Generalization Bounds: Perspectives from Information Theory and PAC-Bayes

Fredrik Hellström
University College London
f.hellstrom@ucl.ac.uk

Giuseppe Durisi
Chalmers University of Technology
durisi@chalmers.se

Benjamin Guedj
Inria and University College London
benjamin.guedj@inria.fr

Maxim Raginsky
University of Illinois
maxim@illinois.edu

Contents

Generalization Bounds: Perspectives from Information Theory and PAC-Bayes

Fredrik Hellström[1], Giuseppe Durisi[2], Benjamin Guedj[3] and Maxim Raginsky[4]

[1] *University College London, UK; frehells@chalmers.se*

[2] *Chalmers University of Technology, Sweden; durisi@chalmers.se*

[3] *Inria and University College London, France and UK; benjamin.guedj@inria.fr*

[3] *University of Illinois, USA; maxim@illinois.edu*

ABSTRACT

A fundamental question in theoretical machine learning is generalization. Over the past decades, the PAC-Bayesian approach has been established as a flexible framework to address the generalization capabilities of machine learning algorithms, and design new ones. Recently, it has garnered increased interest due to its potential applicability for a variety of learning algorithms, including deep neural networks. In parallel, an information-theoretic view of generalization has developed, wherein the relation between generalization and various information measures has been established. This framework is intimately connected to the PAC-Bayesian approach, and a number of results have been independently discovered in both strands.

In this monograph, we highlight this strong connection and present a unified treatment of PAC-Bayesian and information-theoretic generalization bounds. We present techniques and results that the two perspectives have in common, and discuss the approaches and interpretations that differ. In particular, we demonstrate how many proofs in the area share a modular structure, through which the underlying ideas can be intuited. We pay special attention to the conditional mutual information (CMI) framework; analytical studies of the information complexity of learning algorithms; and the application of the proposed methods to deep learning. This monograph is intended to provide a comprehensive introduction to information-theoretic generalization bounds and their connection to PAC-Bayes, serving as a foundation from which the most recent developments are accessible. It is aimed broadly towards researchers with an interest in generalization and theoretical machine learning.

1

Introduction: On Generalization and Learning

Artificial intelligence and machine learning have emerged as driving forces behind transformative advancements in various fields, becoming increasingly pervasive throughout many industries and in our daily lives. As these technologies continue to gain momentum, the need to develop a deeper understanding of their underlying principles, capabilities, and limitations grows larger. In this monograph, we delve into the theory of machine learning, and more specifically statistical learning theory, where a key topic is the generalization capabilities of learning algorithms.

A learning algorithm is a (potentially stochastic) rule for selecting a hypothesis, given a training data set. Generalization bounds for learning algorithms provide guarantees that the performance, as measured by a loss function, is "good enough," given that the training loss is small, when the hypothesis is subjected to new samples that were not necessarily in the training data. Such bounds are useful for several reasons. When applied in a specific use case, a generalization bound provides a certificate that the hypothesis performs well on new data, provided that the assumptions under which the bound was derived are valid. Furthermore, such bounds can serve as inspiration for the design of new learning algorithms, potentially leading to practical improvements.

Finally, on a deeper level, generalization bounds can enable a more complete understanding of learning algorithms.

While the literature on generalization bounds is vast, making an in-depth review of the full field beyond our scope, we will discuss several key references. Valiant (1984) formalized a model of learnability, called Probably Approximately Correct (PAC) learning. Roughly speaking, a problem is PAC learnable if there exists a learning algorithm such that, for any data distribution, the selected hypothesis has satisfactory performance with high probability. In the preceding decade, Vapnik and Chervonenkis (1971) studied the uniform convergence of certain events. They characterized this convergence in terms of a property of the underlying set that would later be termed the Vapnik-Chervonenkis (VC) dimension, which can be thought of as a measure of complexity. Blumer *et al.* (1989) connected these two topics, and demonstrated that the VC dimension of a hypothesis class characterizes its PAC learnability. We discuss these topics and additional results in more detail in Section 1.3.

The two particular strands in the literature on generalization bounds that will be our main focus throughout this monograph are the PAC-Bayesian and information-theoretic lines of research. Despite the great commonality in techniques and concepts, these two fields have evolved in almost parallel tracks until recently. One objective of the present monograph is to give a unified treatment of the two approaches and highlight their similarities, despite the differing origins. The PAC-Bayesian approach—initiated by McAllester (1998, 1999) and Shawe-Taylor and Williamson (1997), with significant later contributions from, *e.g.*, Catoni (2007)—started as a quest to obtain Bayesian-flavored versions of PAC generalization bounds, as the name implies. PAC bounds are independent of the specific learning algorithm used, as they hold uniformly over the class of possible hypotheses. In contrast, PAC-Bayesian bounds take into account the learning algorithm by explicitly incorporating a distribution over hypotheses—hence the Bayesian suffix.

The effort of relating generalization and information, with a broad interpretation of these terms, has a long history. Conventional wisdom, by way of Occam's razor (Blumer *et al.*, 1987), holds that solutions that are "simpler" in some sense tend to generalize better than their more "complex" counterparts. Many different ways of formalizing complexity measures to capture "information" of some kind have been studied, with some of the earliest examples being the Fisher information of Edgeworth (1908) and Fisher and Russell (1922), the information theory of Shannon (1948), and the Kolmogorov complexity of Kolmogorov (1963) and Solomonoff (1964). In seminal works, Yang and Barron (1999) and Leung and Barron (2006) connected such complexity measures to performance guarantees for density estimation. Other notable information notions in the context of learning include the Akaike information criterion of Akaike (1974), the Bayesian information criterion of Schwarz (1978), and the minimum description length principle, studied by, *e.g.*, Rissanen (1978, 1983) and Barron and Cover (1991), Barron *et al.* (1998) (see the book of Grünwald, 2007 for an in-depth treatment). The particular flavor of information-theoretic approach to generalization that we will focus on can be traced back to the work of Zhang (2006), and more recently, to the seminal works of Russo and Zou (2016) and Xu and Raginsky (2017). In this line of work, the learning algorithm is viewed as a communication channel from the training data to the hypothesis. With this interpretation of the statistical learning process, it is clear that quantities that are common in communication applications, such as the mutual information, have an important role to play.

Despite the historical separation between these lines of work—even within the specific strands, at times—the tools and results that appear in these fields have more similarities than differences, and any discrepancy between them is mainly in the motivation and framing of the work. This may be due to the interdisciplinary nature of the field: it can naturally be covered as statistics, computer science, electrical engineering, and physics.[1] Thus, the reader will not be surprised that many of these results were re-discovered and re-interpreted in many separate contexts, evolving independently. Still, the connection between PAC-Bayesian and information-theoretic generalization bounds has been noted and explored by, *e.g.*, Russo and Zou (2016), Banerjee and Montufar (2021), Grünwald *et al.* (2021), and Alquier (2024). One of the aims of the

[1] Noting this deep connection, Catoni (2007) referred to the PAC-Bayesian approach as the "thermodynamics of statistical learning."

present monograph is to solidify the bridge between these strands of the literature, demonstrating the commonalities in the different approaches.

1.1 Notation and Terminology

To set the stage, we introduce the notation that is used throughout this monograph. Unless otherwise stated, capital letters indicate random variables, with lower-case letters indicating their instances. For random vectors, the same applies, but the letters are in bold. We consider the training examples to lie in a set $\mathcal{Z}$, referred to as the *instance space*. In the context of supervised learning, the instance space is a product between a *feature space* $\mathcal{X}$ and a *label space* $\mathcal{Y}$, so that $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. At its disposal, the learning algorithm has a *training set* $\boldsymbol{Z} = (Z_1, \dots, Z_n) \in \mathcal{Z}^n$, consisting of n training examples.[2] Usually, we assume that the training examples are independent and identically distributed (i.i.d.),[3] with each training example being drawn from a data distribution P_Z on $\mathcal{Z}$. We denote the distribution of $\boldsymbol{Z}$, as well as other product distributions, as $P_{\boldsymbol{Z}} = P_Z^n$. Throughout, we will use the shorthand $[n] = \{1, \dots, n\}$ to refer to the indices of the training samples.

Confronted with the training data, the learner selects a hypothesis W from a set $\mathcal{W}$, called the *hypothesis space*. Again, in supervised learning, $\mathcal{W}$ is typically a subset of all functions from $\mathcal{X}$ to $\mathcal{Y}$, or the parameters of such functions, but the general framework can accommodate other notions of hypothesis. The method by which the learner chooses the hypothesis is described by a (probabilistic) mapping from the training set $\boldsymbol{Z}$ to the hypothesis W, denoted by $P_{W|\boldsymbol{Z}}$ and referred to as a *learning algorithm*. Mathematically, it can be seen as a stochastic kernel, which gives rise to a probability distribution on $\mathcal{W}$ for each instance of $\boldsymbol{Z}$. Note that $P_{W|\boldsymbol{Z}}$ is defined for a specific size n of the training set. We usually assume that the learning algorithm can be adapted to training sets of different sizes, *i.e.*, we assume that $P_{W|\boldsymbol{Z}}$

[2] Despite conventionally being called a "set," $\boldsymbol{Z}$ is a vector: its elements are ordered, and elements are allowed to be repeated.

[3] This assumption is classical in statistical learning theory. Nevertheless, we will cover recent results that allow one to relax and even remove it (see Chapters 5 and 9).

is defined for every n. While there is often a natural relation between these conditional distributions for various n, we do not require that they are related in general.

The quality of a specific hypothesis $w \in \mathcal{W}$ with respect to a sample $z \in \mathcal{Z}$ is measured by a *loss function*, $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}^+$. To give some classical examples of loss functions, consider supervised learning, where the sample is decomposed into features and labels (or inputs and outputs) as $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$ and the hypotheses $w \in \mathcal{W}$ are functions $w : \mathcal{X} \to \mathcal{Y}$. For classification, where the label space $\mathcal{Y}$ is discrete, a typical loss function is the classification error $\ell(w, z) = 1\{w(x) \neq y\}$. Here, $1\{\cdot\}$ denotes the indicator function. For regression, where the label space is continuous, a common choice is the squared loss $\ell(w, z) = (w(x) - y)^2$.

The true goal of the learner is to select a hypothesis that performs well on fresh data from the distribution P_Z, as measured by the loss function. This is formalized by the *population loss*

$$L_{P_Z}(w) = \mathbb{E}_{P_Z}[\ell(w, Z)] ,$$

sometimes referred to as the (true) *risk* of a hypothesis. A key feature of the learning problem is that the true data distribution is assumed to be unknown, which implies that the population loss cannot be computed by the learner. However, by averaging the loss function over training data, the learner obtains the *training loss*

$$L_{\boldsymbol{Z}}(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w, Z_i),$$

which serves as an estimate of the population loss. The training loss is also known as the *empirical risk*. A natural procedure for selecting a hypothesis is to minimize the training loss. This is referred to as *empirical risk minimization* (ERM), and is successful in finding a hypothesis with low population loss if the difference between population loss and training loss is small. This is measured by the *generalization error*

$$\mathrm{gen}(w, \boldsymbol{Z}) = L_{P_Z}(w) - L_{\boldsymbol{Z}}(w),$$

which is also called the *generalization gap*.

1.2 Flavors of Generalization

Since the randomized learning algorithm is described by a conditional probability distribution $P_{W|\boldsymbol{Z}}$, bounds on the generalization error $\operatorname{gen}(W, \boldsymbol{Z})$ come in a variety of forms. We now introduce three canonical forms that have been studied in the information-theoretic and PAC-Bayesian literature.

Firstly, one possibility that has been widely considered in the information-theoretic strand of the literature is to bound the average generalization error $\mathbb{E}_{P_{W\boldsymbol{Z}}}[\operatorname{gen}(W, \boldsymbol{Z})]$. Performing an average analysis can often simplify mathematical derivations, and lead to some insights about the studied algorithms. The works of Russo and Zou (2016) and Xu and Raginsky (2017) both focus on this setting, and the mutual information between training data and hypothesis naturally arises as a fundamental quantity in upper bounds for the average generalization error. In Section 2.3, we introduce a first such average generalization bound, as a warm-up to the more general theory presented later in this monograph. The particular features that are relevant specifically for this scenario are discussed in more detail in Chapter 4.

Secondly, in practical situations, we may be given only one instance of a training set, so an arguably more pertinent question is if we can bound the generalization error with high probability over the draw of the data. In the PAC-Bayesian literature, initiated in the works of Shawe-Taylor and Williamson (1997) and McAllester (1998), most bounds are on the generalization error when averaged over the learning algorithm, $\mathbb{E}_{P_{W|\boldsymbol{Z}}}[\operatorname{gen}(W, \boldsymbol{Z})]$, and hold with probability at least $1 - \delta$ under $P_{\boldsymbol{Z}}$ for some confidence parameter $\delta \in (0, 1)$. The change in perspective in the PAC-Bayesian approach, as compared to the classical statistical learning literature, is significant. We no longer ask whether there are specific hypotheses w that perform well: instead, we ask if there are distributions $P_{W|\boldsymbol{Z}}$ over hypotheses that do. To highlight the conceptual connection to Bayesian statistics, the distribution $P_{W|\boldsymbol{Z}}$ is usually termed *posterior*. This distribution is compared, via information-theoretic metrics, to a reference measure Q_W called the *prior*. Another significant feature that is shared among many PAC-Bayesian bounds is that they hold uniformly for all choices of posterior. This, and other important properties of PAC-Bayesian bounds, are detailed in Section 5.2.

Finally, we may be interested in the generalization error when we have a single training set and we use our learning algorithm to select a single hypothesis. Thus, we seek bounds on $\operatorname{gen}(W, \boldsymbol{Z})$ that hold with probability at least $1-\delta$ under $P_{W\boldsymbol{Z}}$. In this monograph, we will call this the *single-draw* setting, following Catoni (2007), since we are concerned with a single draw of both data and hypothesis. This type of bound has appeared sporadically in both the information-theoretic and PAC-Bayesian literature. While this type of bound can arguably be the most relevant in practice—for instance, in deep learning (discussed in Chapter 8), one typically uses a deterministic neural network obtained via one instantiation of a randomized learning algorithm—it comes with some drawbacks. For instance, since the probability is computed with respect to the joint distribution $P_{W\boldsymbol{Z}}$, any single-draw bound is by definition a statement pertaining to a particular posterior $P_{W|\boldsymbol{Z}}$. Thus, we lose uniformity over posteriors. Furthermore, for the information-theoretic bounds that we discuss here, we need a stronger technical requirement on the absolute continuity of the distributions involved—at least for data-dependent bounds. We will discuss this type of bounds in Section 5.3.

It should be stressed that the terminology used here is not universally accepted, and different names are used by different authors. Furthermore, bounds of all types have been studied in both the PAC-Bayesian and information-theoretic strands of the literature. For instance, average bounds have been referred to as "PAC-Bayesian type" bounds (Dalalyan and Salmon, 2012; Salmon and Dalalyan, 2011) or mean approximately correct (MAC)-Bayesian bounds (Grünwald *et al.*, 2021). Single-draw bounds have been referred to as pointwise or de-randomized PAC-Bayesian bounds (Alquier and Biau, 2013; Catoni, 2007; Guedj and Alquier, 2013). The term de-randomized PAC-Bayesian bound has also been used for bounds that specifically apply to the average hypothesis, that is, bounds on $\operatorname{gen}(\mathbb{E}_{P_{W|\boldsymbol{Z}}}[W], \boldsymbol{Z})$ that hold with probability $1-\delta$ under $P_{\boldsymbol{Z}}$ (Banerjee and Montufar, 2021) (such variants will be discussed in Section 5.4). However, throughout this monograph, we will use the terms defined above.

The framework of PAC learnability and the associated uniform-convergence bounds that we mentioned earlier do not fit exactly into any of the flavors that we have mentioned so far (although the single-draw bounds are most closely related). In the following section, we give a formal definition of PAC learnability, and provide an overview of some generalization bounds based on uniform convergence.

1.3 Uniform Convergence-Flavored Generalization Bounds

As previously indicated, demonstrating PAC learnability for a hypothesis class boils down to a very strong type of uniform convergence result. Roughly speaking, PAC learnability requires that for any data distribution P_Z, there is a learning algorithm that, with sufficient training data, is arbitrarily close to the optimal population loss. As it turns out, PAC learnability is equivalent to uniform convergence, defined below (Shalev-Shwartz and Ben-David, 2014, Chapter 4).

Definition 1.1 (Uniform convergence)**.** The hypothesis class $\mathcal{W}$ has the *uniform convergence property* if there exists a function $m : (0,1)^2 \to \mathbb{N}$ such that, for every $\epsilon, \delta \in (0,1)$ and every data distribution P_Z, the following holds: if $\boldsymbol{Z}$ contains $n \geq m(\epsilon, \delta)$ i.i.d. samples from P_Z, we have with probability at least $1 - \delta$ that

$$|L_{\boldsymbol{Z}}(w) - L_{P_Z}(w)| \leq \epsilon \quad \text{for all } w \in \mathcal{W}. \tag{1.1}$$

The function m is called the *sample complexity*.

Thus, if a hypothesis class satisfies the uniform convergence property, we can obtain generalization bounds that are uniform over both data distributions and hypotheses. The attractiveness of these bounds is clear: no matter what data you are dealing with, independent of the learning algorithm you use, you can trust that the training loss gives a good indication of your population loss. At the moment, it unfortunately seems as if such requirements are too strict for many modern machine learning settings, such as deep neural networks.[4] For this model class,

[4] This is not meant to imply that the bounds discussed in this section have no hope of describing modern models, such as deep neural networks. Indeed, promising steps toward this have been taken in the literature (*e.g.*, Negrea *et al.*, 2020; Neyshabur *et al.*, 2019).

some data distributions or some hypotheses lead to poor generalization, while naturally occurring data and commonly used learning algorithms perform well. This motivates the information-theoretic approach of making statements that are specific to the data distribution and learning algorithm in question. Still, the framework of uniform generalization has proven immensely powerful for many domains, and has led to a definitive characterization of when learning is possible in this strict sense for binary classification: the VC dimension. Intuitively, the VC dimension is related to the complexity of a hypothesis class, and measures the size of the biggest data set for which the hypothesis class can induce arbitrary labellings of the features. We give an overview of the VC dimension in Section 1.3.1.

A step towards incorporating data-dependence in the bounds was taken by Bartlett and Mendelson (2001, 2002), Gine and Zinn (1984), Koltchinskii (2001), and Koltchinskii and Panchenko (2000) with the introduction of the Rademacher complexity of a hypothesis class. The Rademacher complexity similarly measures the ability of a hypothesis class to instantiate arbitrary labels, but can be computed empirically on the basis of a training set. Still, it has a uniform flavor in terms of the hypothesis class. We discuss the Rademacher complexity in Section 1.3.2.

Note that we only provide an exceedingly brief overview of uniform convergence-flavored generalization bounds and their history, in order to provide context for the upcoming sections. Since properly covering this vast subject is far beyond the scope of the present monograph, the reader is referred to, for instance, the excellent books by Mohri *et al.* (2018) and Shalev-Shwartz and Ben-David (2014) for further details.

1.3.1 VC Dimension

We will now focus on binary classification, where the sample space decomposes as $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Here, $\mathcal{X}$ is the *feature space*, while $\mathcal{Y} = \{0, 1\}$ is the *label space*. Each hypothesis $w \in \mathcal{W}$ is a map $w : \mathcal{X} \to \{0, 1\}$ that predicts a label for each feature. We will focus on the $0 - 1$ loss function, given by $\ell(w, z) = 1\{w(x) \neq y\}$. Thus, the hypothesis incurs a loss if and only if it predicts the wrong label. For this setting, the VC dimension of $\mathcal{W}$, denoted as d_{VC}, provides a fundamental characterization of uniform convergence (defined in Definition 1.1), and hence of PAC learnability: $\mathcal{W}$ satisfies the uniform convergence property if and only if d_{VC} is finite. In order to define the VC dimension, we need to introduce the growth function of a hypothesis class (Shalev-Shwartz and Ben-David, 2014, Def. 6.5).

Definition 1.2 (Growth function and VC dimension). The *growth function $g_{\mathcal{W}}(m)$* is defined as the maximum number of different ways in which a feature set of size m can be classified using functions from $\mathcal{W}$, that is,

$$\max_{(x_1,\dots,x_m)\in\mathcal{X}^m} \left|\{(w(x_1),\dots,w(x_m)) : w \in \mathcal{W}\}\right| . \tag{1.2}$$

Note that $g_{\mathcal{F}}(m) \leq 2^m$. The *VC dimension* of $\mathcal{W}$, denoted d_{VC}, is the largest integer such that this upper bound holds with equality. Specifically,

$$d_{\mathrm{VC}} = \max\{m \in \mathbb{N} : g_{\mathcal{F}}(m) = 2^m\}. \tag{1.3}$$

If no such integer exists, we say that $d_{\mathrm{VC}} = \infty$. If the VC dimension of a hypothesis class is finite, we will refer to it as a VC class.

Intuitively, VC dimension characterizes uniform convergence for the following reason: if the VC dimension is infinite, we can change the labels of a training set $\boldsymbol{Z}$ arbitrarily and still find a hypothesis that outputs these exact predictions, no matter the size n of the training set. Hence, we can find a hypothesis with a minimal or maximal training loss, independent of the underlying population loss. However, if the VC dimension is finite and $n \gg d_{\mathrm{VC}}$, we cannot adapt arbitrarily to every sample in the training set, but only to d_{VC} of them. Therefore, in some sense, the remaining $n - d_{\mathrm{VC}}$ samples provide a reasonable estimate of the population loss.

Re-producing the full proof is beyond our present scope, but essentially, one proceeds by bounding the generalization gap in terms of the growth function by formalizing the intuition above (see, *e.g.*, Shalev-Shwartz and Ben-David, 2014, Chapter 28). Then, the growth function is controlled using the Sauer-Shelah lemma (Shalev-Shwartz and Ben-

David, 2014, Lemma 6.10), which provides a bound on the growth function in terms of the VC dimension.[5]

Lemma 1.3 (Sauer-Shelah lemma). Let $g_{\mathcal{W}}(\cdot)$ denote the growth function of the function class $\mathcal{W}$. For any function class $\mathcal{W}$ with VC dimension d_{VC},

$$g_{\mathcal{W}}(m) \leq \sum_{i=0}^{d_{\mathrm{VC}}} \binom{m}{i} \leq \begin{cases} 2^{d_{\mathrm{VC}}+1}, & m < d_{\mathrm{VC}}+1, \\ \left(\dfrac{em}{d_{\mathrm{VC}}}\right)^{d_{\mathrm{VC}}}, & m \geq d_{\mathrm{VC}}+1. \end{cases} \tag{1.4}$$

With this, we can obtain the following (Shalev-Shwartz and Ben-David, 2014, Thm. 6.8).

Theorem 1.4 (Generalization from VC dimension). Consider a hypothesis class $\mathcal{W}$ with VC dimension d_{VC}. Then, $\mathcal{W}$ has the uniform convergence property (see Definition 1.1) with sample complexity m, which is upper and lower bounded as

$$C' \frac{d_{\mathrm{VC}} + \log \frac{1}{\delta}}{\epsilon^2} \leq m(\epsilon, \delta) \leq C \frac{d_{\mathrm{VC}} + \log \frac{1}{\delta}}{\epsilon^2} = m_+(\epsilon, \delta), \tag{1.5}$$

for some constants C, C'. In particular, this implies that for all $w \in \mathcal{W}$,

$$|L_{\boldsymbol{Z}}(w) - L_{P_Z}(w)| \leq \sqrt{C \frac{d_{\mathrm{VC}} + \log \frac{1}{\delta}}{n}}. \tag{1.6}$$

This implies that $\mathcal{W}$ is PAC learnable in the following sense: for every distribution P_Z, there exists a deterministic learning algorithm $P_{W|\boldsymbol{Z}}$ such that, for every $\epsilon, \delta \in (0,1)$, we have that with probability at least $1-\delta$ over $P_{\boldsymbol{Z}}$,

$$L_{P_Z}(W) \leq \inf_{w \in W} L_{P_Z}(w) + \epsilon \tag{1.7}$$

provided that $n \geq m_+(\epsilon, \delta)$.

Remarkably, the upper and lower bounds on the sample complexity $m(\varepsilon, \delta)$ differ only by a multiplicative constant, and specifically, the dependence on d_{VC} is identical. Thus, the PAC learnability of a

[5] As we will see in Section 7.3, this is also a key tool for analyzing information-theoretic generalization bounds for the special case of VC classes.

hypothesis class $\mathcal{W}$ is fully determined by its VC dimension d_{VC} in the sense that $\mathcal{W}$ admits a finite sample complexity *if and only if* d_{VC} is finite. As remarked before, PAC learnability is a very strong requirement, as it is equivalent to uniform convergence both with respect to the hypothesis class and the data distribution. Hence, less stringent notions of generalization are of interest, especially distribution- and algorithm-dependent ones.

Under the assumption of realizability, where $\inf_{w \in W} L_{P_Z}(w) = 0$, it is possible to derive a bound similar to (1.6), but with a decay of $1/n$. This is referred to as a *fast* rate, in contrast to the *slow* rate of $1/\sqrt{n}$. For more details on fast rates, the reader is referred to the seminal works of Vapnik and Chervonenkis (1974), Lee *et al.* (1998), Li (1999), and the more recent works of Van Erven *et al.* (2015) and Grünwald and Mehta (2020).

1.3.2 Rademacher Complexity

Another important metric in the theoretical study of generalization is the *Rademacher complexity* (Bartlett and Mendelson, 2001, 2002; Gine and Zinn, 1984; Koltchinskii, 2001; Koltchinskii and Panchenko, 2000). Notably, the Rademacher complexity of a hypothesis class $\mathcal{W}$ is defined with respect to a given data set (although an average version, where an expectation is taken over the data set, is commonly used). We now give the definition of Rademacher complexity (Shalev-Shwartz and Ben-David, 2014, Chap. 26).

Definition 1.5 (Rademacher complexity)**.** Let $\boldsymbol{Z} \in \mathcal{Z}^n$ be a vector of data samples and let $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}^+$ be a loss function. Let σ_i for $i \in [n]$ be independent Rademacher random variables, so that $P_{\sigma_i}[\sigma_i = -1] = P_{\sigma_i}[\sigma_i = +1] = 1/2$. Then, the Rademacher complexity of the function class $\mathcal{W}$ with respect to $\boldsymbol{Z}$ and $\ell(\cdot, \cdot)$ is given by

$$\mathrm{Rad}_{\boldsymbol{Z}}(\mathcal{W}) = \frac{1}{n} \mathbb{E}_{P_{\sigma_1 \cdots \sigma_n}} \left[\sup_{w \in \mathcal{W}} \sum_{i=1}^{n} \sigma_i \ell(w, Z_i) \right]. \tag{1.8}$$

To get some intuition for the Rademacher complexity, one can imagine splitting the data set $\boldsymbol{Z}$ into a training set and a test set uniformly at random. What the Rademacher complexity measures, in a worst-case sense over the hypothesis class, is how big the discrepancy between the loss on the training set and the loss on the test set will be on average. With this interpretation, it is easy to see how the Rademacher complexity is tied to generalization: it is almost a generalization measure by definition. In the following theorem, the connection is made more specific (Shalev-Shwartz and Ben-David, 2014, Thm. 26.5).

Theorem 1.6 (Generalization guarantee from Rademacher complexity)**.** Assume that, for all $z \in \mathcal{Z}$ and all $w \in \mathcal{W}$, we have that $\ell(w, z) \in [0, 1]$. With probability at least $1 - \delta$ over $P_{\boldsymbol{Z}}$, for all $w \in \mathcal{W}$,

$$L_{P_Z}(w) - L_{\boldsymbol{Z}}(w) \leq 2\mathrm{Rad}_{\boldsymbol{Z}}(\mathcal{W}) + \sqrt{\frac{2\log(2/\delta)}{n}}. \tag{1.9}$$

A similar bound holds when the sample-dependent Rademacher complexity is replaced by its expectation under $P_{\boldsymbol{Z}}$.

As discussed by Shalev-Shwartz and Ben-David (2014, Part IV), the Rademacher complexity can be used to derive generalization bounds for relevant hypothesis classes, such as support vector machines, and can also be used to provide tighter bounds for classes with finite VC dimension. One issue with the Rademacher complexity is that, while being data-dependent, it is still a worst-case measure over the hypothesis class. This may typically lead to generalization estimates for modern machine learning algorithms that are overly pessimistic.

1.4 Generalization Bounds from Algorithmic Stability

We conclude our overview of generalization bounds by discussing an example that takes the learning algorithm into account, namely bounds based on algorithmic stability (Devroye and Wagner, 1979; Rogers and Wagner, 1978). As for the section on uniform convergence, we will only provide a very short presentation to provide context for upcoming chapters, as an exhaustive discussion is beyond our scope.

The intuition behind generalization bounds based on algorithmic stability is roughly as follows: if the selected output hypothesis does not depend too strongly on the specific training data it is based on, it should generalize well to unseen samples. Making this intuition precise, and specifically formalizing the notion of "strong dependence," leads to several different notions of stability that can be related to generalization performance. In this section, we will focus only on uniform stability, as studied by, *e.g.*, Bousquet and Elisseeff (2002, Def. 6). There is, however, a whole host of alternatives that have been studied in the literature (see, *e.g.*, the works of Kutin and Niyogi, 2002 and Rakhlin *et al.*, 2005). As shown by Shalev-Shwartz *et al.* (2010), there is also a fundamental relation between stability and uniform convergence in settings beyond standard supervised classification and regression.

We now present a generalization bound for deterministic learning algorithms that satisfy uniform stability (Bousquet and Elisseeff, 2002, Def. 6).

Theorem 1.7 (Uniform stability and generalization). We denote $\boldsymbol{Z}^{\setminus i} = (Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_n)$, and let $W(\boldsymbol{Z}) \in \mathcal{W}$ denote the output of a deterministic learning algorithm given a training set $\boldsymbol{Z}$. Assume that the learning algorithm has uniform stability β in the sense that, for all $\boldsymbol{Z} \in \mathcal{Z}^n$ and all $i \in [n]$,

$$\max_{z' \in \mathcal{Z}} \left\{ \left| \ell(W(\boldsymbol{Z}), z') - \ell(W(\boldsymbol{Z}^{\setminus i}), z') \right| \right\} \leq \beta. \tag{1.10}$$

Then, with probability at least $1 - \delta$ under $P_{\boldsymbol{Z}}$,

$$L_{P_Z}(W(\boldsymbol{Z})) - L_{\boldsymbol{Z}}(W(\boldsymbol{Z})) \leq 2\beta + (4n\beta + 1)\sqrt{\frac{\log \frac{1}{\delta}}{2n}}. \tag{1.11}$$

For many stable algorithms, such as linear regression and classification with support vector machines, the stability parameter β decays with n, implying that the bound in Theorem 1.7 approaches zero as the number of training samples increases. For further details, including the relation to regularization, see, for instance, Shalev-Shwartz and Ben-David (2014, Chapter 13).

While we will not discuss them in detail, other approaches to generalization have been taken in the literature, for instance, based on margins (Shawe-Taylor and Cristianini, 1999) and norms (Neyshabur *et al.*, 2015).

1.5 Outline

This monograph is structured as follows. In Part I, comprising Chapters 2 to 6, we cover the foundations of information-theoretic and PAC-Bayesian generalization bounds for standard supervised learning. Specifically, in Chapter 2, we give an intuitive motivation for why information-theoretic tools are suited for the study of generalization, before presenting and proving a first information-theoretic generalization bound as a gentle introduction to the subsequent chapters. In Chapter 3, we overview the core tools that are used in deriving generalization bounds in the upcoming chapters, in the form of information measures, change of measure techniques, and concentration inequalities. We use these tools to derive generalization bounds in expectation in Chapter 4 and generalization bounds in probability in Chapter 5, including PAC-Bayesian generalization bounds. We conclude Part I by presenting the conditional mutual information (CMI) framework, as well as the generalization bounds that can be derived through it.

In Part II, comprising Chapters 7 to 10, we turn to applications of the generalization bounds from Part I, as well as extensions to settings beyond standard supervised learning. In Chapter 7, we examine the *information complexity* of several learning algorithms, that is, the value of information measures that the learning algorithms induce. In Chapter 8, we focus specifically on iterative methods, wherein the hypothesis is sequentially updated as training progresses. This includes neural networks trained through standard methods, such as variants of gradient descent. In Chapter 9, we derive bounds for alternative learning models, namely meta learning, out-of-distribution generalization, federated learning, and reinforcement learning. Finally, in Chapter 10, we provide concluding remarks and a broader discussion of information-theoretic and PAC-Bayesian generalization bounds as a whole.

Part I

Foundations

2

Information-Theoretic Approach to Generalization

In the previous chapter, we introduced the generalization problem and hinted at an information-theoretic approach to addressing it. In this chapter, we expand upon this connection. We begin by providing a short introduction to information theory, and the flavor of results it provides. While this is only a brief overview of a vast area of study, our goal is to provide a glimpse of the field, which can serve to motivate and contextualize the coming results. After this, we clarify why information theory is a suitable starting point for studying generalization, before finishing the chapter by presenting and proving our first information-theoretic generalization bound. This serves as a warmup for the following chapters, since it allows us to introduce the general tools and concepts with a concrete, simple example.

