparisons and convergence theorems. *IEEE Transactions on Information Theory*, 24(4):422–432, 1978.

[247] S. Song, K. Chaudhuri, and A. D. Sarwate. Stochastic gradient descent with differentially private updates. In *IEEE Global Conference on Signal and Information Processing*, 2013.

[248] N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. In *Advances in neural information processing systems*, 2010.

[249] A. Statnikov, M. Henaff, V. Narendra, K. Konganti, Z. Li, L. Yang, Z. Pei, M. J Blaser, C. F. Aliferis, and A. V. Alekseyenko. A comprehensive evaluation of multicategory classification methods for microbiomic data. *Microbiome*, 1(1):11, 2013.

[250] T. Steinke and J. Ullman. Interactive fingerprinting codes and the hardness of preventing false discovery. In *Conference on Learning Theory*, 2015.

[251] I. Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51(1):128–142, 2005.

[252] R. Sternberg. *Cognitive psychology.* Cengage Learning, 2008.

[253] S. Sun and J. Shawe-Taylor. Sparse semi-supervised learning using conjugate functions. *The Journal of Machine Learning Research*, 11:2423–2455, 2010.

[254] M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'Institut des Hautes Etudes Scientifiques*, 81(1):73–205, 1995.

[255] M. Talagrand. The missing factor in hoeffding's inequalities. *Annales de l'IHP Probabilités et statistiques*, 31(4):689–702, 1995.

[256] M. Talagrand. New concentration inequalities in product spaces. *Inventiones mathematicae*, 126(3):505–563, 1996.

[257] M. Talagrand. A new look at independence. *The Annals of Probability*, 24(1):1–34, 1996.

[258] P. Tchebichef. Des valeurs moyennes. *Journal de mathématiques pures et appliquées*, 2(12):177–184, 1867.

[259] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[260] I. O. Tolstikhin and Y. Seldin. Pac-bayes-empirical-bernstein inequality. In *Neural Information Processing Systems*, 2013.

[261] C. Tomasi. Learning theory: Past performance and future results. *Nature*, 428(6981):378–378, 2004.

[262] A. B. Tsybakov. *Introduction to nonparametric estimation.* Springer Science & Business Media, 2008.

[263] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[264] T. Van Erven. Pac-bayes mini-tutorial: A continuous union bound. *arXiv preprint arXiv:1405.1580*, 2014.

[265] V. N. Vapnik. *Statistical learning theory.* Wiley New York, 1998.

[266] V. N. Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, 1999.

[267] V. N. Vapnik. *The nature of statistical learning theory.* Springer science & business media, 2013.

[268] V. N. Vapnik and A. J. Červonenkis. *Theorie der Zeichenerkennung.* Akademie-Verlag, 1979.

[269] V. N. Vapnik and S. Kotz. *Estimation of dependences based on empirical data*, volume 41. Springer-Verlag New York, 1982.

[270] V. S Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis. State-of-the-art in privacy preserving data mining. *ACM Sigmod Record*, 33(1):50–57, 2004.

[271] M. Wainberg, B. Alipanahi, and B. J. Frey. Are random forests truly the best classifiers? *The Journal of Machine Learning Research*, 17(1):3837–3841, 2016.

[272] Y. Wang, Y. X. Wang, and A. Singh. Differentially private subspace clustering. In *Neural Information Processing Systems*, 2015.

[273] O. Williams and F. McSherry. Probabilistic inference and differential privacy. In *Neural Information Processing Systems*, 2010.

[274] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[275] J. Wolfowitz. Additive partition functions and a class of statistical hypotheses. *The Annals of Mathematical Statistics*, 13(3):247–279, 1942.

[276] D. H. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural computation*, 8(7):1341–1390, 1996.

[277] Q. Wu and D. X. Zhou. Learning with sample dependent hypothesis spaces. *Computers & Mathematics with Applications*, 56(11):2896–2907, 2008.

[278] M. Younsi. Proof of a combinatorial conjecture coming from the pac-bayesian machine learning theory. *arXiv preprint arXiv:1209.0824*, 2012.

[279] C. Zhang, W. Bian, D. Tao, and W. Lin. Discretized-vapnik-chervonenkis dimension for analyzing complexity of real function classes. *IEEE Transactions on Neural Networks and Learning Systems*, 23(9):1461–1472, 2012.

[280] D. X. Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.

[281] X. Zhu and A. B. Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.

[282] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.