2.1 An Exceedingly Brief Introduction to Information Theory

Information theory, as originally developed by Claude E. Shannon (Shannon, 1948) in the late 1940s and early 1950s, provides a rigorous mathematical framework for representing, processing, storing, and transferring information. Many information-theoretic quantities turn out to characterize fundamental limits for this: the entropy characterizes the minimum compressed size at which information can be stored under a perfect-reconstruction requirement; the relative entropy measures the same under a distribution mismatch; and the mutual information characterizes the limit at which information can be reliably transferred over an unreliable medium. The definitions of these quantities are provided in Section 3.1.

In the last decades, the information-theoretic approach of seeking fundamental limits without imposing complexity constraints has found applications in many fields beyond data transmission and storage, including statistical estimation, sparse recovery, and adaptive data analysis. In this monograph, we will see how information-theoretic quantities arise naturally when seeking analytic characterizations of the generalization error of randomized algorithms in the supervised learning setting.

2.2 Why Information-Theoretic Generalization Bounds?

But why is generalization in machine learning related to information theory? Intuitively, generalization should occur when the learning algorithm captures the relevant aspects of the training data, but disregards irrelevant factors. In a sense, this can be seen as a variant of Occam's razor, which says that among learners that perform well on the training set, the one that provides the simplest explanation is to be preferred. One way of interpreting what simplicity means is to say that the learner that extracts the least amount of information from the training data is the simplest one. Information-theoretic generalization bounds make this intuition precise by characterizing the generalization error of (randomized) learning algorithms in terms of information-theoretic metrics. Crucially, unlike the bounds based on uniform convergence in Section 1.3, these information-theoretic bounds do not solely aim to measure the complexity of the hypothesis class under consideration. Instead, they also incorporate dependence on the specific learning algorithm and data distribution. In Chapters 4 to 6, we provide several such results, and discuss their features in terms of assumptions, tightness in various situations, derivations, and relations between them.

Beyond this intuitive appeal, the framework of information-theoretic generalization bounds has several other attractive features. First, it can be used to recover bounds which were originally derived using a wide range of other approaches. In this sense, information-theoretic bounds offer a certain unifying (albeit not all-encompassing) perspective. This is covered in more detail in Chapter 7. Moreover, information-theoretic and PAC-Bayesian bounds have been used to obtain some of the tightest numerical performance guarantees for neural networks to date, indicating a promising avenue for furthering our understanding of these models. New learning algorithms can also be devised on the basis of minimizing the generalization bounds, paving the way for *self-certified* learning—*i.e.*, learning algorithms that use the training data to both learn a hypothesis and provide performance guarantees. We expand on these points in Chapter 8. Finally, as we cover in Chapter 9, the information-theoretic framework is flexible enough to accommodate many settings of interest, beyond the standard learning setting introduced in Chapter 1.

As an introduction, we begin by proving a simple information-theoretic bound in Section 2.3. This enables us to provide concrete instantiations of the tools and concepts that are relevant for deriving and interpreting information-theoretic bounds, before exploring these tools in greater generality in Chapter 3. While the results that are available in the literature vary widely in their details, the general recipe for obtaining them typically includes two crucial steps. The first step is a *change of measure*, which we cover in Section 3.2. The second step is a *concentration inequality*, which we discuss in Section 3.3. Variations of these two steps yield the generalization bounds we will discuss in Chapter 4 and Chapter 5.

When studying generalization, the main object of interest is the error event, which occurs when the hypothesis incurs a large loss on new data samples—*i.e.*, the hypothesis does not generalize well. The probability distribution that governs this event is typically not amenable to direct analysis, because the hypothesis and training sample are jointly distributed. For this reason, it is convenient to *change measure* to an auxiliary distribution that is easier to analyze. The cost of replacing the original distribution with the auxiliary distribution is quantified by an *information measure*, which can be seen as gauging the discrepancy between the two probability distributions.

The auxiliary probability distribution is chosen so that, under this distribution, we can control the error event. This is done by utilizing *concentration of measure inequalities*, which, roughly speaking, characterize the degree to which a random variable tends to deviate from its mean. Thus, by changing measure to a more easy-to-handle auxiliary distribution and applying concentration of measure results, we can obtain generalization bounds expressed through information measures.

2.3 A First Information-Theoretic Generalization Bound

To start us off gently within the broad topic of information-theoretic generalization bounds, we begin by giving a specific instantiation of an average bound. Specifically, in this section, we will present a generalization bound based on the sub-Gaussianity of bounded random variables and the Donsker-Varadhan variational representation of the relative entropy (Csiszar, 1975; Donsker and Varadhan, 1975). Throughout, we will highlight the role played by the different proof ingredients, focusing on intuition and providing indications of how these ingredients can later be generalized.

2.3.1 The Bound

Recall that the notation used here, and throughout this monograph, is detailed in Section 1.1. Before stating our first information-theoretic bound, we need to define the *relative entropy* between two probability distributions, also known as the Kullback-Leibler (KL) divergence. We also need the definition of the *mutual information*, which is the relative entropy between the joint distribution of two random variables and the product of their marginals.

Definition 2.1 (Relative entropy and mutual information). Consider two probability distributions P and Q defined on a common measurable space such that P is absolutely continuous with respect to Q, denoted by $P \ll Q$. The relative entropy between P and Q is given by

$$D(P \,||\, Q) = \mathbb{E}_P\left[\log \frac{\mathrm{d}P}{\mathrm{d}Q}\right]. \tag{2.1}$$

Here, $\frac{\mathrm{d}P}{\mathrm{d}Q}$ denotes the Radon-Nikodym derivative of P with respect to Q. If P is not absolutely continuous with respect to Q, the Radon-Nikodym derivative is undefined and we let $D(P\,||\,Q) = \infty$. We will give a precise definition of the Radon-Nikodym derivative in Theorem 3.16, but for now, it is sufficient to think of it as a likelihood ratio. The relative entropy is non-negative, so that $D(P\,||\,Q) \geq 0$, with equality if and only if $P = Q$. While it is tempting to interpret the relative entropy as a distance between P and Q, it is not a metric: it is not symmetric and it does not satisfy the triangle inequality.

For two random variables X and Y with joint distribution P_{XY} and product of marginals $P_X P_Y$, the mutual information between X and Y is

$$I(X;Y) = D(P_{XY}\,||\,P_X P_Y). \tag{2.2}$$

Note that, if X and Y are independent, $P_{XY} = P_X P_Y$, and $I(X;Y) = 0$. Also note that, if X is a continuous random variable and $Y = f(X)$ is a deterministic function of X, we have $I(X;Y) = \infty$.

We are now ready to state our first information-theoretic generalization bound.

Theorem 2.2. Consider a learning setting where the loss function is bounded, and satisfies $\ell(w,z) \in [0,1]$ for all $(w,z) \in \mathcal{W} \times \mathcal{Z}$. Then,

$$|\mathbb{E}_{P_{W\boldsymbol{Z}}}[\mathrm{gen}(W,\boldsymbol{Z})]| = |\mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{P_Z}(W) - L_{\boldsymbol{Z}}(W)]| \leq \sqrt{\frac{I(W;\boldsymbol{Z})}{2n}}. \tag{2.3}$$

Before proceeding to the proof of this theorem, let us examine the components of (2.3). After removing the absolute value, the average population loss can be upper bounded by two terms: the training loss and a so-called complexity term. Thus, to have any hope of obtaining a small bound on the population loss, we need to achieve a small training loss. Now, consider the complexity term, *i.e.*, the right-hand side of (2.3), whose key component is the mutual information $I(W;\boldsymbol{Z})$ between the hypothesis and the training data. On one hand, if the learning algorithm is oblivious to the training data, so that $P_{W|\boldsymbol{Z}} = P_W$, the mutual information will vanish, and the population loss is guaranteed to equal the training loss (on average). This is not surprising, since the training loss in this case is an unbiased estimator of the population loss. On the other hand, if the hypothesis is a deterministic function of the training data and both W and $\boldsymbol{Z}$ are continuous random variables, the mutual information is unbounded, and Theorem 2.2 provides a vacuous guarantee, meaning that the upper bound is trivial.

Typically, the value of the mutual information depends on the size n of the training set $\boldsymbol{Z}$. For Theorem 2.2 to give bounds that improve with n, the rate of increase of the mutual information with n has to be sublinear. If this is the case, the complexity term in Theorem 2.2 goes to 0 as n approaches infinity, and we guarantee that the population loss of the hypothesis we learn is arbitrarily close to its training loss, given sufficient samples.

It is tempting to compare this result to the channel coding problem, mentioned in Section 2.1. There, a transmitter encodes a message as a codeword X, which after transmission over a noisy channel $P_{Y|X}$ gives rise to the output Y, which is observed by the receiver whose aim is to decode the original message. From a mathematical standpoint, we can identify the training data $\boldsymbol{Z}$ with the codewords X, the learning algorithm $P_{W|\boldsymbol{Z}}$ with the channel law $P_{Y|X}$, and the hypothesis W with the output Y. For channel coding, the communication capacity of a noisy channel is given by the mutual information between the input and output, maximized over the input distribution P_X. By maximizing over the analogue of P_X in (2.3), *i.e.*, the distribution of the training data $P_{\boldsymbol{Z}}$, we obtain a worst-case upper bound for the generalization error.[1]

However, despite these superficial similarities between the two settings, there are fundamental differences between them. For channel communication, the conditional distribution from input to output is considered fixed, and the aim is to find an input that maximizes the mutual information. For machine learning, the input distribution is considered fixed, and the goal is to select a conditional distribution from input to output that minimizes the population loss. More importantly, while the mutual information has a very specific operational meaning in channel coding—it characterizes the maximal rate of reliable communication—its role in Theorem 2.2 is much more spurious. Indeed, it appears as an upper bound simply as a consequence of the particular

[1] In order to match the setting of learning with independent and identically distributed data, we must restrict ourselves to product distributions in this maximization.

change of measure that we use. As we will see in the coming chapters, other changes of measure give rise to upper bounds in terms of other information measures.

2.3.2 Proof of the Bound

We now proceed with proving the information-theoretic generalization bound in Theorem 2.2. For this, we will need two results: the Donsker-Varadhan variational formula for the relative entropy and a concentration result for bounded random variables.[2] As previously mentioned, these are the two main ingredients needed for deriving most information-theoretic generalization bounds. We state the results here without proof and not in the fullest generality possible, and defer further details to Section 3.2 and Section 3.3 respectively.

Theorem 2.3 (Donsker-Varadhan variational formula)**.** Let P and Q be two probability distributions on a common measurable space $\mathcal{X}$ such that $P \ll Q$. Then, for every $f : \mathcal{X} \to \mathbb{R}$ such that $\mathbb{E}_Q\left[e^{f(X)}\right] < \infty$,

$$D(P \,||\, Q) \geq \mathbb{E}_P[f(X)] - \log \mathbb{E}_Q\left[e^{f(X)}\right]. \tag{2.4}$$

Theorem 2.4 (Concentration of bounded random variables)**.** Let X_i, for $i \in [n]$, be independent random variables distributed according to P_X with range $[0, 1]$ and $\mathbb{E}[X_i] = \mu$. Let $X = \sum_{i=1}^{n} X_i/n$ denote the average of the X_i. Then, for every $\lambda \in \mathbb{R}$,

$$\log \mathbb{E}\left[e^{\lambda(\mu - X)}\right] \leq \frac{\lambda^2}{8n}. \tag{2.5}$$

To get an idea of how these results can be used, consider a situation where we want to know how a random variable X behaves under the distribution P, but where it is hard to perform an analysis dealing with P directly. Then, if we have an auxiliary distribution Q that allows easier analysis—for instance, if Theorem 2.4 holds under Q—we can first use Theorem 2.3 with $f(X) = \lambda(\mu - X)$ to change distribution from P to Q, at the price of a relative entropy, and then apply Theorem 2.4

[2] Note that we use the term "concentration result" quite liberally to include bounds on the moment-generating function, as such bounds imply concentration inequalities in the more strict sense.

to bound the term $\log \mathbb{E}_Q\left[e^{\lambda(\mu - X)}\right]$. This is exactly what we will do to prove Theorem 2.2.

Proof of Theorem 2.2. We first apply Theorem 2.3 with $X = (W, \boldsymbol{Z})$, $f(W, \boldsymbol{Z}) = \lambda \mathrm{gen}(W, \boldsymbol{Z})$ for $\lambda \in \mathbb{R}$, $P = P_{W\boldsymbol{Z}}$, and $Q = P_W P_{\boldsymbol{Z}}$. This implies that

$$\mathbb{E}_{P_{W\boldsymbol{Z}}}[\lambda \mathrm{gen}(W, \boldsymbol{Z})] \leq \log \mathbb{E}_{P_W P_{\boldsymbol{Z}}}\left[e^{\lambda \mathrm{gen}(W, \boldsymbol{Z})}\right] + D(P_{W\boldsymbol{Z}} \,||\, P_W P_{\boldsymbol{Z}}). \tag{2.6}$$

Since $D(P_{W\boldsymbol{Z}} \,||\, P_W P_{\boldsymbol{Z}}) = I(W; Z)$, we now see how the mutual information arises from this change of measure. Next, note that, for any fixed $w \in \mathcal{W}$,

$$\mathrm{gen}(w, \boldsymbol{Z}) = L_{P_Z}(w) - \frac{1}{n}\sum_{i=1}^{n} \ell(w, Z_i). \tag{2.7}$$

Since the training losses $\ell(w, Z_i)$ are bounded to $[0, 1]$ and identically distributed with mean $L_{P_Z}(w)$, we can invoke Theorem 2.4 to conclude that, for every $w \in \mathcal{W}$ and $\lambda \in \mathbb{R}$,

$$\mathbb{E}_{P_{\boldsymbol{Z}}}\left[e^{\lambda \mathrm{gen}(w, \boldsymbol{Z})}\right] \leq \exp\left(\frac{\lambda^2}{8n}\right). \tag{2.8}$$

By averaging (2.8) over P_W, we obtain

$$\log \mathbb{E}_{P_W P_{\boldsymbol{Z}}}\left[e^{\lambda \mathrm{gen}(W, \boldsymbol{Z})}\right] \leq \frac{\lambda^2}{8n}. \tag{2.9}$$

We now see that, through the concentration inequality in Theorem 2.4, we are able to control the generalization gap under the distribution $P_W P_{\boldsymbol{Z}}$. By inserting (2.9) into (2.6), we obtain, for $\lambda > 0$,

$$\mathbb{E}_{P_{W\boldsymbol{Z}}}[\mathrm{gen}(W, \boldsymbol{Z})] \leq \frac{\lambda}{8n} + \frac{I(W; \boldsymbol{Z})}{\lambda}. \tag{2.10}$$

All that remains is to select the hitherto unspecified parameter λ, which we do by minimizing the right-hand side of (2.10). To obtain the absolute value, we perform the same procedure for $\lambda < 0$. After this, the result in (2.3) follows. □

The proof of Theorem 2.2 illustrates the key tools needed to establish information-theoretic generalization bounds. In this example, the change of measure was performed via the Donsker-Varadhan variational formula, the resulting information metric is the mutual information $I(W;\boldsymbol{Z})$, and the concentration of measure relied on the boundedness of the involved random variables. In the remainder of this monograph, we will present a more general framework for obtaining information-theoretic generalization bounds, through which alternative techniques can be used to obtain tighter bounds or bounds that hold under different assumptions than the ones in this section.

2.4 Bibliographic Remarks and Additional Perspectives

The specific bound that we present in Theorem 2.2, along with its proof, are based on the work of Xu and Raginsky (2017), which itself extended the results of Russo and Zou (2016) to a more general setting. Arguably, the core of the approach dates back to the work of Shawe-Taylor and Williamson (1997), who derived PAC bounds for Bayesian predictors in terms of a "luckiness" function (which is similar to a prior). This was extended to more general settings by McAllester (1998), leading to a bound that is very similar in form to the bounds discussed in this monograph. The proof technique, however, is quite different: it relies on the "quantifier reversal lemma" which, in some sense, plays the role of a change of measure. This PAC-Bayesian strand of the literature then flourished, with generalizations and tighter bounds by, to only give some examples, Langford and Seeger (2001), McAllester (2003a), Audibert (2004), and Catoni (2007), with proofs of a similar form as we discuss here. A more extensive overview of the PAC-Bayesian literature is given in Chapter 5, specifically in Sections 5.2 and 5.5. The bounds in the PAC-Bayesian literature focused primarily on bounded losses, and in particular, the $0-1$ loss. Around the same time, Zhang (2006) developed generalization bounds for generic loss functions based on a result termed the "information exponential inequality."

We now come to the work of Russo and Zou (2016), who focused on adaptive data analysis, rather than generalization bounds. Specifically, given the data, an analyst computes m different measurements $\phi = \{\phi_i\}_{i\in[m]}$. Then, based on the values of these measurements, they report ϕ_T for some $T \in [m]$. Since the choice of the measurement to report depends on the measurements themselves, this can introduce a significant bias. The main result of Russo and Zou (2016) is a bound on this bias in terms of the mutual information $I(T;\phi)$, under the assumption that the measurements are sub-Gaussian. This setting can be seen to be equivalent to a statistical learning setting, where the measurements correspond to losses and the index T corresponds to a hypothesis from a finite set. While these developments appear to be largely independent from the PAC-Bayesian literature, Russo and Zou (2016) noted the resemblance to PAC-Bayesian bounds, stating that it would be interesting to explore the connections between PAC-Bayes and adaptive data analysis. Xu and Raginsky (2017) made the connection between statistical learning and the results of Russo and Zou (2016) precise, and in particular, extended the argument to uncountable hypothesis classes. Prior to this, Raginsky *et al.* (2016) derived generalization bounds in terms of information-theoretic versions of algorithmic stability, where the bounds were given in terms of the mutual information between the hypothesis and a single training datum, given the rest of the samples, where the sub-Gaussianity assumption was slightly different.

Since we have so far only provided an initial introduction to information-theoretic generalization bounds, we will defer a more detailed discussion and comparison of these results to Chapters 4 and 5.

Another tool from information theory that has received significant attention in machine learning is the information bottleneck (Tishby *et al.*, 1999). While we will not discuss it much in the remainder of this monograph, we will conclude this chapter with a discussion of the application of the information bottleneck in statistical learning. Specifically, consider two random variables X and Y, where X is an input and Y is an output. Assume that we want to find a representation T, which is a compressed version of X, but which should be useful in predicting Y. The idea of the information bottleneck method is that we want to set the conditional distribution $P^*_{T|X}$ of T given X so that, for some parameter $\beta > 0$,

$$P^*_{T|X} = \sup_{P_{T|X}} \left\{ \beta I(T;Y) - I(X;T) \right\}. \tag{2.11}$$

Here, $I(T;Y)$ captures the *sufficiency* of T, in the sense that it is informative of Y, while $I(X;T)$ measures the *minimality* of T, in the sense that it only captures aspects of X that are necessary for predicting Y. The parameter β controls the trade-off between these two objectives. While originally motivated by compression in information theory, Shwartz-Ziv and Tishby (2017) argued that the information bottleneck can also be used to explain phenomena in statistical learning, and in particular neural networks. Specifically, let T denote the activations of an intermediate layer in a neural network. Through empirical studies, Shwartz-Ziv and Tishby (2017) argued that neural network training consisted of a *fitting* phase, where both $I(T;Y)$ and $I(X;T)$ increase and the network achieves good predictive performance, followed by a *compression* phase, where $I(T;Y)$ remains constant but $I(X;T)$ decreases, so that the network learns a compressed, well-generalizing representation. Achille and Soatto (2018) developed this further to derive a regularized training objective that aims to promote learning minimal representations, and connected this with PAC-Bayesian theory. The existence of the fitting and compression phases was questioned by Saxe *et al.* (2019), who argued that these empirical phenomena do not occur in general, and depend heavily on implementation details. More discussion on the information bottleneck and its connection to learning can be found in the works of Goldfeld and Polyanskiy (2020) and Geiger (2021), as well as Kawaguchi *et al.* (2023), who establish generalization bounds.

3

Tools

The proofs of the large majority of information-theoretic generalization bounds in the literature have two key steps in common: a *change of measure* and a *concentration of measure*. In the previous chapter, this was illustrated with a concrete example, leading to our first information-theoretic generalization bound. As we will see in Section 5.2, these same tools are also at the heart of the PAC-Bayesian approach, and these two strands can be unified through this lens.

In this chapter, we will introduce the tools that will be used to derive PAC-Bayesian and information-theoretic generalization bounds in the remainder of the monograph in more detail and generality. Specifically, we will define some common information measures in Section 3.1, discuss change of measure techniques in Section 3.2, and present concentration of measure in Section 3.3.

3.1 Information Measures

In Theorem 2.2, we found that the generalization error of a randomized learning algorithm can be controlled by the mutual information between the training data and the hypothesis. The mutual information is just one example of an *information measure*. Formally, given a measurable space $\mathcal{X}$ and the associated family $\mathcal{M}(\mathcal{X})$ of probability measures on $\mathcal{X}$, an (average) information measure is a mapping $\mathrm{IM} : \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \mathbb{R}$. Typically, for all $P \in \mathcal{M}(\mathcal{X})$, we have $\mathrm{IM}(P, P) = 0$. Thus, an information measure is some way to quantify the discrepancy between two probability measures. Often, these information measures are not metrics in the formal sense, as they may not satisfy symmetry or the triangle inequality. An example of this is the relative entropy from Definition 2.1, which maps the two distributions $P, Q \in \mathcal{M}(\mathcal{X})$ to $D(P \,||\, Q) = \mathbb{E}_P\left[\log \frac{\mathrm{d}P}{\mathrm{d}Q}\right]$. Note that, in general, $D(P \,||\, Q) \neq D(Q \,||\, P)$. In addition to such average information metrics, which only depend on the distributions, we will also consider pointwise versions, which are mappings from $\mathcal{M}(\mathcal{X})^2 \times \mathcal{X}^2$ to $\mathbb{R}$.

Throughout information theory and machine learning, such information measures are exceedingly useful and abundant. In the context of information-theoretic and PAC-Bayesian generalization bounds, they naturally appear in upper bounds on the population loss of learning algorithms, as exemplified by Theorem 2.2. In this section, we will introduce some information measures along with their properties, which will be useful in later chapters. For a more detailed review, the reader is referred to, for example, Cover and Thomas (2006) and Polyanskiy and Wu (2022), upon which much of the material in this section is based.

A basic building block of many information measures is some kind of likelihood ratio. For two probability mass functions P and Q on a common space $\mathcal{X}$, their likelihood ratio at a point $x \in \mathcal{X}$ is defined as $P(x)/Q(x)$. Similarly, if p and q are probability densities, the likelihood ratio is $p(x)/q(x)$. For generic measures P and Q, this concept is captured by the Radon-Nikodym derivative, denoted by $\mathrm{d}P/\,\mathrm{d}Q$. For the cases of discrete or continuous random variables, it reduces to the aforementioned likelihood ratios. The precise meaning of this object is captured by the Radon-Nikodym theorem, a change of measure that relates probabilities of events under P with their probabilities under Q. We will present this result in Theorem 3.16. The Radon-Nikodym derivative exists whenever P is *absolutely continuous* with respect to Q, as described in the following definition.

Definition 3.1 (Absolute continuity)**.** A measure P is absolutely continuous with respect to a measure Q, denoted as $P \ll Q$, if, for every measurable set $\mathcal{E}$ such that $Q(\mathcal{E}) = 0$, we also have $P(\mathcal{E}) = 0$.

For the special case where $P = P_{XY}$ and $Q = P_X P_Y$ are the joint distribution and product of marginal distributions of two random variables X and Y, we will refer to the logarithm of the Radon-Nikodym derivative as the *information density.*

Definition 3.2 (Information density)**.** The information density between two random variables X and Y with joint distribution P_{XY} and marginal distributions P_X and P_Y is given by

$$\imath(X, Y) = \log \frac{\mathrm{d}P_{XY}}{\mathrm{d}P_X P_Y}, \tag{3.1}$$

provided that $P_{XY} \ll P_X P_Y$. The conditional information density between X and Y given Z is

$$\imath(X, Y|Z) = \log \frac{\mathrm{d}P_{XYZ}}{\mathrm{d}P_{X|Z} P_{Y|Z} P_Z}, \tag{3.2}$$

provided that $P_{XYZ} \ll P_{X|Z} P_{Y|Z} P_Z$.

A fundamental information-theoretic quantity is the *entropy.*

Definition 3.3 (Entropy)**.** Let X be a discrete random variable on $\mathcal{X}$ with probability mass function P_X. The entropy of X is given by

$$H(X) = \sum_{x \in \mathcal{X}} P_X(x) \log \frac{1}{P_X(x)}. \tag{3.3}$$

Furthermore, let Y be a discrete random variable on $\mathcal{Y}$ with probability mass function P_Y, such that the joint distribution of X and Y is P_{XY}. Then, the conditional entropy of X given Y is

$$H(X|Y) = \sum_{x,y \in \mathcal{X} \times \mathcal{Y}} P_{XY}(x, y) \log \frac{P_Y(y)}{P_{XY}(x, y)}. \tag{3.4}$$

The entropy satisfies the following key properties:

1. *Non-negativity:* $H(X) \geq 0$, and equality holds if and only if $P_X(x) = 1$ for some $x \in \mathcal{X}$.

2. *Maximum:* $H(X) \leq \log(|\mathcal{X}|)$, and equality holds if and only if P_X is the uniform distribution on $\mathcal{X}$.
3. *Chain rule:* $H(X, Y) = H(Y) + H(X|Y) = H(X) + H(Y|X)$.
4. *Conditioning reduces entropy:* $H(X|Y) \leq H(X)$, and equality holds if and only if X and Y are independent.

There exists an extension to continuous random variables in the form of the differential entropy.

Definition 3.4 (Differential entropy)**.** Let X be a continuous random variable on $\mathcal{X}$ with probability density function p_X. The differential entropy of X is given by

$$h(X) = \int_{\mathcal{X}} p_X(x) \log \frac{1}{p_X(x)} \, \mathrm{d}x. \tag{3.5}$$

Furthermore, let Y be a continuous random variable on $\mathcal{Y}$ with probability density function p_Y, such that the joint density of X and Y is p_{XY}. Then, the conditional differential entropy of X given Y is

$$h(X|Y) = \int_{\mathcal{X}\times\mathcal{Y}} p_{XY}(x, y) \log \frac{p_Y(y)}{p_{XY}(x, y)} \, \mathrm{d}x \, \mathrm{d}y. \tag{3.6}$$

The differential entropy is shift invariant, so that for any $a \in \mathbb{R}$, $h(X) = h(X + a)$. However, the differential entropy does not satisfy many key properties of its discrete counterpart, such as non-negativity, and it is not scale-invariant, meaning that $h(aX) \neq h(X)$ in general.

Several key features of both the discrete and differential entropy can be described using the *relative entropy*, sometimes called the Kullback-Leibler (KL) divergence. This is a very commonly used information measure, which we already introduced in Section 2.3. We repeat its definition below.

Definition 3.5 (The relative entropy)**.** Consider two probability distributions P and Q defined on a common measurable space such that $P \ll Q$. The relative entropy between P and Q is given by

$$D(P \,||\, Q) = \mathbb{E}_P \left[\log \frac{\mathrm{d}P}{\mathrm{d}Q} \right]. \tag{3.7}$$

If P is not absolutely continuous with respect to Q, the Radon-Nikodym derivative is undefined and we set $D(P \,||\, Q) = \infty$.

Given a distribution P_X on $\mathcal{X}$ and two conditional distributions $P_{Y|X}$ and $Q_{Y|X}$ on Y given X, the conditional relative entropy between $P_{Y|X}$ and $Q_{Y|X}$ given P_X is defined as

$$D(P_{Y|X} \,||\, Q_{Y|X} \,|\, P_X) = \mathbb{E}_{P_X}\left[D(P_{Y|X} \,||\, Q_{Y|X})\right]. \tag{3.8}$$

The relative entropy satisfies a useful property called the *chain rule.*

Theorem 3.6 (The chain rule of relative entropy). Given the distributions $P_{XY} = P_X P_{Y|X}$ and $Q_{XY} = Q_X Q_{Y|X}$, we have

$$D(P_{XY} \,||\, Q_{XY}) = D(P_{Y|X} \,||\, Q_{Y|X} \,|\, P_X) + D(P_X \,||\, Q_X). \tag{3.9}$$

When P and Q are a joint distribution and product of marginals of two random variables, the relative entropy between P and Q is referred to as the *mutual information* between the random variables.

Definition 3.7 (Mutual information). The mutual information between two random variables X and Y with joint distribution P_{XY} and marginal distributions P_X and P_Y is given by

$$I(X;Y) = D(P_{XY} \,||\, P_X P_Y) = \mathbb{E}_{P_{XY}}[\imath(X,Y)]. \tag{3.10}$$

The conditional mutual information between two random variables X and Y given Z is given by

$$I(X;Y|Z) = D(P_{XY|Z} \,||\, P_{X|Z}P_{Y|Z} \,|\, P_Z) = \mathbb{E}_{P_{XYZ}}[\imath(X,Y|Z)]. \tag{3.11}$$

We now see the motivation behind the name information density—its average is the mutual information (with an analogous correspondence for the conditional information density). The mutual information is one of the most fundamental quantities in information theory, and famously characterizes the capacity of a noisy communication channel. Recently, as discussed in Section 2.3, the mutual information has garnered interest in the statistical learning community as a measure of generalization.

Since it is a relative entropy, the mutual information inherits a version of the chain rule for relative entropy, which follows directly from Theorem 3.6.

Theorem 3.8 (The chain rule of mutual information)**.** Consider three random variables X, Y, and Z. Then,

$$I(X,Y;Z) = I(X;Z) + I(Y;Z|X) = I(Y;Z) + I(X;Z|Y). \tag{3.12}$$

For discrete random variables, the mutual information can be expressed in terms of entropy as $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$. For continuous random variables, it can be expressed in terms of the differential entropy as $I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X)$.

The relative entropy is a special case of a wider class of information measures called f-divergences.

Definition 3.9 (f-divergence)**.** Let P and Q be two probability distributions on a common measurable space $\mathcal{X}$ such that $P \ll Q$. Let $f : (0,\infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$, extended so that $f(0) = \lim_{x\to 0^+} f(x)$. Then, the f-divergence between P and Q is defined as

$$D_f(P \,||\, Q) = \mathbb{E}_Q\left[f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\right]. \tag{3.13}$$

By setting $f(x) = x \log x$, we recover the relative entropy. Other notable examples include the total variation $\mathrm{TV}(P,Q) = \mathbb{E}_Q\left[\left|\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\right|\right]/2$, obtained by setting $f(x) = |x-1|/2$, and the χ^2-divergence $\chi^2(P \,||\, Q) = \mathbb{E}_Q\left[(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1)^2\right]$, obtained by setting $f(x) = (x-1)^2$.

For many pairs of f-divergences, one can establish comparison inequalities (Polyanskiy and Wu, 2022, Sec. 7.5). For our purposes, comparison inequalities involving the relative entropy and total variation will be of particular importance. We state two such inequalities below (which are discussed more by, *e.g.*, Canonne, 2022).

Theorem 3.10 (Pinsker's inequality and the Bretagnolle-Huber (BH) inequality)**.** Let P and Q be two probability distributions such that $P \ll Q$. Then, Pinsker's inequality states that (see, *e.g.*, Polyanskiy and Wu, 2022, Thm. 7.9)

$$\mathrm{TV}(P,Q) \le \sqrt{\frac{D(P \,||\, Q)}{2}}. \tag{3.14}$$

Furthermore, the BH inequality states that (Bretagnolle and Huber, 1978)

$$\mathrm{TV}(P, Q) \leq \sqrt{1 - \exp(-D(P \,||\, Q))}. \tag{3.15}$$

We now review some useful properties of f-divergences. For proofs, see Polyanskiy and Wu (2022, Thm. 7.4 and 7.5).

Theorem 3.11 (Properties of f-divergences.). For every f-divergence, the following properties hold:

1. *Non-negativity:* $D_f(P \,||\, Q) \geq 0$, and equality holds if and only if $P = Q$.

2. *Data-processing:* Let P_X and Q_X be two distributions on $\mathcal{X}$, and let P_Y and Q_Y be the corresponding distributions on $\mathcal{Y}$ induced by a kernel $P_{Y|X}$, that is, $P_Y(\mathcal{E}) = \int_{\mathcal{X}} \mathrm{d}P_X(x) P_{Y|X}(\mathcal{E}|x)$ and $Q_Y(\mathcal{E}) = \int_{\mathcal{X}} \mathrm{d}Q_X(x) P_{Y|X}(\mathcal{E}|x)$ for every measurable set $\mathcal{E} \subset \mathcal{Y}$. Then,
$$D_f(P_X \,||\, Q_X) \geq D_f(P_Y \,||\, Q_Y). \tag{3.16}$$

3. *Conditioning increases divergence:* Let P_X be a distribution on $\mathcal{X}$, and let P_Y and Q_Y be the distributions induced on $\mathcal{Y}$ by two kernels $P_{Y|X}$ and $Q_{Y|X}$ respectively, *i.e.* $P_Y(\mathcal{E}) = \int_{\mathcal{X}} \mathrm{d}P_X(x) P_{Y|X}(\mathcal{E}|x)$ and $Q_Y(\mathcal{E}) = \int_{\mathcal{X}} \mathrm{d}P_X(x) Q_{Y|X}(\mathcal{E}|x)$ for every measurable set $\mathcal{E} \subset \mathcal{Y}$. The *conditional* f-divergence is defined as
$$D_f(P_{Y|X} \,||\, Q_{Y|X} \,|\, P_X) = \mathbb{E}_{P_X}\Big[D_f(P_{Y|X} \,||\, Q_{Y|X})\Big] \tag{3.17}$$
and it satisfies the inequality
$$D_f(P_Y \,||\, Q_Y) \leq D_f(P_{Y|X} \,||\, Q_{Y|X} \,|\, P_X). \tag{3.18}$$

Notably, unlike the relative entropy, general f-divergences do *not* satisfy the chain rule (cf. Theorem 3.6).

Another special instance of f-divergences is the Rényi divergence, also known as the α-divergence (Van Erven and Harremoës, 2014).

Definition 3.12 (Rényi divergence)**.** Let $\alpha \in (0,1) \cup (1,\infty)$. The Rényi divergence of order α between P and Q is defined as

$$D_\alpha(P \,||\, Q) = \frac{1}{\alpha - 1} \log \mathbb{E}_Q \left[\left(\frac{\mathrm{d}P}{\mathrm{d}Q} \right)^{\alpha} \right]. \tag{3.19}$$

For $\alpha = 1$, motivated by continuity, the Rényi divergence of order 1 coincides with the relative entropy:

$$D_1(P \,||\, Q) = D(P \,||\, Q). \tag{3.20}$$

The conditional Rényi divergence of order α between $P_{Y|X}$ and $Q_{Y|X}$ given P_X is

$$D_\alpha(P_{Y|X} \,||\, Q_{Y|X} \,|\, P_X) = D_\alpha(P_{Y|X} P_X \,||\, Q_{Y|X} P_X). \tag{3.21}$$

When $P = P_{XY}$ and $Q = P_X P_Y$ are the joint distribution of two random variables and the product of their marginals respectively and $\alpha \to \infty$, the Rényi divergence reduces to the maximal leakage, defined below (Issa *et al.*, 2020).

Definition 3.13 (Maximal leakage)**.** The maximal leakage from X to Y is defined as

$$\mathcal{L}(X \to Y) = \log \mathbb{E}_{P_Y} \left[\operatorname*{ess\,sup}_{P_X} \frac{\mathrm{d}P_{XY}}{\mathrm{d}P_X P_Y} \right]. \tag{3.22}$$

Here, the essential supremum of a measurable function $f(\cdot)$ of a random variable X distributed as P_X is defined as

$$\operatorname*{ess\,sup}_{P_X} f(X) = \inf_{a \in \mathbb{R}} \Big[P_X(\{X : f(X) > a\}) = 0 \Big]. \tag{3.23}$$

The conditional maximal leakage from X to Y given Z is defined as

$$\mathcal{L}(X \to Y|Z) = \log \operatorname*{ess\,sup}_{P_Z} \mathbb{E}_{P_{X|Z}} \left[\operatorname*{ess\,sup}_{P_{Y|Z}} \frac{\mathrm{d}P_{XYZ}}{\mathrm{d}P_{X|Z} P_{Y|Z} P_Z} \right]. \tag{3.24}$$

While the maximal leakage is obtained as the infinite limit of the Rényi divergence, the same does not hold for the conditional maximal leakage. Instead, the conditional maximal leakage is the infinite limit of the conditional α-mutual information, defined below (Verdú, 2015).

Definition 3.14 (α-mutual information)**.** For $\alpha \in (0,1) \cup (1,\infty)$, the α-mutual information between X and Y is given by

$$I_\alpha(X;Y) = \frac{1}{\alpha - 1} \log \mathbb{E}^{\alpha}_{P_X} \left[\mathbb{E}^{1/\alpha}_{P_Y} \left[\exp\left(\frac{\mathrm{d}P_{XY}}{\mathrm{d}P_X P_Y} \right)^{\alpha} \right] \right]. \tag{3.25}$$

The conditional α-mutual information between X and Y given Z is

$$I_\alpha(X;Y|Z) = \frac{1}{\alpha - 1} \log \mathbb{E}_{P_Z} \left[\mathbb{E}^{\alpha}_{P_{X|Z}} \left[\mathbb{E}^{1/\alpha}_{P_{Y|Z}} \left[\left(\frac{\mathrm{d}P_{XYZ}}{\mathrm{d}P_{X|Z} P_{Y|Z} P_Z} \right)^{\alpha} \right] \right] \right]. \tag{3.26}$$

It should be noted that the definition of the conditional α-mutual information given here is not the only possible one, and other definitions have been proposed (Esposito *et al.*, 2021b; Tomamichel and Hayashi, 2018). Our main reason for focusing on this particular definition is its role in generalization bounds and its connection to the conditional maximal leakage.

When $\alpha > 1$, the function x^α is convex. Jensen's inequality then implies that, for $\alpha > 1$, the (conditional) α-mutual information is a lower bound to the corresponding (conditional) Rényi divergence. Thus, we have

$$I_\alpha(X;Y) \le D_\alpha(P_{Y|X} \,||\, P_Y \,|\, P_X) \tag{3.27}$$

$$I_\alpha(X;Y|Z) \le D_\alpha(P_{XY} \,||\, P_X P_Y). \tag{3.28}$$

For $\alpha < 1$, the inequalities are reversed, and the two information measures coincide with the (conditional) mutual information for $\alpha \to 1$.

All of the aforementioned information measures rely on the Radon-Nikodym derivative in one way or another. The Wasserstein distance (sometimes called the Kantorovich metric), introduced in the context of optimal transport, is an example of an information measure that does not (Villani, 2008). One appealing consequence of this is that, while most information measures require absolute continuity, the Wasserstein distance does not.

Definition 3.15 (Wasserstein distance)**.** Let $\mathcal{X}$ be a set and ρ be a metric such that $(\mathcal{X}, \rho)$ is a Polish metric space. Let P and Q be two probability distributions on $\mathcal{X}$. Then, for every $p \in [1, \infty]$, the p-Wasserstein distance is

$$\mathbb{W}_p(P, Q) = \left(\inf_{R \in \Pi(P,Q)} \mathbb{E}_{(x,x') \sim R}[\rho(x, x')^p] \right)^{1/p} \tag{3.29}$$

where $\Pi(P, Q)$ denotes the set of joint probability distributions on $\mathcal{X}^2$ with marginal distributions P and Q. We refer to the 1-Wasserstein distance simply as the Wasserstein distance for brevity.

The Wasserstein distance can be understood intuitively as follows. Imagine that the two distributions P and Q describe two different ways of distributing a unit of dirt over $\mathcal{X}$. Then, each coupling between P and Q can be seen as a scheme of moving the dirt to turn one distribution into the other. The Wasserstein distance measures the lowest possible cost (in terms of ρ) at which one distribution of dirt can be turned into the other. This interpretation motivates the alternative name of "Earth Mover's Distance" (Rubner *et al.*, 1998). Note that, due to Jensen's inequality, $\mathbb{W}_p(P, Q)$ is non-decreasing with p.

3.2 Change of Measure

When studying the generalization gap, as aforementioned, the quantity of interest is the error event under the joint distribution of the hypothesis and the data. However, this can be difficult to control directly. Instead, there may be other, auxiliary distributions that allow for direct control of the error event. For instance, when one considers the hypothesis and the data to be drawn independently from each other, there are many situations where the concentration inequalities that we will introduce in Section 3.3 readily apply. The technique of relating the probability of an event under one distribution to its corresponding value under another auxiliary distribution is referred to as *change of measure*. The penalty incurred by replacing the original distribution with the auxiliary one can be expressed through information measures, such as those introduced in the previous section.

In this section, we introduce several change-of-measure results. We begin by introducing the Radon-Nikodym theorem, which is the backbone of many change of measure techniques. After this, we present methods based on variational representations of divergences, starting with the celebrated Donsker-Varadhan variational representation of the relative entropy. This variational representation can be used to relate averages under different distributions, with a penalty given by the relative entropy between the distributions. We then show how the notion of Fenchel conjugates can be used to extend the core idea of the Donsker-Varadhan variational representation to the broad family of f-divergences. Finally, we explain how the framework of optimal transport gives rise to a change of measure, in which the Wasserstein distance appears as the information measure.

3.2.1 The Radon-Nikodym Theorem

For any change of measure technique to be sensible, we need some conditions on the measures (or functions) involved. As an example, consider a random variable X that follows a standard Gaussian distribution, and assume that we are interested in the expectation of a function $f(X)$. Now, assume we want to compute this expectation by drawing samples from a Bernoulli distribution. Of course, this is doomed to fail from the beginning for almost any f. While the true distribution is supported on the real line, our auxiliary Bernoulli distribution is limited to $\{0, 1\}$. Since we have no chance of drawing samples on parts of the space where the Gaussian distribution has a non-zero density, we can only get a good indication from our samples if f is trivial everywhere except $\{0, 1\}$. If we instead were to use another distribution supported on all real numbers as our auxiliary distribution—say, another Gaussian or the t-distribution—we could draw samples from our auxiliary distribution and compute the expectation of f on this basis. For this procedure to give an accurate result, we would need to scale the samples by the probability ratio between the true distribution and our auxiliary one. This is related to importance sampling in statistics, and gives some intuition about the information measures that appear in the results of this section. The intuition described here is formally captured by the concept of absolute continuity, given in Definition 3.1.

Throughout this section, this property will be crucial for virtually every result. The importance of the absolute continuity property is that it guarantees the existence of the Radon-Nikodym derivative that appears in the Radon-Nikodym theorem, sometimes simply referred to as "change of measure." Provided that an absolute continuity requirement holds, the change of measure exactly relates the measure of an event under two distributions (Rudin, 1987, Thm. 6.10(b)).

Theorem 3.16 (Radon-Nikodym theorem)**.** Let P and Q be probability distributions on a common space such that $P \ll Q$. Then, there exists a function f such that, for any measurable event $\mathcal{E}$,

$$P(\mathcal{E}) = \int_{\mathcal{E}} f \, \mathrm{d}Q. \tag{3.30}$$

The function f is referred to as the Radon-Nikodym derivative of P with respect to Q, and is denoted by $\mathrm{d}P/\,\mathrm{d}Q$.

For discrete random variables, $\mathrm{d}P/\,\mathrm{d}Q$ is simply the ratio between the probability mass functions of the two distributions. For continuous random variables, it is the ratio between the probability density functions.

Recall that when the distributions P and Q are chosen as the joint distribution P_{XY} and the product of marginals $P_X P_Y$, the logarithm of the Radon-Nikodym derivative is the information density:

$$\imath(X,Y) = \log \frac{\mathrm{d}P_{XY}}{\mathrm{d}P_X P_Y}. \tag{3.31}$$

This can be used for the following change of measure: assume that we have $f(x,y) = 0$ whenever $\imath(x,y) = -\infty$. Note that, if we assume that $P_X P_Y \ll P_{XY}$, we always have $\imath(x,y) > -\infty$ so that the condition is satisfied for any function f. Then, by Theorem 3.16, we have (Polyanskiy and Wu, 2022, Prop. 18.3)

$$\begin{aligned} \mathbb{E}_{P_X P_Y}[f(X,Y)] &= \mathbb{E}_{P_{XY}}\left[\left(\frac{\mathrm{d}P_{XY}}{\mathrm{d}P_X P_Y}\right)^{-1} f(X,Y)\right] && (3.32) \\ &= \mathbb{E}_{P_{XY}}\left[e^{-\imath(X,Y)} f(X,Y)\right]. && (3.33) \end{aligned}$$

Of course, the same type of result holds if we replace the product of marginals $P_X P_Y$ with an auxiliary distribution Q_{XY}, provided that a suitable absolute continuity assumption holds, and that the information density is replaced with the corresponding logarithm of the Radon-Nikodym derivative.

3.2.2 The Donsker-Varadhan Variational Representation of the Relative Entropy

The celebrated Donsker-Varadhan variational representation of the relative entropy has its origins in the work of Donsker and Varadhan (1975). It has a rich history and is a core tool in both information theory and machine learning. Some alternative names include the shift of measure lemma (McAllester, 2003a) and the compression lemma (Banerjee, 2006). We state this important result below, the proof of which is adapted from Alquier (2024, Lemma 2.2).

Theorem 3.17 (Donsker-Varadhan variational representation)**.** Let Q be a probability distribution on a measurable space $\mathcal{X}$, and let Π denote the set of probability measures such that, for all $P \in \Pi$, we have $P \ll Q$. For every measurable function $f : \mathcal{X} \to \mathbb{R}$ such that $\mathbb{E}_Q\left[e^{f(X)}\right] < \infty$, we have

$$\log \mathbb{E}_Q\left[e^{f(X)}\right] = \sup_{P \in \Pi} \left\{\mathbb{E}_P[f(X)] - D(P \,||\, Q)\right\}. \tag{3.34}$$

The supremum is attained by the *Gibbs distribution* G, defined as

$$\frac{\mathrm{d}G}{\mathrm{d}Q}(X) = \frac{e^{f(X)}}{\mathbb{E}_Q\left[e^{f(X)}\right]}. \tag{3.35}$$

Proof. By straight-forward calculation, we find that for every $P \in \Pi$,

$$\begin{aligned} D(P \,||\, G) &= \mathbb{E}_P\left[\log \frac{\mathrm{d}P}{\mathrm{d}Q}\right] + \mathbb{E}_P\left[\log \frac{\mathrm{d}Q}{\mathrm{d}G}\right] & (3.36) \\ &= D(P \,||\, Q) + \log \mathbb{E}_Q\left[e^{f(X)}\right] - \mathbb{E}_P[f(X)]. & (3.37) \end{aligned}$$

By Theorem 3.11, we have that $D(P \,||\, G) \geq 0$, with equality if and only if $P = G$. Thus, the result follows. □

In the above, we view the function f as fixed, and optimize over the distribution P. Since the final result holds for the supremum over P, this change of measure will later lead to bounds that hold uniformly over learning algorithms—a celebrated key feature of the PAC-Bayesian approach. However, we can also consider an alternative view, where the distribution P is fixed and we allow f to be any function in $\mathcal{F} = \{f :$

$\mathbb{E}_Q\left[e^{f(X)}\right] < \infty\}$. To see this, note that if we let $f = \log \mathrm{d}P/\mathrm{d}Q$, we automatically get $G = P$. Thus, by rearranging (3.37), we find that

$$D(P\,||\,Q) = \sup_{f\in\mathcal{F}} \left\{\mathbb{E}_P[f(X)] - \log \mathbb{E}_Q\left[e^{f(X)}\right]\right\}. \tag{3.38}$$

This provides a dual perspective on Donsker-Varadhan's variational formula. The version in (3.34), with the supremum taken over P, is sometimes referred to as "inverse" Donsker-Varadhan (Polyanskiy and Wu, 2022, Exercise III.6).

Theorem 3.17 relates the expectation of $f(X)$ under P to the moment-generating function of $f(X)$ under Q, via the relative entropy between the two distributions. To see how we can use the theorem in practice for changing measure, consider its weaker form, given in Theorem 2.3, where we consider fixed f and P without performing supremization. As illustrated in the derivation of our first information-theoretic bound in Theorem 2.2, we can use Theorem 3.17 to transition from the joint distribution of W and $\boldsymbol{Z}$ to an auxiliary distribution where they are independent, at the cost of a relative entropy term. This allowed us to obtain an explicit generalization bound by direct application of a concentration inequality.

3.2.3 Variational Representation of f-divergences

As it turns out, the Donsker-Varadhan variational formula can be seen as a special case of a more general family of variational representations, where the relative entropy is replaced by f-divergences (Definition 3.9). This characterization relies on the concept of convex conjugates, sometimes referred to as Fenchel conjugate or Legendre–Fenchel transform (Fenchel, 1949). The convex conjugate is defined as follows.

Definition 3.18 (Convex conjugate)**.** The convex conjugate f^* of a convex function f is defined as

$$f^*(y) = \sup_{x\in\mathbb{R}} \{xy - f(x)\}. \tag{3.39}$$

A useful property of the convex conjugate is *biconjugation*. This means that, if f is convex and lower semi-continuous, $(f^*)^* = f$. For more on convex duality, see, for example, Rockafellar (1970, Part III).

Recall that the f-divergence between two distributions P and Q is given by $D_f(P\,||\,Q) = \mathbb{E}_Q\left[f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\right]$, and that the family of f-divergences includes a number of familiar quantities. We are now ready to state the variational representation of f-divergences (Nguyen *et al.*, 2010).

Theorem 3.19 (Variational representation of f-divergences)**.** Let P and Q be two probability distributions on a common measurable space $\mathcal{X}$ such that $P \ll Q$. Let $f : [0, \infty) \to \mathbb{R}$ be a convex and lower semi-continuous function with $f(1) = 0$. Then,

$$D_f(P\,||\,Q) = \sup_{\phi\in\Phi} \left\{\mathbb{E}_P[\phi] - \mathbb{E}_Q[f^*(\phi)]\right\}. \tag{3.40}$$

Here Φ denotes the set of all functions $\mathcal{X} \to \mathbb{R}$ such that the expectations in (3.40) are defined.

While setting $f(x) = x\log x$ allows us to recover the relative entropy as a special case of the f-divergence, the variational representation given in Theorem 3.19 does not exactly recover the functional form of the Donsker-Varadhan variational representation of the relative entropy in (3.38). Instead, we get

$$D(P\,||\,Q) = \sup_{\phi\in\Phi} \left\{\mathbb{E}_P[\phi] - \frac{1}{e}\,\mathbb{E}_Q\left[e^{\phi}\right]\right\}. \tag{3.41}$$

If we consider a fixed ϕ and use the resulting inequality to bound $\mathbb{E}_P[\phi]$, we obtain a bound that is strictly weaker than what we get from Theorem 3.17. It is possible to derive stronger variational representations of f-divergences which do reduce to the Donsker-Varadhan variational representation of the relative entropy, as is done by Ruderman *et al.* (2012) and Polyanskiy and Wu (2022, Thm. 7.25). However, the resulting identities are more involved, and not amenable to analysis for general f-divergences. For instance, when applied to general α-divergences, Polyanskiy and Wu (2022, Thm. 7.25) does not yield a closed-form solution. However, for some cases, the stronger, *constrained* variational representation due to Ruderman *et al.* (2012) can be used, as we shall see in Section 5.2.2. We state this result below.

Theorem 3.20 (Constrained variational representation of f-divergences)**.** Let P, Q and f be defined as in Definition 3.9. Then,

$$D_f(P\,||\,Q) \geq \sup_{\phi\in\Phi}\left\{\mathbb{E}_P[\phi] - \sup_{p\in\Delta(Q)}\left\{\mathbb{E}_Q[\phi p] + \mathbb{E}_Q[f(p)]\right\}\right\}. \tag{3.42}$$

Here Φ denotes the set of all functions $\mathcal{X}\to\mathbb{R}$ such that the expectations in (3.42) are defined, while $\Delta(Q)$ denotes the set of probability densities with respect to Q.

3.2.4 Optimal Transport

Recall the Wasserstein distance from Definition 3.15. The Wasserstein distance can be used to change measure by using the following result (Villani, 2008, Remark 6.5).

Theorem 3.21 (Kantorovich-Rubinstein duality)**.** Assume that the first moments of P and Q are finite. Then,

$$\mathbb{W}_1(P,Q) = \sup_{f\in 1\text{–Lip}(\rho)} \mathbb{E}_P[f] - \mathbb{E}_Q[f]\,, \tag{3.43}$$

where $1\text{–Lip}(\rho)$ denotes the set of functions $f:\mathcal{X}\to\mathbb{R}$ that are 1-Lipschitz under the metric ρ used to define the Wasserstein distance, that is, $|f(x)-f(y)|\leq\rho(x,y)$ for all $x,y\in\mathcal{X}$.

The usefulness of Theorem 3.21 for changing measure is apparent: if we fix any $f\in 1\text{–Lip}(\rho)$, (3.43) implies that

$$\mathbb{E}_P[f] \leq \mathbb{E}_Q[f] + \mathbb{W}_1(P,Q). \tag{3.44}$$

Thus, knowing the expectation of f under Q immediately yields a bound on its expectation under P, provided that we can characterize the Wasserstein distance $\mathbb{W}_1(P,Q)$. Theorem 3.21 allows us to replace the assumption of absolute continuity with a Lipschitz assumption on the function f.

Note that, by Jensen's inequality, the p-Wasserstein distance is an increasing function in p. Thus, the upper bound above still holds if we use $\mathbb{W}_1(P,Q)\leq\mathbb{W}_p(P,Q)$ for any $p\geq 1$. However, as the 1-Wasserstein distance leads to the tightest bound, this would only be useful if $\mathbb{W}_p(P,Q)$ is easier to control than $\mathbb{W}_1(P,Q)$ when $p>1$.

3.2.5 Hölder's Inequality

Finally, we present Hölder's inequality, which relates the expectation of a product to the product of the separate moments.

Theorem 3.22 (Hölder's inequality)**.** Let $p, q \in [1, \infty)$ be constants such that $1/p + 1/q = 1$. For two random variables X and Y, we have

$$\mathbb{E}[|XY|] \leq \mathbb{E}^{1/p}[|X|^p]\, \mathbb{E}^{1/q}[|Y|^q]\,. \tag{3.45}$$

Here, we use the shorthand $\mathbb{E}^a[X] = (\mathbb{E}[X])^a$.

The utility of Hölder's inequality for the purpose of changing measure is that it relates an expectation under a joint distribution to the product of expectations under the corresponding marginal distributions.

3.3 Concentration of Measure

We now turn to *concentration of measure* techniques, which allow us to control the deviation of a random variable from its mean. Specifically, let X be a random variable with mean μ, and let $S = \frac{1}{n}\sum_{i=1}^{n} X_i$ denote the average of n independent samples distributed as X.[1] Then, a concentration bound controls the probability that S deviates from μ by a certain amount. We will use the term "concentration result" liberally to include bounds on the moment-generating function of X, since they imply a concentration bound in the sense mentioned above (as we will discuss in Theorem 3.25). While the change of measure techniques discussed in Section 3.2 are useful for replacing expectations under a hard-to-handle probability distribution with the corresponding one under an easier auxiliary distribution, concentration of measure results are needed to control the expectation under the auxiliary distribution.

For a more detailed review of this vast topic, we refer the reader to, for example, the works of Boucheron *et al.* (2013), Massart (2007),

[1] While concentration inequalities can be derived for dependent random variables (see, for instance, the work of Marton, 1996, Samson, 2000, Kontorovich and Ramanan, 2008, and Kontorovich and Raginsky, 2017 as well as recent results using information measures from Esposito and Mondelli, 2023), we focus here on independent random variables. In Section 3.3.4, we consider dependent random variables in the form of martingales.

Raginsky and Sason (2013), and Wainwright (2019), where the proofs of the results presented here can be found.

3.3.1 Sub-Gaussian Random Variables

A commonly studied category of random variables is the one of sub-Gaussian random variables. A random variable is said to be sub-Gaussian with parameter σ, or σ-sub-Gaussian, if its moment-generating function is dominated by that of a Gaussian random variable with variance σ^2. This ensures that the random variable inherits many of the desirable properties of Gaussian random variables, and in particular, concentration results. Below, we state the definition of a sub-Gaussian random variable (Wainwright, 2019, Def. 2.2).

Definition 3.23 (Sub-Gaussian random variable)**.** A random variable X is called σ-sub-Gaussian if, for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq e^{\frac{\lambda^2\sigma^2}{2}}. \tag{3.46}$$

A useful property of sub-Gaussian random variables is that the sub-Gaussianity parameter σ behaves like a standard deviation under averaging: if we let S denote the average of n independent σ-sub-Gaussian random variables, then S is $\sigma/\sqrt{n}$-sub-Gaussian. We formalize this property below.

Proposition 3.24 (Averaging sub-Gaussian random variables)**.** Let X be a σ-sub-Gaussian random variable and let $S = \frac{1}{n}\sum_{i=1}^n X_i$ be the average of n independent instances of X. Then, S is $\sigma/\sqrt{n}$-sub-Gaussian.

As indicated earlier in this chapter, a bound on the moment-generating function implies a concentration inequality. This can be shown through the Chernoff method (Wainwright, 2019, Example 2.1). Specifically, for the average of sub-Gaussian random variables, we obtain the following (Wainwright, 2019, Prop. 2.5).

Theorem 3.25 (Sub-Gaussian concentration)**.** Let X be a σ-sub-Gaussian random variable and let $S = \frac{1}{n}\sum_{i=1}^n X_i$ be the average of n independent instances of X. Then

$$P(S - \mathbb{E}[X] > t) \leq e^{-\frac{nt^2}{2\sigma^2}}. \tag{3.47}$$

Thus, the average of independent samples of a sub-Gaussian random variable concentrates around its mean exponentially fast. In later chapters, when deriving generalization bounds, we will typically set S to be the training loss $L_{\boldsymbol{Z}}(W)$ and μ to be the population loss $L_{P_{\boldsymbol{Z}}}(W)$ (see Section 1.1 for definitions), so that $L_{P_{\boldsymbol{Z}}}(W)$ can be controlled in terms of $L_{\boldsymbol{Z}}(W)$ and the information measure that arises from the change of measure. As we will see in Chapters 4 and 5, it is often sufficient in the derivations of generalization bounds to have a bound on the moment-generating function, and we do not need to convert it into a concentration inequality as the one above. Hence, we will focus on bounds on the moment-generating function.

Sub-Gaussian random variables can also be characterized in terms of a bound on the moment-generating function of their square, as we formalized below (Wainwright, 2019, Thm. 2.6).

Proposition 3.26 (Squared sub-Gaussian random variables)**.** Let X be a σ-sub-Gaussian random variable and let $S = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the average of n independent instances of X. Then, for all $\lambda \in [0, 1)$,

$$\mathbb{E}\left[e^{\frac{n\lambda(S-\mathbb{E}[X])^2}{2\sigma^2}}\right] \leq \frac{1}{\sqrt{1-\lambda}}. \tag{3.48}$$

While we will not cover it explicitly, we note that sub-Gaussianity can be relaxed to sub-exponentiality and the related Bernstein condition (Wainwright, 2019, Sec. 2.1.3).

3.3.2 Bounded Random Variables

We now turn to the special case of bounded random variables. Throughout this section, we will, without loss of generality, assume that the range of the random variable is $[0, 1]$—results for generic bounded intervals can be obtained by shifting and scaling as appropriate.

As stated in the following proposition, bounded random variables are sub-Gaussian.

Proposition 3.27 (Bounded random variables are sub-Gaussian)**.** Let X be a random variable whose range is restricted to $[0, 1]$. Then, X is $1/2$-sub-Gaussian.

By directly exploiting the boundedness of a random variable, tighter characterizations of its concentration can be obtained. In the following, we will use the relative entropy between two Bernoulli random variables to obtain a concentration inequality that leads to significantly tighter bounds on the average of X when the observed sample mean is small.

Definition 3.28 (Binary relative entropy). Let $p, q \in [0,1]$. Then $d(q \,||\, p)$ denotes the relative entropy between two Bernoulli random variables with parameters q and p respectively, *i.e.*,

$$d(q \,||\, p) = D(\mathrm{Bern}(q) \,||\, \mathrm{Bern}(p)) \tag{3.49}$$

$$= q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p}. \tag{3.50}$$

Let $\gamma \in \mathbb{R}$. A "relaxed" parametric version of the binary relative entropy can be expressed as

$$d_\gamma(q \,||\, p) = \gamma q - \log(1 - p + pe^\gamma). \tag{3.51}$$

Specifically, one can show that $d(q \,||\, p) = \sup_\gamma d_\gamma(q \,||\, p)$.

The binary relative entropy between a sample mean and its expectation can be shown to display a useful concentration behavior. The following result is due to Maurer (2004).

Theorem 3.29 (Concentration for binary relative entropy). Let X be a random variable with range $[0,1]$ and mean μ. Let $S = \frac{1}{n}\sum_{i=1}^n X_i$ be the average of n independent instances of X. Then,

$$\mathbb{E}\left[e^{nd(S \,||\, \mu)}\right] \le 2\sqrt{n}. \tag{3.52}$$

By using this result, we can obtain upper bounds on the binary relative entropy between the sample average S and the mean μ. Specifically, since S is known, (3.52) leads to a bound on μ, which can be obtained by numerically evaluating the function

$$d^{-1}(S, c) = \sup\{\mu \in [0,1] : d(S \,||\, \mu) \le c\}. \tag{3.53}$$

While (3.53) does not admit an analytical solution, it can be relaxed to obtain the following, more easily interpretable, expression (McAllester, 2003b; Tolstikhin and Seldin, 2013).

Proposition 3.30 (Relaxed inverse of the binary relative entropy)**.** For all $S, c \in [0,1]$, we have

$$d^{-1}(S, c) \leq S + \sqrt{2Sc} + 2c. \tag{3.54}$$

In Chapter 4, we will see that this result is useful to derive accurate generalization bounds for small training losses.

An alternative concentration result can be derived by considering the relaxed binary relative entropy in (3.51). This turns out to be particularly useful in the derivation of average generalization bounds in Chapter 4. The following result is due to McAllester (2013).

Theorem 3.31 (Concentration for parametric binary relative entropy)**.** Let X be a random variable with range $[0,1]$ and mean μ. Let $S = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the average of n independent instances of X. Then, for every $\gamma \in \mathbb{R}$,

$$\mathbb{E}\left[e^{nd_\gamma(S\,||\,\mu)}\right] \leq 1. \tag{3.55}$$

While the upper bound in Theorem 3.29 scales as $\sqrt{n}$, this is constant in Theorem 3.31. This concentration result can be applied for a set of values for γ, for free in the case of bounds in expectation and at the cost of a union bound for bounds in probability. We will discuss this further in Chapter 4 and Chapter 5.

3.3.3 Binary Random Variables

While we previously considered bounded random variables within $[0,1]$, we now restrict our attention to binary random variables within this range. For such random variables, a concentration result on the weighted difference between the random variable and its complement can be derived, which will turn out useful in Chapter 6. The following is due to Steinke and Zakynthinou (2020).

Theorem 3.32 (Concentration of complementary random variables)**.** Let X be a random variable satisfying $P(X = a) = P(X = b) = 1/2$ where $a, b \in [0,1]$. Let $\bar{X} = a + b - X$ denote its complement in the set $\{a, b\}$. Finally, let $\lambda, \gamma > 0$ be constants satisfying $\lambda(1-\gamma) + (e^{\lambda} - 1 - \lambda)(1 + \gamma^2) \leq 0$. Then,

$$\mathbb{E}\left[e^{\lambda\left(X - \gamma\bar{X}\right)}\right] \leq 1. \tag{3.56}$$

3.3.4 Martingales

For all of the concentration results we have discussed so far, we have focused exclusively on *independent* samples. While this assumption often greatly simplifies calculations, it is often not satisfied in practice. One way to allow dependence between the samples, which still enables us to recover essentially the same type of concentration as with sub-Gaussianity, is the martingale property.

Definition 3.33 (Martingale sequences)**.** A sequence of random variables X_i, with $i = 1, \ldots, n$, is a submartingale if

$$\mathbb{E}[X_{n+1} \mid X_1, \ldots, X_n] \geq X_n. \tag{3.57}$$

The sequence is a supermartingale if

$$\mathbb{E}[X_{n+1} \mid X_1, \ldots, X_n] \leq X_n. \tag{3.58}$$

A sequence that is both a submartingale and a supermartingale is called a martingale.

A prototypical example of a martingale is a simple one-dimensional random walk, where $X_i = X_{i-1} + B_{i-1}$, where B_{i-1} is independent and uniformly distributed on $\{-1, +1\}$. By introducing a bias to the walk, it becomes a sub- or super-martingale.

The martingale property allows us to extend essentially sub-Gaussian concentration results to a much broader class of random variables, as shown in the following (Wainwright, 2019, Cor. 2.20).

Theorem 3.34 (Azuma-Hoeffding inequality)**.** Let $\{X_t\}_{t=1}^n$ be a sequence of random variables such that $|X_t - X_{t-1}| \leq c_t$ almost surely for all $t \in [n]$ and some constants $\{c_t\}_{t=1}^n$. Consider the following bound on the moment-generating function for $\lambda \in \mathbb{R}$:

$$\mathbb{E}\left[e^{\lambda(X_n - X_0)}\right] \leq \exp\left(\frac{-\lambda^2 \sum_{t=1}^n c_t^2}{2}\right). \tag{3.59}$$

If $\{X_t\}_{t=1}^n$ is a supermartingale, (3.59) holds for every $\lambda \geq 0$. If $\{X_t\}_{t=1}^n$ is a submartingale, (3.59) holds for every $\lambda \leq 0$. Finally, if $\{X_t\}_{t=1}^n$ is a martingale, (3.59) holds for every $\lambda \in \mathbb{R}$.

Thus, the sum of a bounded martingale sequence satisfies the same kind of sub-Gaussian bound on the moment-generating function (and hence, concentration inequality) as if the sequence were independent. Notably, Theorem 3.29 can similarly be extended to bounded martingale sequences, as shown by Seldin *et al.* (2012b, Lemma 2).[2]

3.3.5 Heavy-Tailed Random Variables

The concentration inequalities that we have discussed so far in this section have all relied on bounds on the moment-generating function. While this does cover many classes of random variables—and in particular encompasses the bounded random variables that appear in classification—there are many scenarios where such bounds are unrealistic. Moreover, even in settings where the moment-generating function is bounded by some parametric function, actually confirming this can be untenable, especially if we want to specify the parameters of the bound (such as σ for the sub-Gaussian random variables in Definition 3.23). Hence, it is of interest to obtain similar results for *heavy-tailed* random variables. While there is no definite consensus regarding the exact definition of this term, it typically refers to random variables for which the moment-generating function does not exist (away from 0). While this precludes the use of the techniques that we have covered so far in this section, it can still be possible to obtain generalization bounds in terms of, for instance, the variance of the involved random variables. We will see this in more detail in, for instance, Section 5.2.2. Finally, an approach to generalization bounds that avoid concentration arguments was taken by Mendelson (2014) and subsequent works by Lecué and Mendelson (2017, 2018) and Mendelson (2018).

[2] While the bound that is explicitly stated in Seldin *et al.* (2012b, Lemma 2) is weaker than Theorem 3.29, this is only for simplicity, as discussed in Seldin *et al.* (2012b, Lemma 13).

4

Generalization Bounds in Expectation

Equipped with the change of measure techniques from Section 3.2 and the concentration inequalities from Section 3.3, we are now ready to derive bounds on the generalization error of learning algorithms. In Section 1.2, we reviewed bounds of different flavors for randomized learning algorithms. Specifically, this included average, PAC-Bayesian, and single-draw bounds. As it turns out, there exists a unified approach to derive bounds of all these flavors simultaneously, sometimes referred to as the exponential stochastic inequality (ESI, see Grünwald and Mehta, 2020, Mhammedi *et al.*, 2019 and Grünwald *et al.*, 2023), or simply as the exponential inequality approach. We will briefly discuss this framework in Section 5.1. However, the details of the derivations that allow us to obtain the tightest possible bounds with the least restrictive assumptions differ somewhat depending on the type of bound under consideration. Hence, we will provide separate treatments for each type of bound. In this chapter, we focus on average generalization bounds—that is, generalization bounds in expectation. In Chapter 5, we discuss generalization bounds in probability—that is, PAC-Bayesian and single-draw bounds.

In order to keep the notation compact, we will use the following shorthands: we denote the average population loss as $L = \mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{P_Z}(W)]$, the average training loss as $\hat{L} = \mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{\boldsymbol{Z}}(W)]$, and their difference—that is, the average generalization error—as $\overline{\text{gen}} = L - \hat{L}$.

4.1 Bounds via Variational Representations of Divergences

As previously mentioned, most information-theoretic generalization bounds are based on a *change of measure* and a *concentration of measure* step. In this section, we will first present a generic result where the change of measure is performed using the Donsker-Varadhan variational representation of the relative entropy, stated in Theorem 3.17. Under different assumptions on the loss function, this can then be instantiated to obtain particular generalization bounds by applying a suitable concentration of measure step. This will allow us to recover the first information-theoretic generalization bound that we derived in Section 2.3, and generalize and improve it in several ways. Proposition 4.1 below is simply a restatement of the Donsker-Varadhan variational representation for the setup of interest in this chapter. A similar result, for convex functions with bounded inputs, was provided by Goyal *et al.* (2017).

Proposition 4.1. Assume that $P_{W\boldsymbol{Z}} \ll Q_W P_{\boldsymbol{Z}}$. Let $f : \mathcal{W} \times \mathcal{Z}^n \to \mathbb{R}$ be a function satisfying $\mathbb{E}_{P_{W\boldsymbol{Z}}}[f(W, \boldsymbol{Z})] < \infty$. Then, the Donsker-Varadhan variational formula for the relative entropy (Theorem 3.17) implies that

$$\mathbb{E}_{P_{W\boldsymbol{Z}}}[f(W, \boldsymbol{Z})] \leq \log \mathbb{E}_{Q_W P_{\boldsymbol{Z}}}\left[e^{f(W,\boldsymbol{Z})}\right] + D(P_{W\boldsymbol{Z}} \,||\, Q_W P_{\boldsymbol{Z}}). \tag{4.1}$$

In particular, when $Q_W P_{\boldsymbol{Z}} = P_W P_{\boldsymbol{Z}}$, we get

$$\mathbb{E}_{P_{W\boldsymbol{Z}}}[f(W, \boldsymbol{Z})] \leq \log \mathbb{E}_{P_W P_{\boldsymbol{Z}}}\left[e^{f(W,\boldsymbol{Z})}\right] + I(W; \boldsymbol{Z}). \tag{4.2}$$

The second term in the right-hand side of (4.1) is minimized by the choice $Q_W P_{\boldsymbol{Z}} = P_W P_{\boldsymbol{Z}}$, as a consequence of the golden formula (Csiszar and Körner, 2011, Eq. (8.7)). However, the resulting relative entropy may not always be possible to compute, while alternative choices of $Q_W P_{\boldsymbol{Z}}$ enable this. We will discuss this in more detail when turning to PAC-Bayesian bounds, where this is an important aspect. Throughout this chapter, we will assume that $Q_W P_{\boldsymbol{Z}} = P_W P_{\boldsymbol{Z}}$, as this allows us to express the information measures as simpler, familiar quantities, but we note that most results hold for arbitrary Q_W.

By suitably choosing the function f, the generic result in (4.2) can be instantiated to obtain different generalization bounds. First, we present the average bound from Xu and Raginsky (2017).

Corollary 4.2. Assume that the loss function $\ell(W, Z)$ is σ-sub-Gaussian under $P_W P_{\boldsymbol{Z}}$ and that $P_{W\boldsymbol{Z}} \ll P_W P_{\boldsymbol{Z}}$. Then,

$$\overline{\text{gen}} \leq \sqrt{\frac{2\sigma^2 I(W; \boldsymbol{Z})}{n}}. \tag{4.3}$$

Proof. We begin by applying (4.2) with

$$f(W, \boldsymbol{Z}) = \lambda \left(L_{\boldsymbol{Z}}(W) - \mathbb{E}_{P_W}[L_{P_Z}(W)] \right). \tag{4.4}$$

Then, by the sub-Gaussianity assumption (3.46), we have

$$\log \mathbb{E}_{P_W P_{\boldsymbol{Z}}} \left[e^{\lambda \left(L_{\boldsymbol{Z}}(W) - \mathbb{E}_{P_W}[L_{P_Z}(W)] \right)} \right] \leq \frac{\lambda^2 \sigma^2}{2n}. \tag{4.5}$$

Finally, we observe that

$$\inf_{\lambda > 0} \left(\frac{\lambda \sigma^2}{2n} + \frac{I(W; \boldsymbol{Z})}{\lambda} \right) = \sqrt{\frac{2\sigma^2 I(W; \boldsymbol{Z})}{n}}, \tag{4.6}$$

from which the result follows. □

This result subsumes Theorem 2.2, which is a special case for bounded random variables. Also, we note that the same result also holds under a slightly different assumption. Instead of the stated sub-Gaussianity and absolute continuity assumptions, one can instead assume that for all $w \in \mathcal{W}$, $\ell(w, Z)$ is σ-sub-Gaussian under P_Z and $P_{\boldsymbol{Z}|W=w} \ll P_{\boldsymbol{Z}}$. Then, instead of the approach used in Proposition 4.1, we consider a random W and use Donsker-Varadhan's variational representation to change measure between $P_{\boldsymbol{Z}|W}$ and $P_{\boldsymbol{Z}}$. The rest of the proof is essentially identical, and we average over P_W in the end. In Section 4.3, we will discuss how one can also derive a slightly tighter bound under this different sub-Gaussianity assumption.

At first glance, Corollary 4.2 seems to suggest that the generalization error decays as $1/\sqrt{n}$. However, when discussing the dependence on n

of the right-hand side of (4.3), one needs to remember that $I(W;\boldsymbol{Z})$ is also an implicit function of n via $P_{W\boldsymbol{Z}}$. In order for this result to be valuable when discussing generalization, we want the upper bound to approach zero as the number of training points goes to infinity, which implies that $I(W;\boldsymbol{Z}) = o(n)$. Therefore, for most settings of interest, we require that $I(W;\boldsymbol{Z})$ is sublinear in n.

For bounded losses, an alternative generalization bound can be derived using the concentration result for the binary relative entropy. We present this below.

Corollary 4.3. Assume that the range of the loss function $\ell(\cdot,\cdot)$ is $[0,1]$ and that $P_{W\boldsymbol{Z}} \ll P_W P_{\boldsymbol{Z}}$. Then,

$$d\Big(\hat{L}\,||\,L\Big) \leq \frac{I(W;\boldsymbol{Z})}{n}. \tag{4.7}$$

Proof. We begin by noting that, due to Jensen's inequality and the convexity of $d_\gamma(\cdot\,||\,\cdot)$,

$$d\Big(\hat{L}\,||\,L\Big) = \sup_\gamma d_\gamma\Big(\hat{L}\,||\,L\Big) \leq \sup_\gamma \mathbb{E}_{P_{W\boldsymbol{Z}}}\left[d_\gamma(L_{\boldsymbol{Z}}(W)\,||\,L_{P_Z}(W))\right]. \tag{4.8}$$

We proceed by applying (4.2) with $f(W,\boldsymbol{Z}) = nd_\gamma(L_{\boldsymbol{Z}}(W)\,||\,L_{P_Z}(W))$. Since $\ell(\cdot,\cdot) \in [0,1]$, we can apply Theorem 3.31 and obtain

$$\log \mathbb{E}_{P_W P_{\boldsymbol{Z}}}\left[e^{nd_\gamma\left(L_{\boldsymbol{Z}}(W)\,||\,L_{P_Z}(W)\right)}\right] \leq 0. \tag{4.9}$$

From this, the result immediately follows. □

We pause here to discuss the fact that we used the $d_\gamma(\cdot\,||\,\cdot)$ function as the starting point of our proof, while the end result is expressed in terms of the regular binary relative entropy $d(\cdot\,||\,\cdot)$. The reason for this is that this approach allowed us to use Theorem 3.31 for the concentration of measure step, instead of Theorem 3.29. Had we not done this, we would end up with an additional $\log(2\sqrt{n})/n$ term in our final result. Crucially, this was possible due to the fact that we derived a generalization bound in expectation, instead of a bound in probability. Had we concerned ourselves with tail bounds, the supremization over γ would have been problematic, and would have necessitated using a union bound (or something to that effect).

The tightest explicit bounds on the population loss that can be obtained based on Corollary 4.3 rely on a numerical inversion of the binary relative entropy. However, by using the upper bound on $d^{-1}(\cdot,\cdot)$ provided in Proposition 3.30, we can obtain a closed-form relaxation that gives some insight into the n-dependence of the bound. In particular, for the case of zero training loss, this relaxation reduces to

$$L \leq \frac{2I(W;\boldsymbol{Z})}{n}. \tag{4.10}$$

Thus, for the case where $I(W;\boldsymbol{Z})$ is sublinear in n, (4.10) gives a faster decay rate with respect to n than the sub-Gaussian bound (4.3), at the cost of a multiplicative constant (the factor 2 in (4.10)).

The bounds that we have discussed so far in this section are all based on the Donsker-Varadhan variational representation of the relative entropy. As previously indicated, alternative changes of measure can be used—for instance, those based on f-divergences in Theorem 3.19. We will now present a generalization bound derived using Theorem 3.19, following Jiao *et al.* (2017).[1]

Theorem 4.4. Let $\|\ell(w,\cdot)\|_\beta = \mathbb{E}_{P_Z}^{1/\beta}\left[|\ell(w,Z) - L_{P_Z}(w)|^\beta\right]$ and assume that $\|\ell(w,\cdot)\|_\beta \leq \sigma_\beta$ for some $\beta > 1$. Also, let $f_\alpha(x) = |x-1|^\alpha$ for some $\alpha \geq 1$ satisfying $\frac{1}{\alpha} + \frac{1}{\beta} = 1$. Then,

$$\overline{\text{gen}} \leq \sigma_\beta D_{f_\alpha}(P_{W\boldsymbol{Z}} \,||\, P_W P_{\boldsymbol{Z}})^{1/\alpha}. \tag{4.11}$$

In particular, if $\alpha = 1$ (so that $\beta \to \infty$),

$$\overline{\text{gen}} \leq \sigma_\infty \text{TV}(P_{W\boldsymbol{Z}}, P_W P_{\boldsymbol{Z}}). \tag{4.12}$$

Proof. We apply Theorem 3.19 with $P = P_{W\boldsymbol{Z}}$, $Q = P_W P_{\boldsymbol{Z}}$, and $\phi = \lambda\text{gen}(w,\boldsymbol{z})$, for some $\lambda > 0$, and obtain

$$D_{f_\alpha}(P_{W\boldsymbol{Z}} \,||\, P_W P_{\boldsymbol{Z}}) \geq \mathbb{E}_{P_{W\boldsymbol{Z}}}[\text{gen}(W,\boldsymbol{Z})] - \mathbb{E}_{P_W P_{\boldsymbol{Z}}}[f_\alpha^*(\text{gen}(w,\boldsymbol{z}))]\,. \tag{4.13}$$

Explicitly computing the convex conjugate f_α^* and optimizing over λ, we obtain the desired result. □

[1] Jiao *et al.* (2017) state their result in terms of adaptive data analysis, in the same vein as Russo and Zou (2016). In Theorem 4.4, we provide a simple adaptation that yields a generalization bound.

As noted by Jiao *et al.* (2017, Sec. 4), an alternative way to derive this bound is through Hölder's inequality. The benefit of this bound is clear: we only needed a bound on the central β-moment of the loss function, which allows for a broader range of loss functions than (*e.g.*) sub-Gaussian ones. Furthermore, the f-divergences that appear in Theorem 4.4 can be bounded even when $P_{W\boldsymbol{Z}} \ll P_W P_{\boldsymbol{Z}}$ does not hold. Therefore, Theorem 4.4 enables us to consider more general distributions and learning algorithms than Corollary 4.2. One drawback, however, is that the dependence of the bound on the number of samples n is less explicit.

4.2 The Randomized-Subset and Individual-Sample Technique

One issue with the bounds from the previous section is that the mutual information that appears in them is infinite if the required absolute continuity criterion does not hold. For instance, consider the bound in Corollary 4.2. If W and $\boldsymbol{Z}$ are separately continuous random variables and W is a deterministic function of $\boldsymbol{Z}$, the absolute continuity criterion $P_{W\boldsymbol{Z}} \ll P_W P_{\boldsymbol{Z}}$ fails to hold. Hence, the mutual information $I(W;\boldsymbol{Z})$ is unbounded. A similar issue arises for many other information measures that appear in information-theoretic and PAC-Bayesian bounds. A separate problem, but which can be solved in the same way, is the lack of an explicit decay with n in Theorem 4.4.

A possible remedy for this is to use the *randomized-subset* technique, wherein the linearity of the expectation operator is used to obtain an average bound for the loss on randomly chosen subsets of the training set, rather than the loss averaged over the full training set. A special case of this is the *individual-sample* technique, where the random subsets are single samples chosen uniformly at random. In the case where W is a deterministic function of $\boldsymbol{Z}$ but not of any individual sample Z_i, this avoids the unboundedness issue. This technique was introduced by Bu *et al.* (2020).

Proposition 4.5 (The randomized-subset technique)**.** Consider a population loss bound ub_n such that $L \leq \mathrm{ub}_n(\mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{\boldsymbol{Z}}(W)], P_{W\boldsymbol{Z}})$ for any $n \in \mathbb{N}^+$. For any subset $\mathcal{M} \subseteq [n]$, let $L_{\boldsymbol{Z}_{\mathcal{M}}}(W) = \sum_{i\in\mathcal{M}} \ell(W, Z_i)/\left|\mathcal{M}\right|$ denote the training loss on the samples $\boldsymbol{Z}_{\mathcal{M}} = \{Z_i\}_{i\in\mathcal{M}}$, Let $P_{\mathcal{M}}$ be an arbitrary probability mass function on subsets of $[n]$, and assume that $\mathcal{M} \sim P_{\mathcal{M}}$. Then,

$$L \le \mathbb{E}_{P_{\mathcal{M}}}\left[\mathrm{ub}_{|\mathcal{M}|}(\mathbb{E}_{P_{W\boldsymbol{Z}_{\mathcal{M}}}}[L_{\boldsymbol{Z}_{\mathcal{M}}}(W)]\,, P_{W\boldsymbol{Z}_{\mathcal{M}}})\right]. \tag{4.14}$$

In particular, if $P_{\mathcal{M}}$ is the uniform distribution on $[n]$,

$$L \le \frac{1}{n}\sum_{i=1}^{n}\left[\mathrm{ub}_1(\mathbb{E}_{P_{WZ_i}}[\ell(W,Z_i)]\,, P_{WZ_i})\right]. \tag{4.15}$$

Proof. For any fixed subset $\mathcal{M}$, the fact that the samples are i.i.d. implies that

$$\mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{\boldsymbol{Z}_{\mathcal{M}}}(W)] = \mathbb{E}_{P_{W\boldsymbol{Z}_{\mathcal{M}}}}[L_{\boldsymbol{Z}_{\mathcal{M}}}(W)] = \mathbb{E}_{P_{W\boldsymbol{Z}_{[|\mathcal{M}|]}}}\left[L_{\boldsymbol{Z}_{[|\mathcal{M}|]}}(W)\right], \tag{4.16}$$

where $[|\mathcal{M}|] = \{1,\dots,|\mathcal{M}|\}$. The result now follows by applying the generalization bound $\mathrm{ub}_{|m|}$ and averaging over M. □

Below, we illustrate the technique by deriving the individual-sample mutual information bound of Bu *et al.* (2020), who introduced the technique.

Corollary 4.6. Assume that the loss function $\ell(W, Z_i)$ is σ-sub-Gaussian under $P_W P_{Z_i}$ and that $P_{WZ_i} \ll P_W P_{Z_i}$. Then,

$$\overline{\mathrm{gen}} \le \frac{1}{n}\sum_{i=1}^{n}\sqrt{2\sigma^2 I(W;Z_i)}. \tag{4.17}$$

Proof. While the result follows immediately by combining Corollary 4.2 and Proposition 4.5, we give a self-contained proof below. By exploiting the linearity of the expectation operator and marginalizing out the data points that do not appear in a given term, we see that

$$\begin{aligned}\overline{\mathrm{gen}} &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{P_Z}(W) - \ell(W,Z_i)] && (4.18)\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{P_{WZ_i}}[\mathrm{gen}(W,Z_i)]\,. && (4.19)\end{aligned}$$

The result now follows by applying Corollary 4.2 to each term in the sum. □

The individual-sample mutual-information bound in Corollary 4.6 is tighter than its full-sample counterpart in Corollary 4.2. Specifically, with $\boldsymbol{Z}_{<i} = (Z_1, \dots, Z_{i-1})$ (where $\boldsymbol{Z}_{<1} = \emptyset$),

$$\frac{1}{n}\sum_{i=1}^{n}\sqrt{2\sigma^2 I(W;Z_i)} \le \sqrt{\frac{2\sigma^2}{n}\sum_{i=1}^{n} I(W;Z_i)} \tag{4.20}$$

$$\le \sqrt{\frac{2\sigma^2}{n}\sum_{i=1}^{n} I(W;Z_i|\boldsymbol{Z}_{<i})} \tag{4.21}$$

$$= \sqrt{\frac{2\sigma^2}{n} I(W;\boldsymbol{Z})}. \tag{4.22}$$

Here, the first step follows from Jensen's inequality, the second step uses the fact that conditioning on independent random variables does not decrease mutual information (here, Z_i and $\boldsymbol{Z}_{<i}$ are independent), and the third follows from the chain rule of mutual information. In fact, as demonstrated by Harutyunyan *et al.* (2021), a wide family of randomized-subset generalization bounds are non-decreasing functions of the subset size.

The benefit of the individual-sample technique is two-fold. First, as mentioned above, it leads to bounds that depend on the sum of functions of the mutual information between the hypothesis and the individual samples, which can sometimes be shown to be tighter than bounds that depend on the mutual information between W and the full training set $\boldsymbol{Z}$. Second, we can obtain an explicit decay with n even from bounds that are constant with respect to n (ignoring the implicit dependence on n of the information measure). For instance, consider the f-divergence bound in Theorem 4.4. By applying the individual-sample technique, we instead obtain

$$\overline{\text{gen}} \le \frac{\sigma_\beta}{n}\sum_{i=1}^{n} D_{f_\alpha}(P_{WZ_i} \,||\, P_W P_{Z_i})^{1/\alpha}. \tag{4.23}$$

While the chain rule of the relative entropy allowed us to recover the full-sample counterparts of generalization bounds from the individual-sample versions, the same does not hold for the f-divergence bound in (4.23), as the f-divergences do not satisfy the chain rule in general.

Hence, it is less clear to what extent the individual-sample technique actually does yield an improvement, and a more careful characterization of the information measures is needed for each case.

The technique just described may similarly be applied to obtain a samplewise version of Corollary 4.3. Now, assume that the learning algorithm is designed so that the loss is zero for all training samples, and assume also that the loss is bounded to $[0, 1]$. Recall that for the sub-Gaussianity-based bound in Corollary 4.2, this implies that $\sigma = 1/2$. Now, while we used the same generalization bound for each subset in Proposition 4.5, it is clear that, instead, we could apply a different generalization bound for each subset. For instance, when using an individual-sample decomposition, we can apply the minimum of the bounds from Corollaries 4.2 and 4.3 to each term to obtain

$$L \leq \frac{1}{n}\sum_{i=1}^{n} \min\left\{\sqrt{I(W; Z_i)/2}, 2I(W; Z_i)\right\}. \tag{4.24}$$

Notice that there is not a general ordering between these two bounds: their ordering depends on the specific value of the mutual information. Specifically, if the mutual information is lower than $1/8$, the minimum is achieved by the second term. Otherwise, the first term achieves the minimum.

While samplewise bounds are powerful tools to obtain average generalization bounds, it can be shown that there are certain formal limitations to how they can be used to obtain bounds on the averaged squared generalization error, as well as bounds in probability, as shown by Harutyunyan *et al.* (2022). We will discuss this further in Chapter 5.

An alternative approach to obtain samplewise bounds, based on the convexity of probability measures, was taken by Aminian *et al.* (2022b). Their approach is based on the following observation: by the linearity of expectation, we can rewrite the average generalization error as

$$\overline{\text{gen}} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{P_W P_{Z_i}}[\ell(W, Z_i)] - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{P_{WZ_i}}[\ell(W, Z_i)] \tag{4.25}$$

$$= \mathbb{E}_{P_W P_{\bar{Z}}}[\ell(W, Z)] - \mathbb{E}_{P_{W\bar{Z}}}[\ell(W, Z)] \tag{4.26}$$

$$= \mathbb{E}_{P_{\bar{Z}}}\left[\mathbb{E}_{P_W}[\ell(W, Z)] - \mathbb{E}_{P_{W|\bar{Z}}}[\ell(W, Z)]\right]. \tag{4.27}$$

Here, $P_{\bar{Z}} = \frac{1}{n}\sum_{i=1}^{n} P_{Z_i}$, $P_{W\bar{Z}} = \frac{1}{n}\sum_{i=1}^{n} P_{WZ_i}$, and $P_{W|\bar{Z}} = \frac{1}{n}\sum_{i=1}^{n} P_{W|Z_i}$. Using this characterization, generalization bounds can be derived by similar techniques as already discussed by changing measure from $P_{W\bar{Z}}$ to $P_W P_{\bar{Z}}$. We will give an explicit example of such bounds in Theorem 4.15 in Section 4.5.

4.3 Disintegration

If the generalization bound is a concave function—as is the case for the square-root bound in (4.3)—moving expectations outside of the generalization bound leads to a tighter characterization, by Jensen's inequality. We used this when showing that the individual-sample technique led to a tighter bound than its full-sample counterpart in (4.20). This insight, referred to as *disintegration* (Negrea *et al.*, 2019), can be exploited further to derive bounds where additional expectations are moved outside a concave bound. Consider the derivation of Corollary 4.6: essentially, a sample index i is fixed, the bound is derived, and only *then* do we average over i. In fact, the same can be done for W, under the slightly altered sub-Gaussianity assumption discussed after Corollary 4.2.

Corollary 4.7. Assume that, for all $w \in \mathcal{W}$, the loss function $\ell(w, Z)$ is σ-sub-Gaussian under P_Z and $P_{\boldsymbol{Z}|W=w} \ll P_{\boldsymbol{Z}}$. Then,

$$\overline{\text{gen}} \le \mathbb{E}_{P_W}\left[\sqrt{\frac{2\sigma^2 D(P_{\boldsymbol{Z}|W} \,||\, P_{\boldsymbol{Z}})}{n}}\right]. \tag{4.28}$$

Proof. By the Donsker-Varadhan variational representation of the relative entropy (Theorem 3.17), we have for all $w \in \mathcal{W}$ that

$$\mathbb{E}_{P_{\boldsymbol{Z}|W=w}}[\lambda \text{gen}(W, \boldsymbol{Z})] \le \log \mathbb{E}_{P_{\boldsymbol{Z}}}\left[e^{\lambda \text{gen}(W,\boldsymbol{Z})}\right] + D(P_{\boldsymbol{Z}|W=w} \,||\, P_{\boldsymbol{Z}}). \tag{4.29}$$

By applying our sub-Gaussianity assumption and optimizing over λ, we get

$$\mathbb{E}_{P_{\boldsymbol{Z}|W=w}}[\text{gen}(W, \boldsymbol{Z})] \le \sqrt{\frac{2\sigma^2 D(P_{\boldsymbol{Z}|W=w} \,||\, P_{\boldsymbol{Z}})}{n}}. \tag{4.30}$$

The result follows after averaging over P_W. □

Naturally, the disintegration approach can be combined with the randomized-subset technique. We highlight that Corollary 4.7 relies on the alternative sub-Gaussianity assumption, and (4.28) does not hold under the sub-Gaussianity assumption used in Corollary 4.2. For some cases, such as the important special case of bounded losses, both assumptions hold.

4.4 Chaining

As previously discussed, one of the main draws of the information-theoretic approach to studying generalization is that it allows us to capture the dependence between the training data and the hypothesis that is induced by the learning algorithm. However, as pointed out by Asadi *et al.* (2018), one relevant aspect that is missing from the bounds that we have seen so far in this chapter is the dependence *between hypotheses*. If one learning algorithm, for different $\boldsymbol{Z}$, selects hypotheses that are distinct but similar to each other in some sense—for instance, as measured by a metric on the hypothesis space—we may expect it to behave very differently from a learning algorithm that selects distinct, highly dissimilar hypotheses. This, however, can go unnoticed by quantities such as the mutual information, since it depends only on the probability measures that are involved, but not the underlying hypothesis space. Hence, if there is a bijection between the output sets for the aforementioned learning algorithms that preserves probabilities, they would be equivalent in terms of mutual information.

One approach to incorporate dependencies between hypotheses is to use the *chaining* technique. Intuitively, this approach consists of looking at the hypothesis space at a coarse level, and approximating the mutual information with increasingly fine granularity. To introduce chaining formally, we need the following definition.

Definition 4.8 (ε-partitions and increasing sequences)**.** Let $\mathcal{W}$ be a set endowed with the metric d. A partition $\mathcal{P} = \{A_1, \dots, A_m\}$, comprising disjoint sets $A_1, \dots, A_m$ such that $\mathcal{W} = \cup_{i=1^m} A_i$, is an ε-partition of $\mathcal{W}$ if for each A_i with $i \in [m]$, there exists a $w_i \in \mathcal{W}$ such that $A_i \subseteq \mathcal{B}_d(w_i, \varepsilon)$, where $\mathcal{B}_d(w_i, \varepsilon) = \{w \in \mathcal{W} : d(w, w_i) \leq \varepsilon\}$ is the ball of radius ε centered around w_i. A sequence of partitions $\{\mathcal{P}_k\}_{k=k'}^{\infty}$ is called *increasing* if, for all $k \geq k'$ and each $A \in \mathcal{P}_{k+1}$, there exists $B \in \mathcal{P}_k$ such that $A \subseteq B$.

In order to state the generalization bound, we also need to define sub-Gaussian processes. The sub-Gaussianity here is essentially the same as we have seen before, with the metric $d(\cdot,\cdot)$ from Definition 4.8 incorporated into the sub-Gaussianity parameter.

Definition 4.9 (Sub-Gaussian process)**.** The random process $\{X_w\}_{w\in\mathcal{W}}$ is *sub-Gaussian* for the metric space $(\mathcal{W}, d)$ if $\mathbb{E}[X_w] = 0$ for all $w \in \mathcal{W}$ and, for all $w, w' \in \mathcal{W}$ and $\lambda \in \mathbb{R}$,

$$\log \mathbb{E}\left[e^{\lambda(X_w - X_{w'})}\right] \leq \frac{\lambda^2 d^2(w,w')}{2}. \tag{4.31}$$

For the result, we also need to assume that the process is separable (Asadi *et al.*, 2018, Definition 2), which is a technical assumption that we refrain from stating explicitly for brevity. We are now ready to state a generalization bound in terms of the chained mutual information for a bounded hypothesis space, due to Asadi *et al.* (2018).

Theorem 4.10. Assume that $\{\mathrm{gen}(w, \boldsymbol{Z})\}_{w\in\mathcal{W}}$ is a separable sub-Gaussian process on $\mathcal{W}$ with metric $d(\cdot,\cdot)$. Furthermore, assume that the diameter of $\mathcal{W}$, defined as $\mathrm{diam}(\mathcal{W}) = \max_{w,w'\in\mathcal{W}} d(w,w')$, is finite. Let $\{\mathcal{P}_k\}_{k=k_1}^{\infty}$ be an increasing sequence of partitions such that, for each $k \geq k_1$, $\mathcal{P}_k$ is a 2^{-k}-partition of $\mathcal{W}$. For each $w \in \mathcal{W}$ and $k \geq k_1$, let $[w]_k$ denote the unique $A \in \mathcal{P}_k$ such that $w \in A$. Then,

$$\overline{\mathrm{gen}} \leq 3\sqrt{2} \sum_{k=k_1}^{\infty} 2^{-k} \sqrt{I([W]_k; \boldsymbol{Z})}. \tag{4.32}$$

As k increases, the mutual information $I([W]_k; \boldsymbol{Z})$ is evaluated on a finer partition of $\mathcal{W}$, which yields an increasingly accurate estimate of the mutual information $I(W; \boldsymbol{Z})$. In fact, the sequence is increasing, and $I([W]_k; \boldsymbol{Z}) \to I(W; \boldsymbol{Z})$ as $k \to \infty$ (Cover and Thomas, 2006, Eq. (8.54)). Thus, as k increases, so does our estimate of the mutual information, but the higher-k contributions are exponentially discounted with 2^{-k}. Relatively speaking, the lower-k contributions are therefore more influential, and due to the coarse partitions, these incorporate dependence between hypotheses, leading to lower mutual information. Indeed, for two distinct $w, w' \in \mathcal{W}$, we have $[w]_k = [w']_k$ if they are sufficiently close as measured by d.

As pointed out by Zhou *et al.* (2022), the bound above has some limitations. Firstly, the hypothesis space is required to be bounded, which precludes many simple settings. The deterministic and hierarchical partitions also impose certain geometric constraints, and can render the bound challenging to compute. This can be addressed by using a *stochastic* chaining procedure. Drawing inspiration from multilevel quantization in data compression, Zhou *et al.* (2022) derive a similar result as Theorem 4.10 in terms of a stochastic partition, as formalized in the following.

Theorem 4.11. Assume that $\{\mathrm{gen}(w, \boldsymbol{Z})\}_{w\in\mathcal{W}}$ is a separable sub-Gaussian process on $\mathcal{W}$ with metric $d(\cdot,\cdot)$. Let $\{W_k\}_{k=k_0}^{\infty}$ be a sequence of random variables on $\mathcal{W}$ such that:

1. $\lim_{k\to\infty} \mathbb{E}_{P_{W_k \boldsymbol{Z}}}[\mathrm{gen}(W_k, \boldsymbol{Z})] = \mathbb{E}_{P_{W\boldsymbol{Z}}}[\mathrm{gen}(W, \boldsymbol{Z})]$,
2. $\mathbb{E}_{P_{W_{k_0} \boldsymbol{Z}}}[\mathrm{gen}(W_{k_0}, \boldsymbol{Z})] = 0$, and
3. $\{\mathrm{gen}(w, \boldsymbol{Z})\}_{w\in\mathcal{W}} - W - W_k - W_{k-1}$ is a Markov chain for every $k > k_0$.

Then,

$$\overline{\mathrm{gen}} \leq \sum_{k=k_0+1}^{\infty} \sqrt{\mathbb{E}[d^2(W_k, W_{k-1})]} \sqrt{2I([W]_k; \boldsymbol{Z})}. \tag{4.33}$$

From the result in Theorem 4.11, we can recover the one in Theorem 4.10 by setting $\{W_k\}_{k=k_0}^{\infty}$ as the deterministic sequence that appears therein. Naturally, this bound can be combined with the disintegration and samplewise techniques, as detailed by Zhou *et al.* (2022).

4.5 Bounds via the Kantorovich-Rubinstein Duality

An alternative approach to obtain bounds that incorporate the dependence between hypotheses and the geometry of the hypothesis class is to use tools from optimal transport. Recall the Wasserstein distance, introduced in Definition 3.15. This information measure is defined in terms of a metric, and suitable choices for this metric enable the possibility of incorporating dependencies between hypotheses. This approach obviates the need for absolute continuity, since the Wasserstein distance is still defined and finite in its absence.

The key tool in deriving these bounds is the Kantorovich-Rubinstein (KR) duality, stated in Theorem 3.21, which relates the difference in expectation under two different distributions to the Wasserstein distance between them. Below, we state a first result based on Theorem 3.21, given by Wang *et al.* (2019a).

Theorem 4.12. Recall that $P_{\bar{Z}} = \frac{1}{n}\sum_{i=1}^n P_{Z_i}$, $P_{W\bar{Z}} = \frac{1}{n}\sum_{i=1}^n P_{WZ_i}$, and $P_{W|\bar{Z}} = \frac{1}{n}\sum_{i=1}^n P_{W|Z_i}$. Assume that the loss function $\ell(\cdot, z)$ is L-Lipschitz on $\mathcal{W}$ for all $z \in \mathcal{Z}$. Then,

$$|\overline{\mathrm{gen}}| \leq L\, \mathbb{E}_{P_{\boldsymbol{Z}}}\Big[\mathbb{W}_1(P_{W|\boldsymbol{Z}}, P_W)\Big]\,. \tag{4.34}$$

Proof. Since the loss is L-Lipschitz, the loss normalized by L is 1-Lipschitz. The result follows immediately by applying Theorem 3.21 with $f = L_{\boldsymbol{Z}}(W)$, $P = P_{W\boldsymbol{Z}}$, and $Q = P_W P_{\boldsymbol{Z}}$. □

For this bound to decay to zero as n approaches infinity, the average Wasserstein distance $\mathbb{E}_{P_{\boldsymbol{Z}}}\Big[\mathbb{W}_1(P_{W|\boldsymbol{Z}}, P_W)\Big]$ would need to be a *decreasing* function of n. We would prefer that the bound explicitly decays with n (when ignoring the Wasserstein distance). This can, for instance, be achieved by observing that if $\ell(w, z)$ is L-Lipschitz under the p-norm, $L_{\boldsymbol{z}}(w)$ is $L/n^{1/p}$-Lipschitz under the p-norm. Under this assumption, the following bound can be derived (Lopez and Jog, 2018, Thm. 1).

Theorem 4.13. Assume that, for some $p \geq 1$, the loss function $\ell(w, \cdot)$ is L-Lipschitz on $\mathcal{Z}$ under the p-norm for all $w \in \mathcal{W}$. Then,

$$|\overline{\mathrm{gen}}| \leq \frac{L}{n^{1/p}}\, \mathbb{E}_{P_{\boldsymbol{Z}}}^{1/p}\Big[(\mathbb{W}_p(P_{\boldsymbol{Z}|W}, P_{\boldsymbol{Z}}))^p\Big]\,. \tag{4.35}$$

Note that this bound is in terms of the “backwards channel” $P_{\boldsymbol{Z}|W}$, which is a result of the Lipschitz assumption being with respect to $\mathcal{Z}$.

While the proof of Lopez and Jog (2018) is a bit more involved, the result for $p = 1$ follows immediately from KR duality.

For generalization bounds in expectation, we have an alternative tool at our disposal: the individual-sample technique. By applying this, as stated in Proposition 4.5, to the bound in Theorem 4.12, we obtain the following (Rodríguez-Gálvez *et al.*, 2021b).

Theorem 4.14. Assume that the loss function $\ell(\cdot, z)$ is L-Lipschitz on $\mathcal{W}$ for all $z \in \mathcal{Z}$. Then,

$$|\overline{\text{gen}}| \leq \frac{L}{n} \sum_{i=1}^{n} \mathbb{E}_{P_{Z_i}} \Big[\mathbb{W}_1(P_{W|Z_i}, P_W) \Big] . \tag{4.36}$$

Proof. We first use the samplewise decomposition from the proof of Corollary 4.6 to obtain

$$\overline{\text{gen}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{P_Z}(W) - \ell(W, Z_i)] \tag{4.37}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{P_W P_{Z_i}}[\ell(W, Z_i)] - \mathbb{E}_{P_{WZ_i}}[\ell(W, Z_i)] . \tag{4.38}$$

Since each term in the sum is a difference between the average of a random variable taken under two different distributions, the result immediately follows by applying the KR duality given in Theorem 3.21. □

As for the mutual information-based bounds, the samplewise Wasserstein bound can be shown to be tighter than its full-sample counterpart (Rodríguez-Gálvez *et al.*, 2021b).

As mentioned earlier, an alternative approach to obtain samplewise bounds is through the convexity of probability measures (Aminian *et al.*, 2022b). Using the decomposition in (4.27) and applying the KR duality, we obtain the following result.

Theorem 4.15. Assume that the loss function $\ell(\cdot, z)$ is L-Lipschitz on $\mathcal{W}$ for all $z \in \mathcal{Z}$. Then,

$$\overline{\text{gen}} \leq L \, \mathbb{E}_{P_{\boldsymbol{Z}}} \Big[\mathbb{W}_1(P_{W|\bar{Z}}, P_W) \Big] . \tag{4.39}$$

This bound is always tighter than the one in Theorem 4.14, due to Jensen's inequality and the convexity of the supremum. For symmetric learning algorithms, where $P_{W|Z_i}$ is the same for all i, the bounds in Theorem 4.14 and Theorem 4.15 coincide.

For bounded losses, the Wasserstein-based bound in Theorem 4.14 can be shown to be tighter than the corresponding slow-rate bound in Corollary 4.6 based on the mutual information. This improvement also has a clear interpretation, as discussed by Rodríguez-Gálvez *et al.* (2021b): the Wasserstein distance can account for structure within the hypothesis class by an appropriate choice of metric. If we use the discrete metric, $\rho_D(x,y) = 1\{x \neq y\}$, we discard this geometric information, but we are able to recover bounds based on the relative entropy. Specifically, for the discrete metric, $\mathbb{W}_1(P,Q) = \mathrm{TV}(P,Q)$ (Villani, 2008, Thm. 6.15). Therefore, we can use either Pinsker's inequality or the BH inequality (Theorem 3.10) to upper-bound the Wasserstein distance in Theorem 4.14. Specifically, consider a bounded loss function, with range restricted to $[0,1]$. Then, it is 1-Lipschitz for any $z \in \mathcal{Z}$ under the discrete metric on $\mathcal{W}$, *i.e.*, $\rho_D(w,w') = 1\{w \neq w'\}$. By applying the upper bound from (3.14) to the Wasserstein distance in Theorem 4.14, we get

$$|\overline{\mathrm{gen}}| \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{P_{Z_i}}\left[\sqrt{\frac{D(P_{W|Z_i} \,||\, P_W)}{2}}\right] \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{I(W;Z_i)}{2}}, \tag{4.40}$$

where the second step is due to Jensen's inequality. This illustrates that the Wasserstein-based bound is tighter for any bounded loss, through the use of Pinsker's inequality. Furthermore, the Wasserstein-based bound is never vacuous for this setting. When the relative entropy is high, the BH inequality in (3.15) gives a tighter upper-bound than Pinsker's inequality in (3.14) does, and in particular, it is never greater than 1.

Similar arguments can also be made in more general settings. For instance, if ρ is an arbitrary metric, we still have

$$\mathbb{W}_1(P,Q) \leq d_\rho(\mathcal{X})\mathrm{TV}(P,Q) \tag{4.41}$$

where $d_\rho(\mathcal{X})$ denotes the diameter of $\mathcal{X}$. The relaxation in terms of the BH and Pinsker's inequality can thus be relevant for other metrics, provided that the diameter of the hypothesis space is bounded.

Furthermore, a Pinsker-type relaxation can also be obtained under a sub-Gaussianity assumption. Specifically, consider a probability distribution P_X on $\mathcal{X}$. If every 1-Lipschitz function $f : \mathcal{X} \to \mathbb{R}$ is σ-sub-Gaussian under P_X, we have (Van Handel, 2016)

$$\mathbb{W}_1(P, Q) \leq \sqrt{2\sigma^2 D(P \,||\, Q)}. \tag{4.42}$$

If the loss function $\ell(W, z)$ is 1-Lipschitz and sub-Gaussian under P_W for all $z \in \mathcal{Z}$, the Wasserstein-based bounds can thus also be shown to be tighter than the corresponding ones based on the relative entropy. Note that this is a different sub-Gaussianity assumption than the one used in Section 4.1, where the loss was instead assumed to be sub-Gaussian under P_Z. Further discussion of this, including the relation to the backward channel, can be found in (Rodríguez-Gálvez *et al.*, 2021b, Sec. B.1).

4.6 Bibliographic Remarks and Additional Perspectives

In Section 2.4, we discussed the history of information-theoretic generalization bounds and the connection to PAC-Bayesian theory. In this section, we will specify how the results that we covered in this chapter are related to existing literature, and give a brief overview of some results that we did not cover in detail. We note that we will not discuss results for the conditional mutual information (CMI) framework, as these will be covered in Chapter 6.

As previously pointed out, Proposition 4.1 is simply a restatement of the Donsker-Varadhan variational representation of the relative entropy. This generic formulation of a generalization bound is similar to many PAC-Bayesian bounds which are stated for generic (convex) functions, like those of Alquier and Guedj (2018), Bégin *et al.* (2016), Germain *et al.* (2009a), and Rivasplata *et al.* (2020). The bound in Corollary 4.2, under the alternative sub-Gaussianity assumption, is due to Xu and Raginsky (2017), but, as discussed in Section 2.4, appeared around the same time in similar forms under slightly different assumptions. The use of the two different assumptions is explicitly discussed by, *e.g.*, Rodríguez-Gálvez *et al.* (2020). The bound with the binary KL divergence in Corollary 4.3 is essentially implicit in the work of McAllester (2013), but only explicitly stated in a looser form. The exact statement was given explicitly in Hellström and Durisi (2022a). Theorem 4.4 was stated by Jiao *et al.* (2017) for adaptive data analysis, but obtaining an analogous generalization bound is straightforward by following the same procedure as Xu and Raginsky (2017) used when adapting the result of Russo and Zou (2016).

The randomized-subset and individual-sample technique of Proposition 4.5 was introduced by Bu *et al.* (2020) and Bu *et al.* (2019), and subsequently applied in more settings by, *e.g.*, Haghifam *et al.* (2020), Hellström and Durisi (2021a), Negrea *et al.* (2019), Rodríguez-Gálvez *et al.* (2020), and Zhou *et al.* (2021), who also introduced the notion of disintegration (Section 4.3). Several key properties of the randomized-subset approach were studied by Harutyunyan *et al.* (2021, 2022), establishing the general dependence on the size of the subsets and the impossibility of obtaining bounds on the average squared generalization gap in terms of the individual-sample mutual information. Aminian *et al.* (2022a) provided an alternative perspective, wherein the randomized subsets were instead considered for the probability measures themselves, leading to improved bounds for some settings.

A detailed discussion of chaining and its use in learning theory, where it is used to derive the tightest generalization bounds in terms of the VC dimension, can be found in the book by Vershynin (2018, Chapter 8). The chaining approach was combined with PAC-Bayesian theory by Audibert and Bousquet (2007). Theorem 4.10 is due to Asadi *et al.* (2018), while Theorem 4.11 is due to Zhou *et al.* (2022).

Bounds based on the Wasserstein distance, discussed in Section 4.5, were obtained by Raginsky *et al.* (2016) for learning algorithms that are stable in terms of Wasserstein distance, in the sense that the distribution of W does not change much (in terms of Wasserstein distance) if one sample of the training set is replaced. Wintenberger (2015) considered weak transport inequalities, and used this to obtain oracle inequalities with fast convergence rates. Results in terms of the Wasserstein distance between the conditional $P_{W|\boldsymbol{Z}}$ and its marginal P_W were derived independently by Lopez and Jog (2018) and Wang *et al.* (2019a). The bounds in Theorems 4.12 and 4.13 are based on these works. Tighter variants of these bounds were obtained by Rodríguez-Gálvez *et al.* (2021b), also through the use of the individual-sample technique. Aminian *et al.* (2022a) demonstrated that through the convexity of probability measures, tighter bounds can be obtained for non-symmetric learning algorithms. Finally, the chaining technique was generalized to information measures beyond the mutual information by Clerico *et al.* (2022b), who obtained bounds in terms of, for instance, the Wasserstein distance and various f-divergences.

We conclude this chapter by mentioning works on average generalization bounds that we did not cover in detail. Alabdulmohsin (2020) considered a notion of uniform generalization over all possible parametric loss functions, and showed that this is equivalent to $\mathrm{TV}(P_{W\boldsymbol{Z}}, P_W P_{\boldsymbol{Z}})$, termed the "variational information." Hafez-Kolahi *et al.* (2020) discussed methods of tightening information-theoretic generalization bounds through the techniques of conditioning and processing, based on a graphical model perspective. Many approaches, such as samplewise bounds and chaining, can be expressed through this framework. Aminian *et al.* (2020) obtained bounds in terms of the Jensen-Shannon divergence, which can be seen as a symmetrized version of the relative entropy. Modak *et al.* (2021) derived variants of the results of Xu and Raginsky (2017) in terms of the Rényi divergence of orders $\alpha \in (0, 1)$, which can potentially be tighter for some settings. Aminian *et al.* (2021b) considered bounds on higher moments of the generalization error, providing bounds in terms of mutual information and other information measures, based on for instance the χ^2 divergence. Raginsky *et al.* (2021) provided a comprehensive discussion of bounds in terms of information-theoretic stability, while Sefidgaran *et al.* (2022b) used tools from rate-distortion theory. Esposito and Gastpar (2022) derived a result that allows them to derive both generalization bounds and transportation-cost inequalities, and used this framework to obtain new bounds in terms of arbitrary divergence measure and recover known bounds in terms of, *e.g.*, the mutual information. Wongso *et al.* (2022, 2023) considered the *sliced* mutual information, based on one-dimensional random projections, and established connections to generalization both theoretically and empirically. Chu and Raginsky (2023) provide a unified approach to deriving information-theoretic bounds via a change of measure and Young's inequality. By incorporating other techniques, such as symmetrization and chaining, they obtain new bounds and recover several existing ones. Finally, Hafez-Kolahi *et al.* (2021) and Xu and Raginsky (2022) derived bounds for the minimum excess risk in Bayesian learning, while Hafez-Kolahi *et al.* (2023) derived information-theoretic bounds for the minimax excess risk and Dogan and Gastpar (2021) derived lower bounds on the expected excess risk.

5

Generalization Bounds in Probability

In the previous chapter, we considered generalization bounds in expectation. While this allowed for compact derivations, and enabled us to use techniques such as disintegration and randomized subsets that are effective only for average bounds, it is not sufficient for answering the most pertinent question that a practitioner may ask regarding generalization. Generalization bounds in expectation give us information about the generalization gap that we incur averaged over all possible datasets and all possible instantiations of our learning algorithm. While this is often sufficient to gain insight, in practice, we usually only have access to a specific, given dataset, and we do not know the distribution that generated it. In this case, we are interested in whether this specific dataset will allow generalization, and not in the performance averaged over other hypothetical datasets. Furthermore, the learning algorithm is often used only once for this given data set, and we only concern ourselves with the performance of the specific hypothesis that this yields, rather than the average performance when running the learning algorithm several times.

Motivated by these considerations, we now turn to generalization bounds in probability. We will focus on two flavors: first, *PAC-Bayesian* *bounds*, which hold with high probability over the draw of the training set, but are averaged over the learning algorithm. These bounds apply when we are concerned with *distributions* over the hypothesis class. Then, we look at *single-draw* bounds, which hold with high probability over the draw of both the dataset and a single hypothesis. This captures the situation that probably occurs most often in practice. Finally, we briefly discuss *mean-hypothesis* bounds, which are a sort of hybrid: high-probability bounds on the generalization error of the average hypothesis output from the learning algorithm, given the dataset.

As mentioned in the beginning of Chapter 4, there is a unified way to derive average, PAC-Bayesian, and single-draw bounds through exponential stochastic inequalities. Hence, many of the bounds that we present in this chapter imply corresponding average bounds. Below, we give a brief exposition of exponential stochastic inequalities, and then proceed with the PAC-Bayesian and single-draw bounds that are the main focus of this chapter.

5.1 Exponential Stochastic Inequality

In this section, we state a basic version of an exponential stochastic inequality, and demonstrate how it can be used to establish bounds of all three flavors (in expectation, PAC-Bayes, and single draw).

Theorem 5.1. Consider two random variables X and Y and two functions f and g such that, for all $\eta > 0$,

$$\mathbb{E}_{P_{XY}}\left[e^{\eta(f(X,Y)-g(X,Y))}\right] \leq 1. \tag{5.1}$$

Then, we have the "average" bound

$$\mathbb{E}_{P_{XY}}[f(X,Y)] \leq \mathbb{E}_{P_{XY}}[g(X,Y)]\,. \tag{5.2}$$

Furthermore, with probability at least $1-\delta$ over P_{XY}, we have the "single-draw" bound

$$f(X,Y) \leq g(X,Y) + \frac{\log\frac{1}{\delta}}{\eta}. \tag{5.3}$$

Finally, with probability at least $1 - \delta$ over P_Y, we have the "PAC-Bayesian" bound

$$\mathbb{E}_{P_{X|Y}}[f(X,Y)] \leq \mathbb{E}_{P_{X|Y}}[g(X,Y)] + \frac{\log \frac{1}{\delta}}{\eta}. \tag{5.4}$$

Proof. To obtain (5.2), we use Jensen's inequality to move the expectation inside the exponential. After re-arranging terms, the result follows. Next, to obtain (5.3), we note that (5.1) and Markov's inequality imply that

$$P_{XY}\left[e^{\eta(f(X,Y)-g(X,Y))} \leq \frac{1}{\delta}\right] \geq 1 - \mathbb{E}_{P_{XY}}\left[e^{\eta(f(X,Y)-g(X,Y))}\right]\delta \tag{5.5}$$

$$\geq 1 - \delta. \tag{5.6}$$

From this, (5.3) follows after re-arranging terms. Finally, to obtain (5.4), we first apply Jensen's inequality only with respect to $P_{X|Y}$. After using Markov's inequality in the same way as above, the stated result follows. □

When applying Theorem 5.1 to obtain generalization bounds, we will typically set $X = W$, $Y = \boldsymbol{Z}$, let f be a function of the generalization gap, and let g be a function of an information measure. The use of exponential inequalities to derive generalization bounds of different flavors can be traced back at least to the work of Zhang (2006) and Catoni (2007). For a more thorough discussion of this approach, see the recent work of Grünwald *et al.* (2023).

5.2 PAC-Bayesian Generalization Bounds

The PAC-Bayesian framework, originating in the seminal works of Shawe-Taylor and Williamson (1997) and McAllester (1998), is concerned with high-probability bounds, under the draw of the data, on the averaged loss of the learning algorithm. The learner, rather than selecting a specific hypothesis given the training data, selects a *distribution* over the hypothesis class. Then, when we want to use the hypothesis for whichever downstream task we are interested in, we draw a hypothesis according to the distribution. This is sometimes referred to as a "Gibbs classifier.” This stochasticity enables us to capture uncertainty in our choice of hypothesis.

In this section, we overview some PAC-Bayesian generalization bounds. Now, it should be noted that the PAC-Bayes literature is rich and varied, and we will only cover some of the main developments herein. In particular, we will highlight how information-theoretic and PAC-Bayesian generalization bounds are closely related via similarities in the derivations and interpretation of the results. For a complementary overview, with additional bounds, details, and historical comments, the reader is encouraged to consult the excellent introduction to PAC-Bayes by Alquier (2024), along with the shorter primer by Guedj (2019).

Throughout, to make the notation more compact, we will use the following shorthands: the PAC-Bayesian population loss, averaged over the randomness of the learning algorithm when trained on the training set $\boldsymbol{Z}$, is denoted by $L(\boldsymbol{Z}) = \mathbb{E}_{P_{W|\boldsymbol{Z}}}[L_{P_Z}(W)]$. Similarly, the PAC-Bayesian training loss is denoted by $\hat{L}(\boldsymbol{Z}) = \mathbb{E}_{P_{W|\boldsymbol{Z}}}[L_{\boldsymbol{Z}}(W)]$. The difference between these, that is, the PAC-Bayesian generalization gap, is denoted as $\overline{\operatorname{gen}}(\boldsymbol{Z}) = L(\boldsymbol{Z}) - \hat{L}(\boldsymbol{Z})$.

5.2.1 Bounds via the Donsker-Varadhan Variational Representation

To begin, we derive a generic PAC-Bayesian bound, analogous to Proposition 4.1, given in terms of a function $f(\cdot,\cdot)$ to be specified later. Similar results are discussed by, *e.g.*, Alquier and Guedj (2018), Bégin *et al.* (2014), Germain *et al.* (2009a), and Rivasplata *et al.* (2020).

Proposition 5.2. Assume that almost surely under $P_{\boldsymbol{Z}}$, $f : \mathcal{W} \times \mathcal{Z}^n \to \mathbb{R}$ is a function satisfying $\mathbb{E}_{P_{W|\boldsymbol{Z}}}[f(W, \boldsymbol{Z})] < \infty$ and $P_{W|\boldsymbol{Z}} \ll Q_{W|\boldsymbol{Z}}$. Then, we have that with probability at least $1 - \delta$ under $P_{\boldsymbol{Z}}$,

$$\mathbb{E}_{P_{W|\boldsymbol{Z}}}[f(W, \boldsymbol{Z})] \le \log \mathbb{E}_{Q_{W\boldsymbol{Z}}}\left[\frac{e^{f(W,\boldsymbol{Z})}}{\delta}\right] + D(P_{W|\boldsymbol{Z}} \,||\, Q_{W|\boldsymbol{Z}}). \tag{5.7}$$

Proof. By applying the Donsker-Varadhan variational representation of the relative entropy, we can conclude that almost surely,

$$\mathbb{E}_{P_{W|\boldsymbol{Z}}}[f(W, \boldsymbol{Z})] \le \log \mathbb{E}_{Q_{W|\boldsymbol{Z}}}\left[e^{f(W,\boldsymbol{Z})}\right] + D(P_{W|\boldsymbol{Z}} \,||\, Q_{W|\boldsymbol{Z}}). \tag{5.8}$$

The result follows by noting that Markov's inequality implies that, with probability at least $1-\delta$ under $P_{\boldsymbol{Z}}$,

$$\mathbb{E}_{Q_{W|\boldsymbol{Z}}}\left[e^{f(W,\boldsymbol{Z})}\right] \leq \mathbb{E}_{Q_{W\boldsymbol{Z}}}\left[\frac{e^{f(W,\boldsymbol{Z})}}{\delta}\right]. \tag{5.9}$$

□

In the PAC-Bayesian vernacular, the distribution $P_{W|\boldsymbol{Z}}$ is referred to as a *posterior* while the distribution $Q_{W|\boldsymbol{Z}}$ is called a *prior*, in line with the historical connection with Bayesian inference. However, we once again emphasize that these distributions are not required to have any relation to actual Bayesian priors and posteriors. We only require that the prior is selected so that the first term in the right-hand side of (5.7) can be controlled and that the posterior and prior satisfy the absolute continuity criterion. Typically, the prior $Q_{W|\boldsymbol{Z}}$ is selected to be independent of the training data $\boldsymbol{Z}$. However, as highlighted by, for instance, Rivasplata *et al.* (2020), this is not technically required, although often convenient.

Clearly, the result in Proposition 5.2 is very similar to the one in Proposition 4.1, and in fact, the two results are connected through an exponential stochastic inequality. Predictably, we can therefore derive a result very similar to Corollary 4.2 for sub-Gaussian losses. However, some care has to be taken. In the derivation of Corollary 4.2, we set $f(W,\boldsymbol{Z}) = \lambda \text{gen}(W,\boldsymbol{Z})$ and applied the concentration result from (3.46), after which we optimized the parameter λ. Such an optimization cannot be performed for Proposition 5.2, since the bound therein holds with probability $1-\delta$ for a *fixed* function $f(W,\boldsymbol{Z})$, and hence, a fixed λ if we set $f(W,\boldsymbol{Z}) = \lambda \text{gen}(W,\boldsymbol{Z})$. Thus, to use a similar approach here, we would need to use some sort of union bound over the set of candidate values for λ. Now, while this can be done—as can be seen, for instance, in Section 9.4.1 and the work of Catoni (2007), Rodríguez-Gálvez *et al.* (2023), and Seldin *et al.* (2012b)—we will take an alternative approach here, and use the bound on the moment-generating function of the square of sub-Gaussian random variables from Proposition 3.26.

Corollary 5.3. Assume that $\ell(w, Z)$ is σ-sub-Gaussian under P_Z for all $w \in \mathcal{W}$. Then, with probability at least $1-\delta$ under $P_{\boldsymbol{Z}}$, we have

$$L(\boldsymbol{Z}) \leq \hat{L}(\boldsymbol{Z}) + \sqrt{2\sigma^2 \left(\frac{D(P_{W|\boldsymbol{Z}} \,||\, Q_W) + \log \frac{\sqrt{n}}{\delta}}{n-1} \right)}. \tag{5.10}$$

Proof. First, we use Proposition 5.2 with $Q_{W|\boldsymbol{Z}} = Q_W$ and

$$f = \frac{(n-1)\mathrm{gen}(W, \boldsymbol{Z})^2}{2\sigma^2}. \tag{5.11}$$

The result then follows by applying (3.48) with $\lambda = (n-1)/n$. □

Thanks to the flexibility of the theoretical framework, which allows the prior and posterior to be freely chosen, the relative entropy term can be used in a number of different ways which may be practical for certain applications. For instance, one can choose the prior to have a desirable property—for instance, sparsity (Alquier and Biau, 2013; Guedj and Alquier, 2013)—and select the posterior by minimizing (5.10) directly. This encourages the same desirable property in the posterior, subject to fitting the training data.

For bounded losses, we can derive results with better dependence on the sample size n, which are useful in the regime of small training losses. To do this, we use the techniques from Section 3.3.2. For instance, by setting $f(W, \boldsymbol{Z}) = nd(L_{\boldsymbol{Z}}(W) \,||\, L_{P_Z}(W))$, we can use Theorem 3.29 to bound the second term in the right-hand side of (5.9). This result, first obtained by Maurer (2004), improves on a previous bound due to Langford and Seeger (2001). It is sometimes referred to as the Maurer-Langford-Seeger (MLS) bound.

Corollary 5.4. Assume that the range of the loss function $\ell(\cdot, \cdot)$ is $[0, 1]$. Then, with probability at least $1-\delta$ under $P_{\boldsymbol{Z}}$,

$$d\Big(\hat{L}(\boldsymbol{Z}) \,||\, L(\boldsymbol{Z})\Big) \leq \frac{D(P_{W|\boldsymbol{Z}} \,||\, Q_W) + \log \frac{2\sqrt{n}}{\delta}}{2n}. \tag{5.12}$$

As noted by Langford (2002) (who attributes this observation to Patrick Haffner), this PAC-Bayesian bound has an appealing dimensional consistency as compared to, say, Theorem 2.2. Both sides are given in terms of logarithms of probabilities, *i.e.*, nats.

In order to use Corollary 5.4 to obtain explicit bounds on the population loss, we somehow need to invert the function $d(L_{\boldsymbol{Z}}(W) \,||\, \cdot)$. As discussed after Corollary 4.3, this can be done via the numerical inverse $d^{-1}(p, \varepsilon)$, defined in (3.53). In words, given a training loss $\hat{L}(\boldsymbol{Z})$ and an upper-bound on $d\Big(\hat{L}(\boldsymbol{Z}) \,||\, L(\boldsymbol{Z})\Big)$, this "inverse" of the binary relative entropy gives the highest possible value of $L(\boldsymbol{Z})$ that is consistent with the upper bound and training loss. While it does not admit an analytical expression, it can be found efficiently via numerical search. Analytical relaxations can be obtained either by using Pinsker's inequality (Theorem 3.10) or the more refined bound in Proposition 3.30.

In the PAC-Bayesian literature, a distinction is sometimes made between parametric and non-parametric bounds. The MLS bound in Corollary 5.4, for instance, is an example of a non-parametric bound. It admits a parametric counter-part due to Catoni (2007) and McAllester (2013). Unsurprisingly, this is obtained by using the parametric version of the concentration result for binary relative entropy from Theorem 3.31.

Corollary 5.5. Assume that the range of the loss function $\ell(\cdot, \cdot)$ is $[0, 1]$. Then, with probability at least $1 - \delta$ under $P_{\boldsymbol{Z}}$, for any constant γ,

$$d_\gamma\Big(\hat{L}(\boldsymbol{Z}) \,||\, L(\boldsymbol{Z})\Big) \leq \frac{D(P_{W|\boldsymbol{Z}} \,||\, Q_W) + \log \frac{1}{\delta}}{n}. \tag{5.13}$$

We see that, as compared to the MLS bound, this parametric version saves a $\log(2\sqrt{n})/n$-term. However, this comes at the cost of having to choose the constant γ appropriately (and in a data-independent way). As discussed following Proposition 5.2, we cannot simply optimize over γ, since the bound is probabilistic. Catoni (2007) discusses how to select γ by constructing a dyadic grid of candidate values and optimizing over it, while McAllester (2013) advises a heuristic set of candidates values over which one can optimize.

A relaxation of Corollary 5.5, which more clearly reveals how the parameter γ can be used to control a trade-off between the training loss and the relative entropy, can be obtained as follows (McAllester, 2013).

Corollary 5.6. Assume that the range of the loss function $\ell(\cdot, \cdot)$ is $[0, 1]$.

For any fixed $\lambda > 1$, with probability $1-\delta$ under $P_{\boldsymbol{Z}}$, we have

$$L(\boldsymbol{Z}) \leq \lambda \hat{L}(\boldsymbol{Z}) + \frac{\lambda \left(D(P_{W|\boldsymbol{Z}} \,||\, Q_W) + \log \frac{1}{\delta} \right)}{2n(1-1/\lambda)}. \tag{5.14}$$

Proof. Starting from (5.13) and using (3.51), we obtain

$$\gamma \hat{L}(\boldsymbol{Z}) - \log(1 + (e^\gamma - 1)L(\boldsymbol{Z})) \leq \underbrace{\frac{D(P_{W|\boldsymbol{Z}} \,||\, Q_W) + \log \frac{1}{\delta}}{n}}_{B}. \tag{5.15}$$

Now, assume that $\gamma \in (-2, 0)$. Then, (5.15) implies that

$$L(\boldsymbol{Z}) \leq \frac{1 - \exp(\gamma \hat{L}(\boldsymbol{Z}) - B)}{1 - e^\gamma}. \tag{5.16}$$

When $\gamma \in (-2, 0)$, we have $e^\gamma \geq 1 + \gamma$ and $e^\gamma \leq 1 + \gamma + \gamma^2/2$, so that

$$L(\boldsymbol{Z}) \leq \frac{1 - \exp(\gamma \hat{L}(\boldsymbol{Z}) - B)}{1 - e^\gamma} \leq \frac{\hat{L}(\boldsymbol{Z}) - B/\gamma}{1 + \gamma/2}. \tag{5.17}$$

Finally, let $\lambda = 1/(1+\gamma/2)$, and note that $\gamma \in (-2, 0)$ implies $\lambda > 1$, from which the result follows. □

In Section 4.2, we introduced the individual-sample technique for generalization bounds in expectation. Given any bound on the average population loss depending on the joint distribution of the hypothesis and the training set $P_{W\boldsymbol{Z}}$, this technique allowed us to obtain a bound depending on the joint distribution of the hypothesis and each individual sample, P_{WZ_i}. In most cases, this allowed us to obtain tighter bounds in expectation, sometimes enabling us to turn a vacuous bound into a nonvacuous one. A natural question is then the following: can we similarly derive PAC-Bayesian individual-sample bounds? Unfortunately, the answer is, in general, negative. As shown by Harutyunyan *et al.* (2022), there exists a counter-example for which W is independent of each Z_i, but where the PAC-Bayesian generalization gap is high with non-negligible probability. It is, however, possible to derive such bounds based on subsets of size $m \geq 2$.

5.2.2 PAC-Bayesian Bounds Beyond the Relative Entropy

The bounds that we discussed so far are all based on the Donsker-Varadhan variational representation of the relative entropy. Similar to the average bound case, PAC-Bayesian bounds in terms of other information measures have also been considered. Notably, these bounds often allow for heavy-tailed losses. We will give a brief exposition of two approaches, one based on Hölder's inequality and the other based on the variational representation of f-divergences. Additional works on this topic are mentioned in the bibliographic remarks of Section 5.5, and a more detailed discussion can be found in Alquier (2024, Chapter 5).

The basic idea behind using Hölder's inequality to obtain generalization bounds is as follows. First, we consider an expectation of a quantity of interest related to generalization under the true, joint distribution. By using the Radon-Nikodym theorem (Theorem 3.16), this can be turned into an expectation under an auxiliary distribution, at the cost of a Radon-Nikodym derivative appearing. Finally, Hölder's inequality can be used to disentangle the Radon-Nikodym derivative and the generalization quantity, which can then be handled separately. This was done for bounded losses by Bégin *et al.* (2016), and extended to unbounded losses by Alquier and Guedj (2018). We present the result of Alquier and Guedj (2018, Thm. 1) below.

Theorem 5.7. Let $\boldsymbol{Z}$ denote a training set, the samples of which are allowed to be dependent and drawn from different distributions. Furthermore, let $L_{P_{\boldsymbol{Z}}}(W) = \mathbb{E}_{P_{\boldsymbol{Z}}}[L_{\boldsymbol{Z}}(W)]$.[1] For some $p > 1$, we let $q = p/(p-1)$, and set $f_\alpha(x) = x^\alpha$. Assume that $P_{W|\boldsymbol{Z}} \ll Q_W$ almost surely. Then, with probability at least $1-\delta$ under $P_{\boldsymbol{Z}}$,

$$|\overline{\text{gen}}(\boldsymbol{Z})| \leq \left(\frac{\mathbb{E}_{Q_W P_{\boldsymbol{Z}}}[|L_{P_Z}(W) - L_{\boldsymbol{Z}}(W)|^q]}{\delta} \right)^{1/q} \times (D_{f_p - 1}(P_{W|\boldsymbol{Z}} \,||\, Q_W) + 1)^{1/p}. \quad (5.18)$$

Proof. Let $\Delta(W, \boldsymbol{Z}) = |L_{P_Z}(W) - L_{\boldsymbol{Z}}(W)|$. Then, by Jensen's inequal-

[1] While this reduces to the previously defined population loss $L_{P_Z}(W)$ for i.i.d. data, this does not hold in general.

ity and the Radon-Nikodym theorem (Theorem 3.16),

$$|\overline{\mathrm{gen}}(\boldsymbol{Z})| \leq \mathbb{E}_{P_{W|\boldsymbol{Z}}}[\Delta(W,\boldsymbol{Z})] = \mathbb{E}_{Q_W}\left[\Delta(W,\boldsymbol{Z})\frac{\mathrm{d}P_{W|\boldsymbol{Z}}}{\mathrm{d}Q_W}\right]. \tag{5.19}$$

Then, by Hölder's inequality (Theorem 3.22),

$$\mathbb{E}_{Q_W}\left[\Delta(W,\boldsymbol{Z})\frac{\mathrm{d}P_{W|\boldsymbol{Z}}}{\mathrm{d}Q_W}\right] \leq \mathbb{E}^{1/q}_{Q_W}[\Delta(W,\boldsymbol{Z})^q]\,\mathbb{E}^{1/p}_{Q_W}\left[\left(\frac{\mathrm{d}P_{W|\boldsymbol{Z}}}{\mathrm{d}Q_W}\right)^p\right]. \tag{5.20}$$

Finally, it follows from Markov's inequality that with probability at least $1-\delta$,

$$\mathbb{E}^{1/q}_{Q_W}[\Delta(W,\boldsymbol{Z})^q] \leq \mathbb{E}^{1/q}_{Q_W}\left[\mathbb{E}_{P_{\boldsymbol{Z}}}\left[\frac{\Delta(W,\boldsymbol{Z})^q}{\delta}\right]\right]. \tag{5.21}$$

Note that $\mathbb{E}_{Q_W}\left[\left(\frac{\mathrm{d}P_{W|\boldsymbol{Z}}}{\mathrm{d}Q_W}\right)^p\right] = D_{f_p-1}(P_{W|\boldsymbol{Z}}\,||\,Q_W)+1$. Thus, the desired result follows after combining the steps above. □

As shown by Alquier and Guedj (2018) and Bégin *et al.* (2016), this bound can be specialized to settings such as i.i.d. data with bounded variance, and even auto-regressive data with finite moments. One drawback is the linear dependence on the inverse confidence parameter $1/\delta$, in contrast to the more benign logarithmic dependence of the previous bounds in this chapter.

An alternative route can be taken based on the unconstrained (Theorem 3.19) or constrained (Theorem 3.20) variational representations for f-divergences. This was done by Ohnishi and Honorio (2021), who derived explicit bounds in terms of a whole host of divergences under various assumptions. For instance, they obtained tighter versions of some bounds from Alquier and Guedj (2018) for heavy-tailed losses. To illustrate the benefit of the constrained representation of Theorem 3.20, one can compare the two results in Lemma 2 and 3 of Ohnishi and Honorio (2021). Using the unconstrained representation in Theorem 3.19, Ohnishi and Honorio (2021) obtain the change of measure result

$$\mathbb{E}_P[\phi] \leq \chi^2(P\,||\,Q) + \mathbb{E}_Q[\phi] + \frac{1}{4}\mathbb{E}_Q\left[\phi^2\right], \tag{5.22}$$

where $\chi^2(P\,||\,Q) = \mathbb{E}_Q\left[(\frac{\mathrm{d}P}{\mathrm{d}Q}-1)^2\right]$ is the χ^2 divergence, which can be expressed as an f-divergence (see Definition 3.9) by setting $f(x) = (\sqrt{x}-$

$1)^2$. In contrast, using the constrained representation in Theorem 3.20, this can be improved to

$$\mathbb{E}_P[\phi] \leq \chi^2(P \,||\, Q) + \mathbb{E}_Q[\phi] + \frac{1}{4}\sqrt{\mathbb{E}_Q[(\phi - \mathbb{E}_Q[\phi])^2]}. \tag{5.23}$$

Since the variance is upper-bounded by the second moment, this is always tighter.

Below, we state a bound in terms of the Rényi divergence for sub-Gaussian losses from Ohnishi and Honorio (2021, Prop. 6), to illustrate that the variational representation for f-divergences allows for the derivation of bounds with a more benign logarithmic dependence on $1/\delta$.

Theorem 5.8. Assume that the loss function $\ell(w, Z)$ is σ-sub-Gaussian under P_Z for all $w \in \mathcal{W}$. Fix $\alpha > 1$. Then, with probability at least $1-\delta$,

$$\overline{\text{gen}}(\boldsymbol{Z}) \leq \sqrt{\frac{2\sigma^2}{m}\log\left(\frac{2}{\delta}\right)\left(\alpha(\alpha-1)D_\alpha(P_{W|\boldsymbol{Z}} \,||\, Q_W)\right)^{1/\alpha}}. \tag{5.24}$$

This result is obtained by specializing Theorem 3.19 to the Rényi divergence, from which one obtains (Ohnishi and Honorio, 2021, Lemma 5)

$$\mathbb{E}_P[\phi] \;\leq\; D_\alpha(P \,||\, Q) \;+\; \frac{(\alpha-1)^{\frac{\alpha}{\alpha-1}}}{\alpha}\mathbb{E}_Q\left[\phi^{\frac{\alpha}{\alpha-1}}\right] \;+\; \frac{1}{\alpha(\alpha-1)}. \tag{5.25}$$

The remaining steps follow after setting $P = P_{W|\boldsymbol{Z}}$, $Q = Q_W$, $\phi = \lambda\text{gen}(w, \boldsymbol{z})$, and using a sub-Gaussian concentration argument (Definition 3.23).

5.2.3 Data-Dependent Priors

So far, we have only covered generalization bounds with data-independent priors. However, as we have indicated when discussing Proposition 5.2, data-dependent priors can also be considered. There are several ways of obtaining generalization bounds with data-dependent priors. The approach that is perhaps simplest, but which can lead to very tight bounds in practice, is the data-splitting technique (Ambroladze *et al.*, 2006; Dziugaite *et al.*, 2021). It does not actually involve any new tool—we simply need to apply the tools we have already introduced in a slightly different way. The idea is to split the training set into two parts as $\boldsymbol{Z} = (\boldsymbol{Z}_B, \boldsymbol{Z}_P)$, where $|\boldsymbol{Z}_B| = m$ and $|\boldsymbol{Z}_P| = n - m$. Then, the training loss in the bound is evaluated only on $\boldsymbol{Z}_B$, which means that we are free to use $\boldsymbol{Z}_P$ to inform our prior. To be clear: the full training data $\boldsymbol{Z}$ is still used as input to the *posterior* (*i.e.*, the learning algorithm). The only difference is in how we evaluate the generalization bound itself; the learning procedure remains the same. Since the prior, which can now be written $Q_{W|\boldsymbol{Z}_P}$, is independent of the data $\boldsymbol{Z}_B$ used to compute the training loss in the bound, we can apply the same concentration arguments as in the case of a data-free prior.

As an example, we apply this approach to the bound in Corollary 5.4. Note that it can be applied to all other PAC-Bayesian bounds reviewed in this chapter (and in fact, also to the average generalization bounds discussed in Chapter 4).

Corollary 5.9. With probability at least $1 - \delta$ under $P_{\boldsymbol{Z}}$,

$$d\Big(\mathbb{E}_{P_{W|\boldsymbol{Z}}}[L_{\boldsymbol{Z}_B}(W)] \,||\, L(\boldsymbol{Z})\Big) \leq \frac{D(P_{W|\boldsymbol{Z}} \,||\, Q_{W|\boldsymbol{Z}_P}) + \log \frac{2\sqrt{m}}{\delta}}{2m}. \quad (5.26)$$

Here, a trade-off emerges between two factors that affect the tightness of the bound. Evaluating the training loss based only on the m samples in $\boldsymbol{Z}_B$ means that we divide the right-hand side by a smaller overall factor. However, this is compensated by the fact that the relative entropy term can be significantly lower, since there may exist a posterior with low training loss that is close (in terms of relative entropy) to a suitably chosen data-dependent prior.

Data-dependent priors with the data-splitting technique can be connected to a class of learning algorithms called *compression schemes*. We discuss this more in Section 6.3. Furthermore, data-dependent priors have been used to obtain numerically accurate generalization bounds for neural networks. We cover this in more detail in Section 8.2.

We conclude by noting that there are other ways to obtain data-dependent priors—for instance, through differential privacy (Dziugaite and Roy, 2018b) or algorithmic stability (Rivasplata *et al.*, 2018).

5.3 Single-Draw Generalization Bounds

The PAC-Bayesian bounds in Section 5.2 apply to losses that are averaged over the posterior, $P_{W|\boldsymbol{Z}}$. In practice, it is common to instead use a randomized learning algorithm to select a single hypothesis, and then use this specific instance of W for future inference. In the PAC-Bayesian literature, bounds for this scenario are often termed *de-randomized* or *pointwise* PAC-Bayes bounds. We will refer to this scenario, and bounds that apply for it, as *single-draw*, following the terminology of Catoni (2007): the bounds apply to a single draw of the training data and a single draw from the stochastic learning algorithm. In this section, we present several such single-draw bounds.

5.3.1 Bounds via Variational Representations of Divergences

As we did for both the average case in Chapter 4 and the PAC-Bayesian setting in Section 5.2, we begin by deriving a generic single-draw inequality for a function $f(\cdot,\cdot)$ to be specified later. However, this will require slightly stronger absolute continuity assumptions than we had before.

Proposition 5.10. Assume that $P_{W\boldsymbol{Z}} \ll Q_{W\boldsymbol{Z}}$ and $Q_{W\boldsymbol{Z}} \ll P_{W\boldsymbol{Z}}$. For any function $f(\cdot,\cdot)$, with probability at least $1-\delta$ under $P_{W\boldsymbol{Z}}$,

$$f(W,\boldsymbol{Z}) \le \log \mathbb{E}_{Q_{W\boldsymbol{Z}}}\left[\frac{e^{f(W,\boldsymbol{Z})}}{\delta}\right] + \log \frac{\mathrm{d}P_{W\boldsymbol{Z}}}{\mathrm{d}Q_{W\boldsymbol{Z}}}. \tag{5.27}$$

Proof. From Polyanskiy and Wu (2022, Proposition 18.3), we have

$$\mathbb{E}_{Q_{W\boldsymbol{Z}}}\left[e^{f(W,\boldsymbol{Z})}\right] = \mathbb{E}_{P_{W\boldsymbol{Z}}}\left[\exp\left(f(W,\boldsymbol{Z}) - \log \frac{\mathrm{d}P_{W\boldsymbol{Z}}}{\mathrm{d}Q_{W\boldsymbol{Z}}}\right)\right]. \tag{5.28}$$

Rewriting this, we obtain

$$\mathbb{E}_{P_{W\boldsymbol{Z}}}\left[\exp\left(f(W,\boldsymbol{Z}) - \log \mathbb{E}_{Q_{W\boldsymbol{Z}}}\left[e^{f(W,\boldsymbol{Z})}\right] - \log \frac{\mathrm{d}P_{W\boldsymbol{Z}}}{\mathrm{d}Q_{W\boldsymbol{Z}}}\right)\right] = 1. \tag{5.29}$$

Applying Markov's inequality (in the same way as in (5.5)) to (5.29), we conclude that with probability at least $1-\delta$ under $P_{W\boldsymbol{Z}}$,

$$\exp\left(f(W,\boldsymbol{Z}) - \log \mathbb{E}_{Q_{W\boldsymbol{Z}}}\left[e^{f(W,\boldsymbol{Z})}\right] - \log \frac{\mathrm{d}P_{W\boldsymbol{Z}}}{\mathrm{d}Q_{W\boldsymbol{Z}}}\right) \le \frac{1}{\delta}. \tag{5.30}$$

The result follows by taking the logarithm and rearranging terms. □

As previously mentioned, for the specific choice of $Q_{W\boldsymbol{Z}} = P_{W\boldsymbol{Z}}$, the logarithm of the Radon-Nikodym derivative reduces to the information density $\imath(W, \boldsymbol{Z})$.

By making the same specific choices for $f(W, \boldsymbol{Z})$ and using the same sub-Gaussian concentration inequality as in Corollary 5.3, we can obtain the following analogous single-draw generalization bound.

Corollary 5.11. With probability at least $1 - \delta$ under $P_{W\boldsymbol{Z}}$, we have

$$L_{P_Z}(W) \le L_{\boldsymbol{Z}}(W) + \sqrt{2\sigma^2 \left(\frac{\log \frac{\mathrm{d}P_{W\boldsymbol{Z}}}{\mathrm{d}Q_W P_{\boldsymbol{Z}}} + \log \frac{\sqrt{n}}{\delta}}{n-1} \right)}. \tag{5.31}$$

In this bound, the population loss, training loss, and information metric $\log \frac{\mathrm{d}P_{W\boldsymbol{Z}}}{\mathrm{d}Q_W P_{\boldsymbol{Z}}}$ all depend on the specific instances of W and $\boldsymbol{Z}$. One benefit of this is that the bound is fully empirical: all quantities that appear in the bound can be computed given the training data and hypothesis. Another benefit of the pointwise information measure $\log \frac{\mathrm{d}P_{W\boldsymbol{Z}}}{\mathrm{d}Q_W P_{\boldsymbol{Z}}}$ is that it can be evaluated in closed form for a wider class of distributions than, say, the relative entropy. For example, as long as the distributions $P_{W\boldsymbol{Z}}$ and $Q_W P_{\boldsymbol{Z}}$ have densities, the Radon-Nikodym derivative can easily be evaluated in closed form. In contrast, the relative entropy has a closed form only for a limited number of probability distributions.

All PAC-Bayesian bounds that are derived through an exponential stochastic inequality approach admit a single-draw counterpart—provided that the more stringent absolute continuity criterion in Proposition 5.10 is satisfied. Hence, we can obtain single-draw variants of Corollaries 5.4 to 5.9, with the PAC-Bayesian losses replaced by their single-draw counterparts and with the relative entropy replaced by the logarithm of the corresponding Radon-Nikodym derivative.

5.3.2 Using Hölder's Inequality

An alternative way to obtain single-draw generalization bounds is through the use of Hölder's inequality (Theorem 3.22), via an approach introduced by Esposito *et al.* (2021a). We start by providing the following general theorem.

Theorem 5.12. Assume that $P_{W\boldsymbol{Z}} \ll P_W P_{\boldsymbol{Z}}$. For any constants $\alpha, \alpha', \gamma, \gamma'$ such that $1/\alpha + 1/\gamma = 1/\alpha' + 1/\gamma' = 1$ and all measurable sets $\mathcal{E} \in \mathcal{W} \times \mathcal{Z}^n$, we have

$$P_{W\boldsymbol{Z}}[\mathcal{E}] \leq \mathbb{E}_{P_W}^{1/\gamma'}\left[P_{\boldsymbol{Z}}(\mathcal{E}_W)^{\gamma'/\gamma}\right] \mathbb{E}_{P_W}^{1/\alpha'}\left[\mathbb{E}_{P_{\boldsymbol{Z}}}^{\alpha'/\alpha}\left[\left(\frac{\mathrm{d}P_{W\boldsymbol{Z}}}{\mathrm{d}P_W P_{\boldsymbol{Z}}}\right)^{\alpha}\right]\right]. \tag{5.32}$$

Here, $\mathcal{E}_W = \{\boldsymbol{Z} : (\boldsymbol{Z}, W) \in \mathcal{E}\}$.

Proof. Let $1_{\mathcal{E}}$ denote the indicator function of $\mathcal{E}$. By the Radon-Nikodym theorem we have

$$\begin{aligned} P_{W\boldsymbol{Z}}[\mathcal{E}] &= \mathbb{E}_{P_{W\boldsymbol{Z}}}[1_{\mathcal{E}}] & (5.33)\\ &= \mathbb{E}_{P_W P_{\boldsymbol{Z}}}\left[1_{\mathcal{E}}\frac{\mathrm{d}P_{W\boldsymbol{Z}}}{\mathrm{d}P_W P_{\boldsymbol{Z}}}\right] & (5.34)\\ &= \mathbb{E}_{P_W}\left[\mathbb{E}_{P_{\boldsymbol{Z}}}\left[1_{\mathcal{E}_W}\frac{\mathrm{d}P_{W\boldsymbol{Z}}}{\mathrm{d}P_W P_{\boldsymbol{Z}}}\right]\right]. & (5.35) \end{aligned}$$

By applying Hölder's inequality twice, we get

$$\begin{aligned} P_{W\boldsymbol{Z}}[\mathcal{E}] &\leq \mathbb{E}_{P_W}\left[\mathbb{E}_{P_{\boldsymbol{Z}}}^{1/\gamma}\left[1_{\mathcal{E}_W}^{\gamma}\right]\mathbb{E}_{P_{\boldsymbol{Z}}}^{1/\alpha}\left[\left(\frac{\mathrm{d}P_{W\boldsymbol{Z}}}{\mathrm{d}P_W P_{\boldsymbol{Z}}}\right)^{\alpha}\right]\right] & (5.36)\\ &\leq \mathbb{E}_{P_W}^{1/\gamma'}\left[\mathbb{E}_{P_{\boldsymbol{Z}}}^{\gamma'/\gamma}[1_{\mathcal{E}_W}]\right]\mathbb{E}_{P_W}^{1/\alpha'}\left[\mathbb{E}_{P_{\boldsymbol{Z}}}^{\alpha'/\alpha}\left[\left(\frac{\mathrm{d}P_{W\boldsymbol{Z}}}{\mathrm{d}P_W P_{\boldsymbol{Z}}}\right)^{\alpha}\right]\right] & (5.37) \end{aligned}$$

from which the result follows. □

By choosing $\mathcal{E}$ and the parameters in Theorem 5.12 appropriately, we can derive many generalization bounds. We will focus on a bound for sub-Gaussian losses which is expressed in terms of the α-mutual information (see Definition 3.14).

Corollary 5.13. Assume that the loss function $\ell(w, Z)$ is σ-sub-Gaussian under P_Z for all $w \in \mathcal{W}$. Furthermore, assume that $P_{W\boldsymbol{Z}} \ll P_W P_{\boldsymbol{Z}}$. Then, with probability at least $1-\delta$ under $P_{W\boldsymbol{Z}}$, for any $\alpha > 1$,

$$|\mathrm{gen}(W, \boldsymbol{Z})| \leq \sqrt{\frac{2\sigma^2}{n}\left(I_\alpha(W; \boldsymbol{Z}) + \log 2 + \frac{\alpha}{\alpha-1}\log\frac{1}{\delta}\right)}. \tag{5.38}$$

In particular, when $\alpha \to \infty$,

$$|\mathrm{gen}(W, \boldsymbol{Z})| \leq \sqrt{\frac{2\sigma^2}{n}\left(\mathcal{L}(\boldsymbol{Z} \to W) + \log\frac{2}{\delta}\right)}. \tag{5.39}$$

Proof. Let $\alpha' \to 1$, implying that $\gamma' \to \infty$. Consider the error event $\mathcal{E} = \{W, \boldsymbol{Z} : |L_{P_Z}(W) - L_{\boldsymbol{Z}}(W)| > \varepsilon\}$. By sub-Gaussianity (see Theorem 3.25), for each $w \in \mathcal{W}$ we have

$$P_{\boldsymbol{Z}}[\mathcal{E}_w] \leq 2\exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right). \tag{5.40}$$

Furthermore, note that

$$\mathbb{E}_{P_W}\left[\mathbb{E}_{P_{\boldsymbol{Z}}}^{1/\alpha}\left[\frac{\mathrm{d}P_{W\boldsymbol{Z}}}{\mathrm{d}P_W P_{\boldsymbol{Z}}}\right]\right] = \exp\left(\frac{\alpha-1}{\alpha} I_\alpha(W;\boldsymbol{Z})\right). \tag{5.41}$$

By setting $\delta = P_{W\boldsymbol{Z}}[\mathcal{E}]$ and solving for ε, we obtain the desired result. □

Notice that, unlike the single-draw bounds we presented previously, the right-hand side of this bound is a constant: it does not depend on the specific instances of W and $\boldsymbol{Z}$ in the left-hand side. A drawback of this is that the bound is no longer empirical, and requires knowledge of P_Z to be computed exactly. An advantage is that the same bound holds regardless of the specific data or output of the algorithm, and the bound can thus be stated *a priori*.

5.4 Mean-Hypothesis Generalization Bounds

Before concluding this chapter, we will briefly mention a third flavor of generalization bounds in probability: bounds for the mean hypothesis. To this end, we consider a stochastic learning algorithm $P_{W|\boldsymbol{Z}}$ with a fixed, randomly drawn $\boldsymbol{Z}$. There are many ways to define deterministic classifiers based on this, by averaging over the randomness of the learning algorithm in various ways. For instance, in Banerjee and Montufar (2021), the goal is to bound, with high probability over $P_{\boldsymbol{Z}}$, the population loss of the average hypothesis output by the learning algorithm: $w^* = \mathbb{E}_{P_{W|\boldsymbol{Z}}}[W]$. The motivation for this is that, when evaluating PAC-Bayesian generalization bounds, it is common to start from a deterministic learning algorithm and make it stochastic by adding zero-mean Gaussian noise to the parameters. We will see this in more detail in Chapter 8. While this can allow the bounds to be computed, the new classifier with added noise often has degraded performance relative to the underlying deterministic one. However, since the mean of this randomized classifier is the underlying, original hypothesis, bounds on w^* apply to this deterministic classifier.

For binary classifiers, we can obtain various type of majority voting algorithms on the basis of the posterior. Following the discussion from Seeger (2002), we assume that W denotes the parameters of a map $f_W : \mathcal{X} \to \mathbb{R}$. The goal is to predict the binary label of X, and we use the sign of $f_W(X)$ to achieve this. For the stochastic predictors we considered previously, the output is given by $\text{sign}(f_W(X))$, where $W \sim P_{W|\boldsymbol{Z}}$. However, we can also consider the majority vote classifier given by $f_{\text{mv}(X)} = \text{sign}\left(\mathbb{E}_{P_{W|\boldsymbol{Z}}}[\text{sign}(f_W(X))]\right)$, as well as the averaging classifier, given by $f_{\text{BPM}(X)} = \text{sign}\left(\mathbb{E}_{P_{W|\boldsymbol{Z}}}[f_W(X)]\right)$. These are referred to as the Bayes voting classifier and the Bayes classifier, respectively, by Seeger (2002). As noted by Langford and Shawe-Taylor (2002), a bound on the population loss of the stochastic predictor based on $P_{W|\boldsymbol{Z}}$ leads to a bound on the population loss of the BPM classifier, at the price of a factor 2. This is based on the observation that, for any sample for which f_{BPM} incurs a loss, the underlying stochastic classifier must incur a loss with probability at least $1/2$.

5.5 Bibliographic Remarks and Additional Perspectives

In this section, we discuss how the presented results relate to the literature, and briefly mention some results that we did not cover in detail. As the literature is vast, in particular for PAC-Bayesian bounds, this brief overview will not be exhaustive. See Section 2.4 for further discussion of the early history of PAC-Bayesian bounds, and the monograph of Alquier (2024).

The underlying concepts of the exponential stochastic inequality framework of Section 5.1 can be traced to the works of Zhang (2006) and Catoni (2007). This was formalized using ESI notation by Grünwald and Mehta (2020), Koolen *et al.* (2016), and Mhammedi *et al.* (2019), and was recently given an exhaustive treatment by Grünwald *et al.* (2023).

The generic PAC-Bayesian bound in Proposition 5.2 is similar to statements given by Bégin *et al.* (2014) and Germain *et al.* (2009a), while this exact form is due to Rivasplata *et al.* (2020). While Corollary 5.3 is very similar to earlier results, such as the one from McAllester (2003a), this exact form is from Hellström and Durisi (2020a). The bound in Corollary 5.4 is due to Maurer (2004), where the logarithmic factor is improved compared to the result of Langford and Seeger (2001). See the work of Foong *et al.* (2021) for an in-depth discussion of the tightness of this bound and whether this logarithmic dependence can be improved further. Corollary 5.5 is implicit in the work of McAllester (2013) (who in turn describes the result as a corollary of statements from Catoni (2007)), while the loosened version in Corollary 5.6 is stated explicitly. The role of the generic convex function in the bound is studied in (Hellström and Guedj, 2024), where optimal choices are established. Jang *et al.*, 2023 used the coin-betting framework from online learning to improve PAC-Bayesian bounds for bounded losses.

Theorem 5.7 is due to Alquier and Guedj (2018), and is an extension of a result from Bégin *et al.* (2016) for bounded losses. Ohnishi and Honorio (2021) provided a comprehensive treatment of change of measure inequalities with f-divergences and their application in PAC-Bayesian bounds, including Theorem 5.8, as well as a whole host of additional results.

Data-dependent priors based on data splitting were introduced by Ambroladze *et al.* (2006), and have since been extended and used in various ways by, for instance, Dziugaite *et al.* (2021), Dziugaite and Roy (2018b), Mhammedi *et al.* (2019), Parrado-Hernández *et al.* (2012), and Rivasplata *et al.* (2020, 2018). Seeger (2002) used a similar technique, whereby an independent set of “model selection” samples is used to learn the prior and the model class. However, unlike in the works mentioned above, this set is disjoint from the training set used to find the posterior. Data-dependent priors through differential privacy were studied by Dziugaite and Roy (2018b), while Rivasplata *et al.* (2020) used algorithmic stability. Distribution-dependent priors are discussed by, *e.g.*, Catoni (2007) and Lever *et al.* (2010, 2013). We discuss data-dependent priors further in Sections 6.3 and 8.2.

While mainly focusing on PAC-Bayesian bounds, Catoni (2007, Thm. 1.2.7) mentioned in passing that similar techniques can be used to obtain bounds for single draws from the posterior, which is the basis for our terminology of "single-draw." The generic inequality in Proposition 5.10 is due to Rivasplata *et al.* (2020), while Corollary 5.11 can be found in Hellström and Durisi (2020a). Explicit derivations of more single-draw generalization bounds can be found in Hellström and Durisi (2021a,b).

Theorem 5.12 and Corollary 5.13 are due to Esposito *et al.* (2021a), who also presented several additional bounds and results beyond learning theory. In (Hellström and Durisi, 2020a), the "strong converse" lemma from binary hypothesis testing is used to obtain single-draw bounds in terms of the tail of the information density. Xu and Raginsky (2017) adapted the monitor technique from Bassily *et al.* (2016) to convert their average generalization bound to a single-draw one, albeit with a detrimental linear dependence on the inverse confidence parameter $1/\delta$.

Langford and Shawe-Taylor (2002) pointed out that certain mean-hypothesis generalization bounds follow immediately from standard PAC-Bayesian bounds, stating that this was essentially folklore, with further discussion in the work of Seeger (2002). PAC-Bayesian bounds for aggregated predictors have been studied by, *e.g.*, Alquier and Biau (2013), Dalalyan and Salmon (2012), Dalalyan and Tsybakov (2007, 2008, 2012), Guedj and Alquier (2013), Leung and Barron (2006), and Salmon and Dalalyan (2011). Further discussion of this can be found in Alquier (2024, Sec. 2.2). Germain *et al.* (2015) introduced the celebrated $\mathcal{C}$-bound, which studies the behavior of majority votes in binary classification. Bounds for voting classifiers are also discussed by Lacasse *et al.* (2006), while Biggs *et al.* (2022) and Zantedeschi *et al.* (2021) consider stochastic majority votes.

Finally, we provide some pointers to results that we did not explicitly cover. As mentioned, a complementary overview of PAC-Bayesian bounds can be found in the introduction by Alquier (2024), as well as the primer by Guedj (2019). Two particularly notable topics that we did not cover are oracle bounds and the localization technique of Catoni (2007). Oracle bounds, also called excess risk bounds, bound the difference between the population loss of the posterior (or hypothesis, depending on the flavor under consideration) and the minimal achievable loss for the given hypothesis class. Such bounds are covered in Alquier (2024, Chapter 4). Some notable works proving oracle bounds are Alquier and

Guedj (2017), Alquier and Lounici (2011), Dalalyan and Salmon (2012), Dalalyan and Tsybakov (2008, 2012), Rigollet and Tsybakov (2012), and Salmon and Dalalyan (2011). The localization technique of Catoni (2007) is a method for selecting a suitable distribution-dependent prior, and is discussed in Alquier (2024, Sec. 4.5).

Tolstikhin and Seldin (2013) used (3.30) to obtain a relaxation of Corollary 5.4 to obtain a bound that interpolates between a fast and slow rate, depending on the value of the training loss, while Thiemann *et al.* (2017) considered a relaxation of Corollary 5.4, and provided a procedure for minimizing it. The connection between PAC-Bayesian bounds and Bayesian inference is discussed by Germain *et al.* (2016a), while the connection to KL-regularized objective functions is covered by Germain *et al.* (2009b). PAC-Bayesian bounds for sub-exponential random variables are discussed by, *e.g.*, Catoni (2004b). Alquier (2006, 2008) used truncated losses in order to handle unbounded loss functions, while Catoni and Giulini (2018) used a robust loss function to handle heavy-tailed distributions. Holland (2019) derived PAC-Bayesian bounds for heavy-tailed losses, obtaining a novel Gibbs posterior on this basis. Biggs and Guedj (2023) obtained tighter bounds based on the excess risk by using the underlying difficulty of the problem. Herbrich and Graepel (2002) and Langford and Shawe-Taylor (2002) derived bounds in terms of the margins of the learned predictor—an approach recently used by Biggs and Guedj (2022b) to establish derandomized generalization bounds. Audibert and Bousquet (2007) combined the chaining technique (discussed in Section 4.4) with PAC-Bayesian bounds. Similarly, Asadi and Abbe (2020) derived bounds based on a multilevel relative entropy, while Clerico *et al.* (2022b) derived an alternative chained bound. Yang *et al.* (2019) derived fast-rate PAC-Bayesian bounds through the use of Rademacher processes. Saunshi *et al.* (2019) derived generalization bounds involving Rademacher complexities for contrastive unsupervised representation learning (CURL), the state-of-the-art technique to learn representations (as a set of features) from unlabelled data. Their results were generalized by Nozawa *et al.* (2020) to obtain PAC-Bayesian generalization bounds for CURL, holding for non-i.i.d. data and allowing for new representation learning algorithms. Mhammedi *et al.* (2020) noted that while many works study bounds for the expected risk, *i.e.*, the mean performance of an algorithm, this might not be the relevant metrics in many problems (*e.g.*, medical, environmental or sensitive engineering tasks). Motivated by this, they presented a PAC-Bayesian generalization bound for the Conditional Value at Risk (CVaR).[2] Chérief-Abdellatif *et al.* (2022) analyzed Variational Auto-Encoders (VAE) (Kingma and Welling, 2019), a popular generative model, through PAC-Bayesian generalization bounds on the reconstruction error of the VAE, and used it to study the regularization effect of classical VAE objectives. Mbacke *et al.* (2023b) provided further PAC-Bayesian bounds for VAEs, while Mbacke *et al.* (2023a) studied adversarial generative models. Haddouche *et al.* (2021) considered losses with a hypothesis-dependent range, and obtained bounds for these through the use of self-bounding functions. Haddouche and Guedj (2023a) developed bounds for heavy-tailed loss functions through the use of supermartingales. Amit *et al.* (2022) derived bounds in terms of *integral probability metrics* (IPM), which includes the total variation and the Wasserstein distance. This is achieved by essentially using the definition of IPMs as a change of measure (which is similar to the Kantorovich-Rubinstein duality). Notably, this can be used to convert uniform convergence bounds, as those discussed in Section 1.3, into algorithm-dependent bounds where the uniform convergence bound is multiplied by a total variation between the posterior and a prior. Recently, Haddouche and Guedj (2023b) and Viallard *et al.* (2023) proposed PAC-Bayesian generalization bounds given in terms of a Wasserstein distance. These bounds hold for unbounded (possibly heavy-tailed) losses, and are used as training objectives.

[2] For any $\alpha \in (0, 1)$ and any random variable Z, $\mathrm{CVaR}_\alpha(Z)$ measures the expectation of Z conditioned on the event that Z is greater than its $(1 - \alpha)$-th quantile. See, for instance, Pflug (2000).

6

The CMI Framework

In previous chapters, the majority of the results that we presented required an absolute continuity assumption to be satisfied. The reason for this requirement is that without it, quantities such as the mutual information in Chapter 4 and the relative entropy in Section 5.2 would be infinite. This absolute continuity requirement is not satisfied when both the training data and the hypothesis are continuous random variables and the hypothesis is a deterministic function of the training data. For average bounds, this issue can be alleviated by the individual-sample technique of Bu *et al.* (2020), as discussed in Section 4.2. However, this approach still yields a vacuous generalization bound when the hypothesis is a deterministic function of a single training sample. So, while the individual-sample technique mitigates the problem, the fundamental issue still remains: the information carried by a single training sample can be infinite.

These considerations motivate the conditional mutual information (CMI) approach, introduced to the literature of information-theoretic generalization bounds by Steinke and Zakynthinou (2020). An essentially equivalent approach was introduced in the PAC-Bayesian context much earlier by Audibert (2004) and Catoni (2007), under the name of “almost exchangeable priors" and "transductive learning," the motivation of which was to reduce the variance of PAC-Bayesian generalization bounds. The terminology used to describe the CMI framework is not uniform—it has also been referred to as the random-subset setting (Hellström and Durisi, 2020b), randomized-subsample setting (Rodríguez-Gálvez *et al.*, 2020), and the supersample setting (Wang and Mao, 2023c). Here, we will stick with the terms "CMI approach" or "CMI framework." These names are motivated by one of the main end-products of the approach: generalization bounds in expectation given in terms of a conditional mutual information. As we will show, many of the techniques covered in the preceding chapters are readily extended to this new setting.

An intuitive view of the CMI framework is that, rather than asking whether one can identify a given training sample based on the chosen hypothesis, we instead ask if, given two candidate samples, we can figure out which one was used for training. Whereas the first question can reveal infinite information, the second one is a *binary* question, so the answer can carry at most 1 bit. From a technical standpoint, this will guarantee that the desired absolute continuity criterion is always satisfied. We now introduce the CMI framework more formally.

6.1 Definition of the CMI Framework

The CMI framework consists of the following elements. First, we assume that we generate a *supersample* $\tilde{\boldsymbol{Z}} = (\tilde{Z}_1, \dots, \tilde{Z}_{2n}) \in \mathcal{Z}^{2n}$ consisting of $2n$ samples drawn i.i.d. from P_Z. Only half of these samples are actually used for training, as determined by a *membership vector* $\boldsymbol{S} \in \{0,1\}^n$, consisting of n Bernoulli-1/2 random variables that are independent of each other and $\tilde{\boldsymbol{Z}}$. Specifically, the ith training sample $Z_i(S_i)$ is given by $\tilde{Z}_{i+S_i n}$, *i.e.*, the Bernoulli-1/2 random variable S_i determines whether $\tilde{Z}_i$ or $\tilde{Z}_{i+n}$ is used for training. Through this procedure, the training set $\boldsymbol{Z}_{\boldsymbol{S}} = (Z_1(S_1), \dots, Z_n(S_n))$ is built, and the hypothesis W is chosen based on this training set. This leads to the Markov chain $(\tilde{\boldsymbol{Z}}, \boldsymbol{S})$—$\boldsymbol{Z}_{\boldsymbol{S}}$—$W$ (*i.e.*, W and $(\tilde{\boldsymbol{Z}}, \boldsymbol{S})$ are conditionally independent given $\boldsymbol{Z}_{\boldsymbol{S}}$). We denote the entry-wise modulo-2 complement of $\boldsymbol{S}$ as $\bar{\boldsymbol{S}}$, *i.e.*, the ith element of $\bar{\boldsymbol{S}}$ is given by $\bar{S}_i = 1 - S_i$. Note that $\boldsymbol{Z}_{\bar{\boldsymbol{S}}}$ is conditionally independent from W given $\boldsymbol{Z}_{\boldsymbol{S}}$, and can hence be considered a test set.

We will use the notation $P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}} = P_{W|\boldsymbol{Z_S}}$ to refer to the conditional distribution on $\mathcal{W}$ that characterizes the learning algorithm and $P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}} = P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}}P_{\tilde{\boldsymbol{Z}}\boldsymbol{S}}$ for the induced joint distribution on W, $\tilde{\boldsymbol{Z}}$ and $\boldsymbol{S}$.

It is important to note that this framework is actually just a reformulation of the standard learning setting from before. Indeed, the training set $\boldsymbol{Z_S}$ still consists of n i.i.d. samples from P_Z, on the basis of which we select W according to our learning algorithm. Here, the supersample $\tilde{\boldsymbol{Z}}$ can be viewed as a “ghost sample,” which is used purely for the purpose of analysis.

The remainder of this chapter is structured as follows. First, we present generalization bounds in expectation using the CMI framework. Then, we review PAC-Bayesian bounds in the CMI framework, with a particular focus on the connection to data-dependent priors, before briefly discussing single-draw bounds. We end the chapter with some extensions of the CMI framework. Specifically, we present bounds in terms of the so-called evaluated and functional CMI, which improve upon the standard CMI bounds due to the data-processing inequality. Finally, we present the leave-one-out setting, where the supersample has size $n+1$ instead of $2n$. This turns out to be closely related to the concept of leave-one-out validation.

6.2 Generalization Bounds in Expectation

We now derive generalization bounds in expectation using the structure of the CMI framework. In order to keep the notation more compact, we will use the following shorthands: the average population loss is $L = \mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}[L_{P_Z}(W)]$, the average training loss is $\hat{L} = \mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}[L_{\boldsymbol{Z_S}}(W)]$, and the average generalization gap is $\overline{\text{gen}} = L - \hat{L}$. When originally introducing the CMI framework, Steinke and Zakynthinou (2020) derived the following bound.

Theorem 6.1. Assume that the range of the loss function $\ell(\cdot,\cdot)$ is $[0,1]$. Then,

$$|\overline{\text{gen}}| \leq \sqrt{\frac{2I(W;\boldsymbol{S}|\tilde{\boldsymbol{Z}})}{n}}. \tag{6.1}$$

Proof. The proof is very similar to that of Corollary 4.2, but with some minor modifications. We begin by noting that, in expectation, the test loss, *i.e.*, the loss evaluated on the test set $\boldsymbol{Z}_{\bar{\boldsymbol{S}}}$, equals the population loss:

$$\mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\Big[L_{\boldsymbol{Z}_{\bar{\boldsymbol{S}}}}(W)\Big] = \mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}[L_{P_Z}(W)] = L. \tag{6.2}$$

Hence, a bound on the average difference between the training and test loss is also a bound on the average generalization gap. To this end, let $\operatorname{gen}(W, \tilde{\boldsymbol{Z}}, \boldsymbol{S}) = L_{\boldsymbol{Z}_{\bar{\boldsymbol{S}}}}(W) - L_{\boldsymbol{Z}_{\boldsymbol{S}}}(W)$. Note that this quantity satisfies the symmetry property $\operatorname{gen}(W, \tilde{\boldsymbol{Z}}, \boldsymbol{S}) = -\operatorname{gen}(W, \tilde{\boldsymbol{Z}}, \bar{\boldsymbol{S}})$. Hence, for any W and $\boldsymbol{Z}$, we have

$$\mathbb{E}_{P_{\boldsymbol{S}}}\Big[\operatorname{gen}(W, \tilde{\boldsymbol{Z}}, \boldsymbol{S})\Big] = \mathbb{E}_{P_{\boldsymbol{S}}}\Big[L_{\boldsymbol{Z}_{\bar{\boldsymbol{S}}}}(W) - L_{\boldsymbol{Z}_{\boldsymbol{S}}}(W)\Big] = 0. \tag{6.3}$$

Furthermore, since $\ell(\cdot,\cdot)$ is bounded to $[0, 1]$, it follows that for each i, the loss difference $\ell(W, Z_i(S_i)) - \ell(W, Z_i(\bar{S}_i))$ is bounded to $[-1, 1]$. Hence, the loss difference is a 1-sub-Gaussian random variable under $P_{\boldsymbol{S}}$ (as well as under $Q_{W\tilde{\boldsymbol{Z}}}P_{\boldsymbol{S}}$ for every distribution $Q_{W\tilde{\boldsymbol{Z}}}$ of $(W, \tilde{\boldsymbol{Z}})$). Since $\operatorname{gen}(W, \tilde{\boldsymbol{Z}}, \boldsymbol{S})$ is an average of n such terms, it is $1/\sqrt{n}$-sub-Gaussian.

Next, by the Donsker-Varadhan variational representation of the relative entropy, we have

$$\lambda\overline{\operatorname{gen}} = \mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\Big[\lambda\operatorname{gen}(W, \tilde{\boldsymbol{Z}}, \boldsymbol{S})\Big] \tag{6.4}$$

$$\leq \log \mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}}P_{\boldsymbol{S}}}\Big[e^{\lambda\operatorname{gen}(W, \tilde{\boldsymbol{Z}}, \boldsymbol{S})}\Big] + D(P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}} \,||\, P_{W\tilde{\boldsymbol{Z}}}P_{\boldsymbol{S}}). \tag{6.5}$$

Note that $D(P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}} \,||\, P_{W\tilde{\boldsymbol{Z}}}P_{\boldsymbol{S}}) = I(W; \boldsymbol{S}|\tilde{\boldsymbol{Z}})$. The rest of the argument follows the same lines as the proof of Corollary 4.2: specifically, we apply the sub-Gaussian concentration inequality and optimize over λ, from which the result follows. □

The benefit of the CMI framework can now be clearly seen. Notice that we did not need to impose any absolute continuity assumption. Since $I(W; \boldsymbol{S}|\tilde{\boldsymbol{Z}}) = D(P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}} \,||\, P_{W\tilde{\boldsymbol{Z}}}P_{\boldsymbol{S}} \,|\, P_{\tilde{\boldsymbol{Z}}\boldsymbol{S}})$, we need $P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}$ to be absolutely continuous with respect to $P_{W\tilde{\boldsymbol{Z}}}P_{\boldsymbol{S}}$. But since $P_{W|\tilde{\boldsymbol{Z}}}$ is obtained by marginalising $P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}}P_{\boldsymbol{S}}$ over the discrete random variable $\boldsymbol{S}$, this is automatically guaranteed. More specifically, we actually have $I(W; \boldsymbol{S}|\tilde{\boldsymbol{Z}}) \leq H(\boldsymbol{S}) = n \log 2$, where $H(\boldsymbol{S})$ is the entropy of $\boldsymbol{S}$ (see Definition 3.3). This confirms the motivation for introducing the framework: the training set, consisting of n samples, cannot carry more information than n bits. However, in the worst case scenario where the trivial upper bound holds with equality—which can occur when $P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}}$ represents a deterministic learning algorithm and gives distinct outputs for each value of $\boldsymbol{S}$—the resulting generalization bound is vacuous, since $\sqrt{2\log 2} > 1$.

The following interesting observation regarding the connection between CMI and mutual information was noted by Haghifam *et al.* (2020). The motivation for having a supersample consisting of $2n$ data samples was to normalize the information carried by each training sample to 1 bit. However, we could have a different scheme where $\tilde{\boldsymbol{Z}}$ consisted of kn samples, for an integer $k > 2$, and instead have S_i uniformly distributed on an index set of size k. While such a construction leads to looser bounds than using $k = 2$, it can be shown that, when the hypothesis space $\mathcal{W}$ is finite, the resulting conditional mutual information $I(W; \boldsymbol{S}|\tilde{\boldsymbol{Z}})$ equals the mutual information $I(W; \boldsymbol{Z}_{\boldsymbol{S}})$ in the limit $k \to \infty$.

We note that the assumption of bounded loss can be somewhat relaxed, as shown in Steinke and Zakynthinou (2020, Thm. 5.1). Specifically, assume that there exists a function $\Delta : \mathcal{Z}^2 \to \mathbb{R}$ such that, for all $z_1, z_2 \in \mathcal{Z}$ and $w \in \mathcal{W}$, we have $|\ell(w, z_1) - \ell(w, z_2)| \leq \Delta(z_1, z_2)$. Furthermore, define $\bar{\Delta} = \sqrt{\mathbb{E}_{Z_1, Z_2 \sim P_Z^2}[\Delta(Z_1, Z_2)^2]}$. Due to boundedness, it is clear that $\ell(w, z_i(S_i)) - \ell(w, z_i(\bar{S}_i))$ is $\Delta(z_i(1), z_i(0))$-sub-Gaussian under P_{S_i} for all $w \in \mathcal{W}$, $i \in [n]$, and $\tilde{\boldsymbol{z}} \in \mathcal{Z}^{2n}$. By following the same argument as above, this therefore leads to the bound

$$\overline{\text{gen}} \leq \sqrt{\frac{2\bar{\Delta}^2 I(W; \boldsymbol{S}|\tilde{\boldsymbol{Z}})}{n}}. \tag{6.6}$$

For simplicity, we will assume a bounded loss throughout this chapter, but we note that all bounds that are derived through a sub-Gaussianity argument can be generalized in this way.

While the bound in Theorem 6.1 achieves a slow $1/\sqrt{n}$-rate with respect to the training set size, this can, just as before, be improved at the cost of worse multiplicative constants. In this vein, Steinke and Zakynthinou (2020) also presented the following average bound. The result essentially follows along the same lines as Theorem 6.1, but using Theorem 3.32 for the concentration step.

Theorem 6.2. Assume that the range of the loss function $\ell(\cdot,\cdot)$ is $[0,1]$. For all constants $\gamma, \lambda > 0$ satisfying $\lambda(1-\gamma) + (e^\lambda - 1 - \lambda)(1+\gamma^2) \leq 0$, we have

$$L \leq \gamma \hat{L} + \frac{I(W; \boldsymbol{S}|\tilde{\boldsymbol{Z}})}{\lambda n}. \tag{6.7}$$

Under the assumption that the learning algorithm interpolates the training data almost surely, meaning that it achieves zero training loss, the constants in the bound can be improved.

Theorem 6.3. Assume that the range of the loss function $\ell(\cdot,\cdot)$ is $[0,1]$. Furthermore, assume that $\hat{L} = 0$, meaning that the algorithm interpolates the data almost surely. Then, we have

$$L \leq \frac{I(W; \boldsymbol{S}|\tilde{\boldsymbol{Z}})}{n \log 2}. \tag{6.8}$$

We will postpone the proof of this result, and instead prove it in Section 6.5, when we introduce the *evaluated* CMI. While it is possible to prove it without reference to evaluated CMI, as was done by Steinke and Zakynthinou (2020), the proof becomes somewhat shorter once we introduce it.

The constant $\log 2$ in the bound can be shown to be sharp. Indeed, as mentioned before, the conditional mutual information is trivially upper-bounded as $n \log 2$. Inserting this bound into (6.8) yields a population loss bound of 1. Thus, if the constant could be improved, we would have a non-trivial generalization bound that holds for any algorithm, which is not possible.

We now prove a generalization bound with the binary relative entropy on the left-hand side, as before. The functional form may appear surprising at first glance, but the reason for this quickly becomes apparent in the proof.

Theorem 6.4. Assume that the range of the loss function $\ell(\cdot,\cdot)$ is $[0,1]$. Then, for every $\gamma \in \mathbb{R}$,

$$d_\gamma\left(\hat{L} \,\middle\|\, \frac{\hat{L}+L}{2}\right) \leq d\left(\hat{L} \,\middle\|\, \frac{\hat{L}+L}{2}\right) \leq \frac{I(W; \boldsymbol{S}|\tilde{\boldsymbol{Z}})}{n}. \tag{6.9}$$

Proof. First, by Jensen's inequality and the definition of the parametrized binary relative entropy, we have

$$\begin{aligned} d\left(\hat{L} \,\middle\|\, \frac{\hat{L}+L}{2}\right) &= \sup_\gamma d_\gamma\left(\hat{L} \,\middle\|\, \frac{\hat{L}+L}{2}\right) \\ &\leq \sup_\gamma \mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[d_\gamma\left(L_{\boldsymbol{Z}_{\boldsymbol{S}}}(W) \,\middle\|\, \frac{L_{\boldsymbol{Z}_{\boldsymbol{S}}}(W)+L_{\boldsymbol{Z}_{\bar{\boldsymbol{S}}}}(W)}{2}\right)\right] \\ &\leq \sup_\gamma \mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[d_\gamma(L_{\boldsymbol{Z}_{\boldsymbol{S}}}(W) \,||\, L_{\tilde{\boldsymbol{Z}}}(W))\right]. \end{aligned} \tag{6.10}$$

In the last step, we used that for any W and $\boldsymbol{Z}$, the value of $(L_{\boldsymbol{Z}_{\boldsymbol{S}}}(W) + L_{\boldsymbol{Z}_{\bar{\boldsymbol{S}}}}(W))/2 = L_{\tilde{\boldsymbol{Z}}}(W)$ is actually independent of $\boldsymbol{S}$—it is just the average loss on the entire supersample. In fact, for fixed $W, \tilde{\boldsymbol{Z}}$, we have

$$\mathbb{E}_{P_{\boldsymbol{S}}}[L_{\boldsymbol{Z}_{\boldsymbol{S}}}(W)] = L_{\tilde{\boldsymbol{Z}}}(W). \tag{6.11}$$

Therefore, under $P_{\boldsymbol{S}}$, the second argument of the binary relative entropy in (6.10) is the mean of the first argument—this motivates the form of the bound. As per usual, we use the Donsker-Varadhan variational representation of the relative entropy to change measure from $P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}$ to $P_{W\tilde{\boldsymbol{Z}}}P_{\boldsymbol{S}}$. Then, we use Theorem 3.31 to find that

$$\log \mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}}P_{\boldsymbol{S}}}\left[e^{nd_\gamma\left(L_{\boldsymbol{Z}_{\boldsymbol{S}}}(W) \,||\, L_{\tilde{\boldsymbol{Z}}}(W)\right)}\right] \leq 0. \tag{6.12}$$

The final result follows after reorganizing terms. □

Without the CMI approach, the corresponding bound in Corollary 4.3 had the training loss as the first argument and the population loss as the second. Here, we instead have the arithmetic mean of the training and population loss as the second argument. The reason for this, as seen in the proof, is that the averaging is done over $\boldsymbol{S}$ rather than $\boldsymbol{Z}$, necessitating a different form in order to use the concentration result for the binary relative entropy. This gives rise to an additional factor of 2—similar to how this extra factor arose in the sub-Gaussian argument in the proof of Theorem 6.1 where we had to apply sub-Gaussianity to a bounded random variable with range $[-1, 1]$ instead of $[0, 1]$.

Naturally, the bound in (6.9) can be relaxed as before to obtain a result that more clearly illustrates the scaling of the bound, by following the same recipe used to derive Corollary 5.6. This yields a result very similar to Theorem 6.2, albeit with slightly different constants.

We conclude this section by discussing the application of the individual-sample technique and disintegration to the CMI framework. When introducing the individual-sample technique in Section 4.2, one of the main motivations was to avoid infiniteness of the mutual information. Now, as mentioned before, this problem has been solved with the CMI, which is always finite. However, if the CMI reaches its maximum value, our bounds are still vacuous—although finite. Hence, it is still of interest to apply these techniques in the CMI framework. This was done by, for instance, Haghifam *et al.* (2020). We present a bound incorporating these techniques below without proof—as expected, the bound is derived by suitably adapting the proof methods from Sections 4.2 and 4.3 to the CMI bound in Theorem 6.1.

Theorem 6.5. Assume that the range of the loss function $\ell(\cdot,\cdot)$ is $[0,1]$. Then,

$$\overline{\text{gen}} \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{P_{\tilde{\boldsymbol{Z}}}}\left[\sqrt{\frac{I^{\tilde{\boldsymbol{Z}}}(W;S_i)}{n}}\right], \tag{6.13}$$

where $I^{\tilde{\boldsymbol{Z}}}(W;S_i) = D(P_{W|\tilde{\boldsymbol{Z}}S_i} \,||\, P_{W|\tilde{\boldsymbol{Z}}} \,|\, P_{S_i})$. Similar extensions can be obtained for the other CMI bounds in this section.

The application of the individual-sample technique in Theorem 6.5 can be naturally extended as follows. While we have $I(W;\boldsymbol{S}|\tilde{\boldsymbol{Z}}) \leq I(W;\boldsymbol{Z}_{\boldsymbol{S}})$, meaning that the CMI-based generalization bounds improve on their mutual information-based counterparts (up to constants), the same does not hold true when comparing Theorem 6.5 to its individual-sample counterpart in Corollary 4.6. The issue is that conditioning on the entire supersample can reveal an unnecessarily large amount of information, so that there are scenarios where we are better off not using any conditioning. The technical reason behind this is that, in the derivation of Theorem 6.5, parts of $\tilde{\boldsymbol{Z}}$ that can be marginalized out in the samplewise decomposition of the generalization error are not marginalized. This issue was noted by both Rodríguez-Gálvez *et al.* (2020) and Zhou *et al.* (2021), who rectified it to obtain the following result.

Theorem 6.6. Assume that the range of the loss function $\ell(\cdot,\cdot)$ is $[0,1]$. Then,

$$\overline{\text{gen}} \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{P_{\tilde{Z}_i\tilde{Z}_{i+n}}}\left[\sqrt{\frac{2I^{\tilde{Z}_i\tilde{Z}_{i+n}}(W;S_i)}{n}}\right]. \tag{6.14}$$

The proof of this is essentially the same as for Theorem 6.5, but with more care taken with regards to marginalization. Due to the data-processing inequality, this always improves on Corollary 4.6, up to a constant factor.

6.3 PAC-Bayesian Generalization Bounds

So far, we have used the CMI framework to obtain generalization bounds in expectation. Indeed, this has been the main focus in the recent CMI literature. However, as in Chapter 5, we can also derive bounds in probability, that is, bounds on the PAC-Bayesian or single-draw loss. PAC-Bayesian bounds were the focus of Audibert (2004) and Catoni (2007) in their use of almost exchangeable priors. In this section, we will discuss such PAC-Bayesian generalization bounds in the CMI framework. These results—which can be seen as PAC-Bayesian analogues of the results in Section 6.2 or CMI analogues of the results in Section 5.2—can be derived for all manner of bounds discussed previously. Here, we will present one such extension of a previous bound, as well as a simplified version of an excess risk bound due to Grünwald *et al.* (2021) based on the Bernstein condition. We will also discuss the relation between the "CMI prior" and the data-dependent prior mentioned in Section 5.2, and touch upon some connections to other topics in learning theory. These latter points will be fleshed out further in Chapter 7.

We begin by presenting a PAC-Bayesian version of Theorem 6.1, which can also be seen as a CMI version of Corollary 5.3.

Theorem 6.7. Assume that the range of the loss function $\ell(\cdot,\cdot)$ is $[0,1]$ and assume that $P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}} \ll Q_{W|\tilde{\boldsymbol{Z}}}$. Then, with probability at least $1-\delta$

under $P_{\tilde{\boldsymbol{Z}}\boldsymbol{S}}$,

$$\mathbb{E}_{P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\Big[\mathrm{gen}(W,\tilde{\boldsymbol{Z}},\boldsymbol{S})\Big] \leq \sqrt{\frac{2}{n-1}\left(D(P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}}\,||\,Q_{W|\tilde{\boldsymbol{Z}}}) + \log\frac{\sqrt{n}}{\delta}\right)}. \tag{6.15}$$

The role of the unused data points, $\boldsymbol{Z}(\bar{\boldsymbol{S}})$, and the auxiliary conditional distribution $Q_{W|\tilde{\boldsymbol{Z}}}$, which acts as the prior in the PAC-Bayesian bounds for the CMI framework, merit some discussion. In the average bounds presented earlier in this section, these samples are purely hypothetical "ghost samples," and the data-generation process can be seen as a thought experiment that is just used for the proofs. In the left-hand side, averaging $\mathrm{gen}(W,\tilde{\boldsymbol{Z}},\boldsymbol{S})$ over $P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}$ transforms the test loss $L_{\boldsymbol{Z}_{\bar{\boldsymbol{S}}}}(W)$ into the ordinary population loss $L_{P_Z}(W)$. On the right-hand side, the information measure is the conditional mutual information $I(W;\boldsymbol{S}|\tilde{\boldsymbol{Z}})$, where these ghost samples are averaged out.

In contrast, for the data-dependent bound in Theorem 6.7, the left-hand side depends on the PAC-Bayesian test loss $\mathbb{E}_{P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\Big[L_{\boldsymbol{Z}_{\bar{\boldsymbol{S}}}}(W)\Big]$. Since this is an unbiased estimate of the population loss, one can convert this into a bound on the PAC-Bayesian population loss through the triangle inequality (Hellström and Durisi, 2020b, Thm. 3). Furthermore, the right-hand side actually explicitly depends on these unused training samples. An upside of this is that this leads to bounds that are actually manageable to compute. Indeed, given a set of $2n$ training samples, one can just implement the subset-selection procedure in practice, use the obtained $\boldsymbol{Z}_{\boldsymbol{S}}$ to select a hypothesis, and select the prior freely based on $\tilde{\boldsymbol{Z}}$—provided that one does not use any knowledge of $\boldsymbol{S}$. This shares many similarities with the data-splitting approach for data-dependent priors in PAC-Bayes, discussed in Section 5.2.3, wherein one splits the training data into two parts: one part $\boldsymbol{Z}_P$ for selecting the prior, and one part $\boldsymbol{Z}_B$ for evaluating the training loss in the generalization bound. Crucially, in the data-splitting approach, the selected hypothesis is still allowed to depend on all of the training data. The two approaches are explained pictorially in Fig. 6.1 and Fig. 6.2 respectively.

From a practical standpoint, there are many reasons to prefer the data-splitting approach. The most obvious difference is that with the

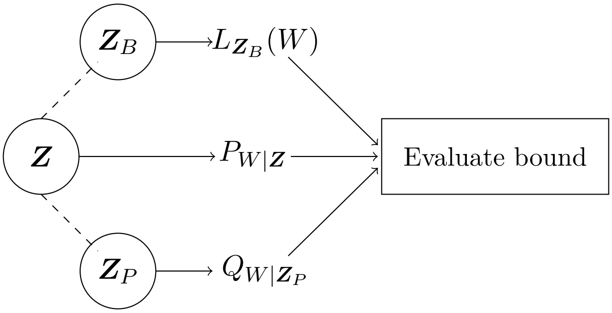

Figure 6.1: The data-splitting approach to data-dependent priors, discussed in Section 5.2.3.

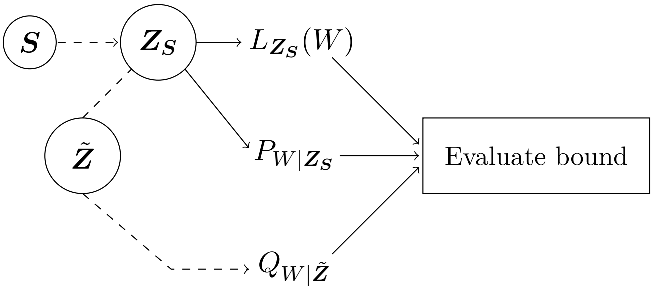
$\boldsymbol{S}$
$\boldsymbol{Z}_{\boldsymbol{S}}$
$L_{\boldsymbol{Z}_{\boldsymbol{S}}}(W)$
$\tilde{\boldsymbol{Z}}$
$P_{W|\boldsymbol{Z}_{\boldsymbol{S}}}$
Evaluate bound
$Q_{W|\tilde{\boldsymbol{Z}}}$

Figure 6.2: The CMI approach to data-dependent priors.

data-splitting approach, all available samples can be used as input to the learning algorithm, which typically leads to better performance. Furthermore, the PAC-Bayesian bound can be directly optimized, meaning that the information measure between posterior and prior is used as a regularizer. For the CMI approach, such regularization would introduce illegal dependencies between W and $\tilde{\boldsymbol{Z}}$, violating the Markov property upon which the proof is based. Thus, the resulting generalization bounds would no longer hold.

However, if the motivation is to theoretically understand learning algorithms, rather than to derive risk certificates for practical hypotheses or to devise the best algorithm, the difference is more conceptual. The data-splitting approach can be seen as a generalization of *compression schemes* (Shalev-Shwartz and Ben-David, 2014). Roughly speaking, a learning algorithm is a compression scheme of size k if its output on any training set $\boldsymbol{Z}$ with $n > k$ samples is the same as its output on $\boldsymbol{Z}_C$, consisting of k samples from $\boldsymbol{Z}$. Indeed, a version of the standard generalization bounds for stable compressors can be derived from the data-splitting PAC-Bayesian bound by setting $\boldsymbol{Z}_P = \boldsymbol{Z}_C$, paying a union bound cost for the $\binom{n}{k}$ possible choices of $\boldsymbol{Z}_C$. Since the output of the learning algorithm based on $\boldsymbol{Z}$ can be obtained entirely on the basis of $\boldsymbol{Z}_C$, the relative entropy will vanish, as the data-dependent prior exactly matches the posterior. So, in this compression-related approach, the hypothesis is allowed to depend strongly on a few samples, as long as the dependence on the remaining samples is weak.

The CMI approach, on the other hand, can be seen as drawing on the notion of algorithmic stability, discussed in Section 1.4. Intuitively, algorithmic stability measures how sensitive the output hypothesis is to the inclusion of any one sample in the training set. In a sense, this is closely related to the CMI. Indeed, the individual-sample CMI $I(W; S_i|Z_i, Z_{i+n})$ measures how strong the dependence of the hypothesis is on the specific ith sample. This connection is explored in more detail by Harutyunyan *et al.* (2021). We will discuss all this further in Chapter 7, where the information complexity of specific algorithms is evaluated. Of course, by repeating the arguments from Section 5.2.3, we can combine the data-splitting technique and the CMI approach. This ensures that the prior and posterior have a fixed set of samples in common, which are absent from the training loss in the bound, while the remaining samples are randomly selected through the CMI procedure.

We conclude this section by stating a variant of a result of Grünwald *et al.* (2021, Cor. 1), which provides a PAC-Bayesian excess risk bound with potentially fast rates. Since the statement of the result and its proof are quite involved, we will only provide a simplified version without proof. The full details, including extensions to average bounds through an exponential stochastic inequality and other variants, can be found in the work of Grünwald *et al.* (2021).

Theorem 6.8. Assume that the range of the loss function $\ell(\cdot,\cdot)$ is $[0,1]$. Furthermore, assume that the β-Bernstein condition is satisfied, *i.e.*, for some $\beta \in [0,1]$, there exists a $w^* \in \mathcal{W}$ such that, for all $w \in \mathcal{W}$,

$$\mathbb{E}_{P_Z}\left[(\ell(w,Z)-\ell(w^*,Z))^2\right] \le 4\left(\mathbb{E}_{P_Z}[\ell(w,Z)-\ell(w^*,Z)]\right)^{\beta}. \qquad (6.16)$$

Then, for some $C_1, C_2 \in \mathbb{R}^+$ with probability at least $1-\delta$ under $P_{\boldsymbol{Z}_{\boldsymbol{S}}}$,

$$\mathbb{E}_{P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[\mathrm{gen}(W,\tilde{\boldsymbol{Z}},\boldsymbol{S})\right] \le \min(1,2\beta)\left(\mathbb{E}_{P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}}}[L_{\boldsymbol{Z}_{\boldsymbol{S}}}(W)] - L_{\boldsymbol{Z}_{\boldsymbol{S}}}(w^*)\right)$$
$$+ C_1\left(\frac{\mathbb{E}_{P_{\boldsymbol{Z}_{\bar{\boldsymbol{S}}}}}\left[D(P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}}\,||\,Q_{W|\tilde{\boldsymbol{Z}}})\right] + \log(\sqrt{n})}{n}\right)^{\frac{1}{2-\beta}} + \frac{C_2\log\frac{1}{\delta}}{\sqrt{n}}. \quad (6.17)$$

The first term on the right-hand side of (6.17) is the (scaled) *empirical excess risk*, *i.e.*, the degree to which the training loss of the learning algorithm exceeds the training loss of the optimal hypothesis with respect to the population loss. For many algorithms this is negligible, and for empirical risk minimizers, it is guaranteed to be non-positive. There are several notable aspects of this result. Since the loss is bounded, the β-Bernstein condition always holds with $\beta = 0$, which means that the slow rate of $1/\sqrt{n}$ can be obtained (in which case the empirical excess risk does not enter the bound). However, for smooth losses such as the squared or logistic loss, it also holds with $\beta = 1$, enabling the relative entropy-dependent term to decay faster. Furthermore, the relative entropy enters only averaged over the test data. While this leads to the relative entropy being non-empirical, in the sense that it cannot be computed based on the training set and learning algorithm, it can still be shown to be bounded in several cases, such as for hypothesis classes with bounded VC dimension. Indeed, this is the main focus of Grünwald *et al.* (2021): deriving a fast-rate bound that is well-behaved for VC classes. We will discuss this further in Section 7.3. We emphasize again that the result in Theorem 6.8 is not stated in its full generality nor tightness, but has been significantly simplified in terms of assumptions, constants, and logarithmic dependencies in the interest of brevity.

6.4 Single-Draw Generalization Bounds

Before moving on to extensions of the CMI framework, we turn to single-draw bounds. As for the average and PAC-Bayesian bounds, we can also derive CMI versions of the single-draw bounds from Section 5.3.

We begin by stating a basic single-draw bound, now given in terms of the conditional information density $\imath(W,\boldsymbol{S}|\tilde{\boldsymbol{Z}})$. When averaged over the joint distribution $P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}$, the conditional information density gives the CMI $I(W;\boldsymbol{S}|\tilde{\boldsymbol{Z}})$—hence the name. As the proof is a straightforward adaptation of previous derivations, we do not give it explicitly.

Theorem 6.9. Assume that the range of the loss function $\ell(\cdot,\cdot)$ is $[0,1]$. Then, with probability at least $1-\delta$ under $P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}$,

$$\left|\mathrm{gen}(W,\tilde{\boldsymbol{Z}},\boldsymbol{S})\right| \le \sqrt{\frac{2}{n-1}\left(\imath(W,\boldsymbol{S}|\tilde{\boldsymbol{Z}}) + \log\frac{\sqrt{n}}{\delta}\right)}. \tag{6.18}$$

The techniques from Section 5.3.2, where repeated uses of Hölder's inequality were used to obtain a generic bound on the probability of an event under one distribution in terms of another, can also be extended to the CMI framework. In terms of the proof, we need to consider three random variables, perform the change of measure conditioned on one of them, and make use of Hölder's inequality an additional time. We present the generic result below.

Theorem 6.10. For all constants $\alpha,\gamma,\alpha',\gamma',\tilde{\alpha},\tilde{\gamma}>1$ such that $1/\alpha+1/\gamma = 1/\alpha'+1/\gamma' = 1/\tilde{\alpha}+1/\tilde{\gamma} = 1$ and all measurable sets $\mathcal{E} \in \mathcal{W}\times\mathcal{Z}^{2n}\times\{0,1\}^n$,

$$\begin{aligned} P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}[\mathcal{E}] \le{}& \mathbb{E}_{P_{\tilde{\boldsymbol{Z}}}}^{1/\tilde{\gamma}}\left[\mathbb{E}_{P_{W|\tilde{\boldsymbol{Z}}}}^{\tilde{\gamma}/\gamma'}\left[P_{\boldsymbol{S}}^{\gamma'/\gamma}[\mathcal{E}_{W\tilde{\boldsymbol{Z}}}]\right]\right]\times \\ & \mathbb{E}_{P_{\tilde{\boldsymbol{Z}}}}^{1/\tilde{\alpha}}\left[\mathbb{E}_{P_{W|\tilde{\boldsymbol{Z}}}}^{\tilde{\alpha}/\alpha'}\left[\mathbb{E}_{P_S}^{\alpha'/\alpha}\left[e^{\alpha\imath(W,\boldsymbol{S}|\tilde{\boldsymbol{Z}})}\right]\right]\right]. \end{aligned} \tag{6.19}$$

Here, $\mathcal{E}_{W\tilde{\boldsymbol{Z}}} = \{\boldsymbol{S} : (W,\tilde{\boldsymbol{Z}},\boldsymbol{S})\in\mathcal{E}\}$.

Proof. First, we rewrite $P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}[\mathcal{E}]$ in terms of the expectation of the indicator function $1_{\mathcal{E}}$ and perform a change of measure:

$$P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}[\mathcal{E}] = \mathbb{E}_{P_{W|\tilde{\boldsymbol{Z}}}P_{\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[1_{\mathcal{E}}\cdot\frac{\mathrm{d}P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}{\mathrm{d}P_{W|\tilde{\boldsymbol{Z}}}P_{\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\right] \tag{6.20}$$

$$= \mathbb{E}_{P_{W|\tilde{\boldsymbol{Z}}}P_{\tilde{\boldsymbol{Z}}}P_{\boldsymbol{S}}}\left[1_{\mathcal{E}}\cdot e^{\imath(W,\boldsymbol{S}|\tilde{\boldsymbol{Z}})}\right]. \tag{6.21}$$

To obtain the desired result, we apply Hölder's inequality thrice. Let α, γ, α', γ', $\tilde{\alpha}$, $\tilde{\gamma}>1$ be constants such that $1/\alpha+1/\gamma = 1/\alpha'+1/\gamma' =$

$1/\tilde{\alpha} + 1/\tilde{\gamma} = 1$. Then,

$$P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}[\mathcal{E}] \le \mathbb{E}_{P_{W|\tilde{\boldsymbol{Z}}}P_{\tilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{1/\gamma}_{P_{\boldsymbol{S}}}\left[1_{\mathcal{E}_{W\tilde{\boldsymbol{Z}}}}\right]\cdot \mathbb{E}^{1/\alpha}_{P_{\boldsymbol{S}}}\left[e^{\alpha\imath(W,\boldsymbol{S}|\tilde{\boldsymbol{Z}})}\right]\right] \tag{6.22}$$

$$\le \mathbb{E}_{P_{\tilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{1/\gamma'}_{P_{W|\tilde{\boldsymbol{Z}}}}\left[P_{\boldsymbol{S}}^{\gamma'/\gamma}[\mathcal{E}_{W\tilde{\boldsymbol{Z}}}]\right]\cdot \mathbb{E}^{1/\alpha'}_{P_{W|\tilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{\alpha'/\alpha}_{P_{\boldsymbol{S}}}\left[e^{\alpha\imath(W,\boldsymbol{S}|\tilde{\boldsymbol{Z}})}\right]\right]\right]$$

$$\le \mathbb{E}^{1/\tilde{\gamma}}_{P_{\tilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{\tilde{\gamma}/\gamma'}_{P_{W|\tilde{\boldsymbol{Z}}}}\left[P_{\boldsymbol{S}}^{\gamma'/\gamma}[\mathcal{E}_{W\tilde{\boldsymbol{Z}}}]\right]\right]\cdot \mathbb{E}^{1/\tilde{\alpha}}_{P_{\tilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{\tilde{\alpha}/\alpha'}_{P_{W|\tilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{\alpha'/\alpha}_{P_{\boldsymbol{S}}}\left[e^{\alpha\imath(W,\boldsymbol{S}|\tilde{\boldsymbol{Z}})}\right]\right]\right].$$

□

Many different types of bounds can be obtained by making different choices for the three free parameters in Theorem 6.10. We will focus on a choice that leads to bounds in terms of a version of the conditional α-mutual information (Definition 3.14).

We emphasize two properties of the conditional α-mutual information. First, in the limit $\alpha \to \infty$, it reduces to the conditional maximal leakage (Issa *et al.*, 2020, Thm. 6):

$$\mathcal{L}(\boldsymbol{S} \to W|\tilde{\boldsymbol{Z}}) = \log \operatorname*{ess\,sup}_{P_{\tilde{\boldsymbol{Z}}}} \mathbb{E}_{P_{W|\tilde{\boldsymbol{Z}}}}\left[\operatorname*{ess\,sup}_{P_{\boldsymbol{S}|\tilde{\boldsymbol{Z}}}} e^{\imath(W,\boldsymbol{S}|\tilde{\boldsymbol{Z}})}\right]. \tag{6.23}$$

Second, for $\alpha > 1$, one can see that the conditional α-mutual information is upper-bounded by the conditional Rényi divergence of order α, as shown in (3.27).

After this aside, we return to presenting the generalization bound in terms of the conditional α-mutual information.

Corollary 6.11. Assume that the range of the loss function $\ell(\cdot,\cdot)$ is $[0, 1]$. Then, for any fixed $\alpha > 1$, the following holds with probability at least $1 - \delta$ under $P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}$:

$$\left|\text{gen}(W, \tilde{\boldsymbol{Z}}, \boldsymbol{S})\right| \le \sqrt{\frac{2}{n}\left(I_\alpha(W; \boldsymbol{S} \mid \tilde{\boldsymbol{Z}}) + \log 2 + \frac{\alpha}{\alpha - 1}\log\frac{1}{\delta}\right)}. \tag{6.24}$$

Proof. In (6.19), set $\tilde{\alpha} = \alpha$ and let $\alpha' \to 1$, which implies that $\tilde{\gamma} = \gamma$ and $\gamma' \to \infty$. Also, let $\mathcal{E}$ be the error event

$$\mathcal{E} = \{W, \tilde{\boldsymbol{Z}}, \boldsymbol{S} : \left|\text{gen}(W, \tilde{\boldsymbol{Z}}, \boldsymbol{S})\right| > \varepsilon\}. \tag{6.25}$$

For this choice of parameters, the second factor in (6.19) reduces to

$$\mathbb{E}^{1/\alpha}_{P_{\tilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{\alpha}_{P_{W|\tilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{1/\alpha}_{P_S}\left[\exp\Big(\alpha\imath(W,\boldsymbol{S}|\tilde{\boldsymbol{Z}})\Big)\right]\right]\right]$$
$$= \exp\left(\frac{\alpha-1}{\alpha}I_\alpha(W;\boldsymbol{S}\,|\,\tilde{\boldsymbol{Z}})\right). \quad (6.26)$$

Furthermore, we can bound $P_{\boldsymbol{S}}[\mathcal{E}_{W\tilde{\boldsymbol{Z}}}]$ in the first factor in (6.19) by using sub-Gaussianity to find that, for all W and $\tilde{\boldsymbol{Z}}$,

$$P_{\boldsymbol{S}}[\mathcal{E}_{W\tilde{\boldsymbol{Z}}}] \le 2\exp\left(-\frac{n\varepsilon^2}{2}\right). \quad (6.27)$$

Using this in the first factor of (6.19), we conclude that

$$\lim_{\gamma'\to\infty}\mathbb{E}^{1/\gamma}_{P_{\tilde{\boldsymbol{Z}}}}\left[\mathbb{E}^{\gamma/\gamma'}_{P_{W|\tilde{\boldsymbol{Z}}}}\left[P^{\gamma'/\gamma}_{\boldsymbol{S}}[\mathcal{E}_{W\tilde{\boldsymbol{Z}}}]\right]\right] = \mathbb{E}^{1/\gamma}_{P_{\tilde{\boldsymbol{Z}}}}\left[\left(\operatorname*{ess\,sup}_{P_{W|\tilde{\boldsymbol{Z}}}} P^{1/\gamma}_{\boldsymbol{S}}[\mathcal{E}_{W\tilde{\boldsymbol{Z}}}]\right)^{\gamma}\right]$$
$$\le \left(2\exp\left(-\frac{n\varepsilon^2}{2}\right)\right)^{1/\gamma}. \quad (6.28)$$

By substituting (6.26) and (6.28) into (6.19), noting that $1/\gamma = (\alpha - 1)/\alpha$, we conclude that

$$P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}[\mathcal{E}] \le \left(2\exp\left(-\frac{n\varepsilon^2}{2}\right)\right)^{\frac{\alpha-1}{\alpha}}\cdot\exp\left(\frac{\alpha-1}{\alpha}I_\alpha(W;\boldsymbol{S}\,|\,\tilde{\boldsymbol{Z}})\right). \quad (6.29)$$

We obtain the desired result by requiring the right-hand side of (6.29) to equal δ and solving for ε. □

By letting $\alpha\to\infty$, we obtain a generalization bound in terms of the conditional maximal leakage:

$$\left|\operatorname{gen}(W,\tilde{\boldsymbol{Z}},\boldsymbol{S})\right| \le \sqrt{\frac{2}{n}\left(\mathcal{L}(\boldsymbol{S}\to W|\tilde{\boldsymbol{Z}}) + \log\frac{2}{\delta}\right)}. \quad (6.30)$$

It can be shown that $\mathcal{L}(\boldsymbol{S}\to W|\tilde{\boldsymbol{Z}}) \le \mathcal{L}(\boldsymbol{Z}_{\boldsymbol{S}}\to W)$ (Hellström and Durisi, 2020b, Thm. 5). Thus, up to constants and the penalty term incurred to obtain a bound on the population loss (as per Hellström and Durisi, 2020b, Thm. 3), (6.30) improves on (5.39).

6.5 Evaluated CMI and f-CMI

As noted by Steinke and Zakynthinou (2020), there is a potential deficiency that comes with measuring information as captured by the hypothesis W itself. For instance, if W is a real number, we can take the output of an algorithm with low CMI, and change W so that it encodes the training set in its insignificant digits. For most settings, this change should have a negligible effect on the generalization of the algorithm, but the CMI will be maximized, leading to vacuous bounds. Another way to see the issue is to consider the case where W consists of the weights of a neural network. Neural networks (which will be discussed in Chapter 8) typically possess many symmetries, such as permutation and scaling invariance, so that different values of W can represent the exact same function. Ideally, we would want to obtain generalization bounds in terms of a measure that is more directly related to the predictions our hypothesis produces and the losses that they incur.

As it turns out, this can be accomplished in a straightforward way. In fact, we barely need to change the derivations we have used so far. Consider, for instance, the derivation of Theorem 6.1. In the proof, W only appears in a "processed" version, either through the loss on a training sample or the loss on a test sample. Hence, the derivation can be adapted so that no explicit mention is made of W, but instead, only the losses that it incurs on the supersample appear in both the derivation and the final result. Motivated by this, we introduce the notation $\mathbf{\Lambda} \in [0,1]^{2n}$ to denote the random vector that contains the losses that the hypothesis incurs on the entire supersample. Specifically, the ith element of $\mathbf{\Lambda}$ is $\Lambda_i = \ell(W, \tilde{Z}_i)$. Now, we proceed as in the proof of Theorem 6.1, with $\mathbf{\Lambda}$ replacing W. As observed by Steinke and Zakynthinou (2020), this leads to a bound in terms of $I(\mathbf{\Lambda}; \boldsymbol{S} | \tilde{\boldsymbol{Z}})$, referred to as the *evaluated* CMI, or e-CMI for short. In fact, as pointed out by Haghifam *et al.* (2022), we can even avoid any explicit reference to the supersample, leading to a bound in terms of the evaluated mutual information $I(\mathbf{\Lambda}; \boldsymbol{S})$, abbreviated as e-MI. For Theorem 6.1, this can be taken even further, in fact, by noting that only the *difference* between training and test losses actually enters the derivation, as done by Wang and Mao (2023c). To this end, we define the vector of loss differences $\mathbf{\Delta}$, with elements given by $\Delta_i = \ell(W, \tilde{Z}_{i+n}) - \ell(W, \tilde{Z}_i)$, and the resulting loss-difference mutual information $I(\boldsymbol{\Delta}; \boldsymbol{S})$, or ld-MI for short. This gives rise to the following three upper bounds on the average generalization error. While we only present the bounds for the full-sample square-root bound, analogous results can be derived for other comparator functions and using individual samples.

Theorem 6.12. Assume that the range of the loss function $\ell(\cdot, \cdot)$ is $[0, 1]$. Then,

$$\overline{\text{gen}} \leq \sqrt{\frac{2I(\boldsymbol{\Delta}; \boldsymbol{S})}{n}} \leq \sqrt{\frac{2I(\boldsymbol{\Lambda}; \boldsymbol{S})}{n}} \leq \sqrt{\frac{2I(\boldsymbol{\Lambda}; \boldsymbol{S}|\tilde{\boldsymbol{Z}})}{n}}. \tag{6.31}$$

Proof. We start off by rewriting the generalization gap in terms of $\boldsymbol{\Delta}$:

$$\mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[L_{\boldsymbol{Z}_{\bar{\boldsymbol{S}}}}(W) - L_{\boldsymbol{Z}_{\boldsymbol{S}}}(W)\right] = \mathbb{E}_{P_{\boldsymbol{\Lambda}\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\Lambda_{i+\bar{S}_i n} - \Lambda_{i+S_i n}\right)\right] \tag{6.32}$$

$$= \mathbb{E}_{P_{\boldsymbol{\Lambda}\boldsymbol{S}}}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\Lambda_{i+\bar{S}_i n} - \Lambda_{i+S_i n}\right)\right] \tag{6.33}$$

$$= \mathbb{E}_{P_{\boldsymbol{\Delta}\boldsymbol{S}}}\left[\frac{1}{n}\sum_{i=1}^{n}(-1)^{S_i}\Delta_i\right]. \tag{6.34}$$

The remainder of the proof proceeds by changing measure to $P_{\boldsymbol{\Delta}}P_{\boldsymbol{S}}$ through the use of the Donsker-Varadhan variational representation of the relative entropy. The remaining steps of the proof are identical to those used in deriving Theorem 6.1. The relaxation in terms of the e-MI follows due to the data-processing inequality. Finally, since $\boldsymbol{S}$ and $\tilde{\boldsymbol{Z}}$ are independent, the relaxation in terms of the e-CMI follows since conditioning on independent random variables does not decrease mutual information. □

Expressing generalization bounds in terms of ld-MI, e-MI, and e-CMI can have drastic consequences for the tightness of the resulting bound. This new approach guarantees that any two hypotheses that lead to the same losses on the supersample will be considered equivalent by our information measure. While we will not present it explicitly, it is of course possible to apply this approach to the other bounds discussed in the previous sections, including the PAC-Bayesian and single-draw bounds.

By the data-processing inequality, bounds in terms of the CMI from earlier in this chapter can be re-obtained from the ld-MI. For supervised learning, where the learning algorithm implements a function $f_W : \mathcal{X} \to \mathcal{Y}$, we can also consider the *functional* CMI (f-CMI), studied by Harutyunyan *et al.* (2021). Specifically, if we assume that each sample consists of label—example pairs $\tilde{Z}_i = (\tilde{X}_i, \tilde{Y}_i)$, and let $\mathbf{F} = (F_1, \dots, F_{2n})$ denote the vector of predictions induced by W, *i.e.*, $F_i = f_W(\tilde{X}_i)$, we get the following chain of inequalities:

$$I(\boldsymbol{\Delta}; \boldsymbol{S}) \leq I(\boldsymbol{\Lambda}; \boldsymbol{S}) \leq I(\boldsymbol{\Lambda}; \boldsymbol{S}|\tilde{\boldsymbol{Z}}) \leq I(\mathbf{F}; \boldsymbol{S}|\tilde{\boldsymbol{Z}}) \leq I(W; \boldsymbol{S}|\tilde{\boldsymbol{Z}}). \tag{6.35}$$

Here, each step is a consequence of the data-processing inequality (or conditioning on independent random variables). Thus, the tightest of these bounds is the one in terms of the ld-MI, and all the others can be obtained in a straightforward way from this. It should be noted that the bounds in terms of evaluated mutual informations are, in a sense, more restrictive: they only apply to a specific loss function. In contrast, bounds in terms of the f-CMI and CMI bounds apply to any bounded loss function. Finally, bounds in terms of the CMI only require knowledge of the hypothesis itself. In Chapters 7 and 8, we will provide interpretations of each of these information-theoretic quantities as components of generalization bounds.

We end this section by providing the promised proof of the sharp generalization bound for interpolating learning algorithms, as mentioned after Theorem 6.3. We will do this through a communication-inspired proof, due to Wang and Mao (2023c).

Theorem 6.13. Assume that the loss function is binary, meaning that for all $w \in \mathcal{W}$ and $z \in \mathcal{Z}$, $\ell(w, z) \in \{0, 1\}$. Consider an interpolating learning algorithm, so that the training loss $\hat{L} = 0$. Then,

$$L = \frac{1}{n} \sum_{i=1}^{n} \frac{I(\Delta_i; S_i)}{\log 2} \leq \frac{I(\boldsymbol{\Delta}; \boldsymbol{S})}{n \log 2}. \tag{6.36}$$

Proof. Consider the weighted directed graph in Fig. 6.3, depicting the communication channel between S_i and $\Delta_i = \ell(W, \tilde{Z}_{i+n}) - \ell(W, \tilde{Z}_i)$ that is induced by the learning problem.

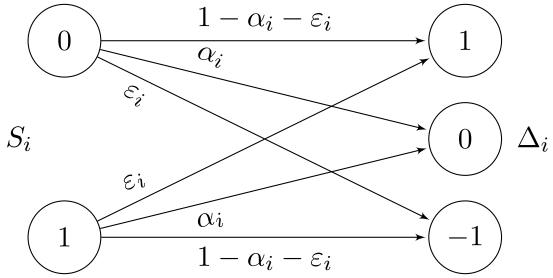

Figure 6.3: Communication channel from S_i to Δ_i induced by the learning algorithm.

We can now interpret the meaning of these transitions, and hence their probabilities. First, notice that both $0 \to -1$ and $1 \to 1$ imply that the learning algorithm incurs a loss on the training sample, which contradicts the interpolating assumption. Hence, we have $\varepsilon_i = 0$. Next, note that any transition to $\Delta_i = 0$ means that the test loss is zero, so that we do not incur a loss. Only the transitions $0 \to 1$ and $1 \to -1$ represent situations where a loss is incurred. Hence, for each input, the probability of incurring a loss on the test sample is $1 - \alpha_i$. Furthermore, for the specified communication channel, it can be shown that the Shannon capacity (with $\varepsilon_i = 0$) is $(1-\alpha_i)\log 2$ (Cover and Thomas, 2006, Problem 7.13). It is well-known that this equals the mutual information between input and output for a uniform input distribution, *i.e.*, $I(\Delta_i; S_i)$. Hence,

$$L = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{P_{\boldsymbol{\Delta S}}}\Big[(-1)^{S_i}\Delta_i\Big] = \frac{1}{n}\sum_{i=1}^{n}(1-\alpha_i) = \frac{1}{n}\sum_{i=1}^{n}\frac{I(\Delta_i; S_i)}{\log 2}. \quad (6.37)$$

The full-sample relaxation follows by the chain rule and conditioning on independent random variables. □

Thus, remarkably, for the case of a binary loss and interpolating learning algorithm, the average population loss can be exactly characterized in terms of the samplewise, loss-difference mutual information. By progressively upper-bounding this information measure through the data-processing inequality and the chain rule of mutual information, we can go step by step all the way to bounds in terms of the mutual information between the hypothesis and the training data. More on this result, as well as several interesting extensions, can be found in the work of Wang and Mao (2023c).

Now, Theorem 6.13 and data-processing do not directly imply Theorem 6.3, since the latter assumes a generic bounded loss. In order to relate these two results, we need the following observations. Let $\ell(\cdot,\cdot)$ be a generic loss function bounded to $[0,1]$, and let $\tilde{\ell}(\cdot,\cdot)$ denote a binarized version of the underlying loss function, given by $\tilde{\ell}(w,z) = 1\{\ell(w,z) > 0\}$. Let $\tilde{L}$ denote the population loss with respect to $\tilde{\ell}(\cdot,\cdot)$, and let $\tilde{\boldsymbol{\Delta}}$ denote the loss-difference vector with respect to $\tilde{\ell}(\cdot,\cdot)$. Since $\tilde{\ell}(\cdot,\cdot)$ is an upper bound to $\ell(\cdot,\cdot)$, we have $L \leq \tilde{L}$. Furthermore, since $\tilde{\boldsymbol{\Delta}}$ is a processed version of $\boldsymbol{\Delta}$, we have $I(\tilde{\Delta}_i; S_i) \leq I(\Delta_i; S_i)$. Thus, Theorem 6.13 can be relaxed in order to obtain Theorem 6.3.

6.6 Leave-One-Out CMI

When we introduced the CMI framework, the size of the supersample being $2n$ was natural: the aim was to normalize the information carried by each sample to 1 bit. Then, the bounds were derived by comparing the loss on the randomly selected samples, which gives a training loss, and the loss on the parts of the supersample that were *not* selected, which gives a test loss. In a sense, however, the n unused samples of $\tilde{\boldsymbol{Z}}$ seem quite wasteful. Indeed, if we instead were to use all but one sample for the training set, the single remaining sample would suffice for a test loss that provides an unbiased estimate of the population loss.

As it turns out, a variant of the CMI framework where the supersample is of size $n+1$ is possible, as demonstrated independently by Haghifam *et al.* (2022) and Rammal *et al.* (2022). In order to avoid confusion with the $2n$-supersample setup, we will use a slightly different notation. Specifically, we let $\dot{\boldsymbol{Z}} = (\dot{Z}_1, \dots, \dot{Z}_{n+1})$ denote a vector of $n+1$ samples drawn independently from P_Z, and let U be drawn uniformly at random from $[n+1] = \{1, \dots, n+1\}$. Based on this, the training set $\boldsymbol{Z}_{\bar{U}}$ is formed by removing the Uth element from $\dot{\boldsymbol{Z}}$, while the Uth element Z_U is a test sample. We denote the vector of losses on the supersample by $\dot{\boldsymbol{\Lambda}}$, with elements given by $\dot{\Lambda}_i = \ell(W, \dot{Z}_i)$. Throughout, we assume that the range of $\ell(\cdot,\cdot)$ is $[0,1]$. Using this setup, we can derive bounds in terms of the leave-one-out CMI $I(W; U|\dot{\boldsymbol{Z}})$, or loo-CMI for short, or analogous variants as we have seen before, such as the evaluated loo-MI $I(\dot{\mathbf{\Lambda}}; U)$. The name is due to its connections to the leave-one-out loss cross validation error, defined as

$$\text{loo-cv}(\dot{\mathbf{\Lambda}}, U) = \frac{1}{n} \sum_{i \neq U} \dot{\Lambda}_i - \dot{\Lambda}_U. \tag{6.38}$$

Note that, when averaged over the joint distribution of the random variables involved, the leave-one-out cross validation error equals the generalization gap $\overline{\text{gen}}$.

Compared to the CMI quantities from before, the loo-CMI is significantly less complex to compute. When computing the CMI (and evaluated versions of it), one needs to average over the 2^n possible values of $\boldsymbol{S}$. In contrast, for the loo-CMI, U can only take $n+1$ possible values—an exponential reduction in the number of cases that need to be considered.

With this notation in place, let us derive generalization bounds. For the change of measure step, there are no surprises: we can simply use, for instance, the Donsker-Varadhan variational representation of the relative entropy to replace the true joint distribution with one where U is independent from the other random variables. For the concentration of measure step, we need the following result.

Lemma 6.14. Assume that the range of the loss function $\ell(\cdot, \cdot)$ is $[0, 1]$. For all $\dot{\mathbf{\Lambda}}$ and $\lambda \in \mathbb{R}$,

$$\mathbb{E}_{P_U}\left[\exp\left(\lambda \text{loo-cv}(\dot{\mathbf{\Lambda}}, U) - \frac{\lambda^2 (n+1)^2}{8n^2}\right)\right] \leq 1. \tag{6.39}$$

This looks very similar to the exponential inequality for sub-Gaussian random variables from Definition 3.23, and can be used in an analogous way to derive generalization bounds. The proof of this result is given by Rammal *et al.* (2022). A simpler bound (of the same order), as used by Haghifam *et al.* (2022), can be obtained by simply noting that loo-cv$(\dot{\mathbf{\Lambda}}, U)$ is bounded to $[-1, 1]$ and using the fact that bounded random variables are sub-Gaussian. This gives

$$\mathbb{E}_{P_U}\left[\exp\left(\lambda \text{loo-cv}(\dot{\mathbf{\Lambda}}, U) - \frac{\lambda^2}{2}\right)\right] \leq 1, \tag{6.40}$$

which can also be directly obtained from (6.39) by using the fact that $(n+1)/n \leq 2$.

As before, by following the recipe from the proof of Corollary 4.2, we can obtain the following generalization bound.

Theorem 6.15. Assume that the range of the loss function $\ell(\cdot, \cdot)$ is $[0, 1]$. Then,

$$|\overline{\text{gen}}| \leq \frac{n+1}{n}\sqrt{\frac{I(\dot{\mathbf{\Lambda}}; U)}{2}}. \tag{6.41}$$

Proof. We begin from (6.39). By averaging over $P_{\dot{\mathbf{\Lambda}}}$, changing measure to $P_{\dot{\mathbf{\Lambda}}U}$, and using Jensen's inequality, we get, for $\lambda > 0$,

$$\mathbb{E}_{P_{\dot{\mathbf{\Lambda}}U}}\left[\text{loo-cv}(\dot{\mathbf{\Lambda}}, U)\right] \leq \frac{\lambda(n+1)^2}{8^2} + \frac{I(\dot{\mathbf{\Lambda}}; U)}{\lambda}. \tag{6.42}$$

As previously mentioned, the average of the leave-one-out cross-validation loss is the generalization gap. The result follows by optimizing over λ and repeating the argument for $\lambda < 0$. □

Unlike most generalization bounds that we reviewed so far, the result in Theorem 6.15 does not decay with n (ignoring the n-dependence of the information measure). Since U can take at most $n+1$ values, a trivial upper bound on the evaluated loo-MI is $I(\dot{\mathbf{\Lambda}}; U) \leq \log(n+1)$, so the bound could grow logarithmically with n in the worst case.

Finally, we present a bound for interpolating learning algorithms in terms of the evaluated loo-MI due to Haghifam *et al.* (2022). As for Theorem 6.13, this result can be derived through an argument that, essentially, just uses the Shannon capacity formula of a suitably chosen discrete memoryless communication channel. The proof below, which follows the one of Haghifam *et al.* (2022), proceeds without explicit reference to such a channel, and instead simply relies on the manipulation of information-theoretic quantities.

Theorem 6.16. Assume that the loss function is binary, meaning that for all $w \in \mathcal{W}$ and $z \in \mathcal{Z}$, $\ell(w, z) \in \{0, 1\}$. Consider an interpolating learning algorithm, so that the training loss $\hat{L} = 0$. Then,

$$L = \frac{I(\dot{\mathbf{\Lambda}}; U)}{\log(n+1)}. \tag{6.43}$$

Proof. To prove this result, we will simply compute the entropies in the decomposition $I(\dot{\mathbf{\Lambda}}; U) = H(\dot{\mathbf{\Lambda}}) - H(\dot{\mathbf{\Lambda}}|U)$. For $i \in [n+1]$, let $0^{(i)}$ denote the $n+1$ vector with $0^{(i)}_j = 0$ for $j \neq i$ and $0^{(i)}_i = 1$, and let $0^{(0)}$ denote the all-zeros vector of size $n+1$. Since the learning algorithm is interpolating, $\dot{\mathbf{\Lambda}}$ can incur a loss for at most one element of $\dot{\boldsymbol{Z}}$, and hence, the support of $\dot{\mathbf{\Lambda}}$ is the set $\{0^{(i)} : i \in \{0, \dots, n+1\}\}$. For $i > 0$, $P[\dot{\mathbf{\Lambda}} = 0^{(i)}]$ is the probability of not training on the ith sample times the probability of incurring a loss on that sample if it is not used for training—*i.e.*, the test loss. Hence,

$$P[\dot{\mathbf{\Lambda}} = 0^{(i)}] = \frac{1}{n+1} P[\dot{\mathbf{\Lambda}} = 0^{(i)} | U = i] = \frac{L}{n+1}. \tag{6.44}$$

Furthermore, $P[\dot{\mathbf{\Lambda}} = 0^{(0)}]$ is the probability of not incurring a loss on the test sample, *i.e.*, $1 - L$. Hence, we can calculate the entropy of $\dot{\mathbf{\Lambda}}$ as

$$\begin{aligned} H(\dot{\mathbf{\Lambda}}) &= -\sum_{i=0}^{n+1} P[\dot{\mathbf{\Lambda}} = 0^{(i)}] \log\Big(P[\dot{\mathbf{\Lambda}} = 0^{(i)}]\Big) && (6.45)\\ &= -(1-L)\log(1-L) - L\log\left(\frac{L}{n+1}\right). && (6.46) \end{aligned}$$

Through a similar calculation, we get

$$H(\dot{\mathbf{\Lambda}}|U) = -(1-L)\log(1-L) - L\log(L)\,. \tag{6.47}$$

Putting it all together, we find that

$$I(\dot{\mathbf{\Lambda}}; U) = H(\dot{\mathbf{\Lambda}}) - H(\dot{\mathbf{\Lambda}}|U) = L\log(n+1), \tag{6.48}$$

from which the result follows. □

As noted after Theorem 6.13, this result can be leveraged to obtain a bound for generic bounded losses.

We thus have two characterizations of the binary loss of interpolating learning algorithms that hold with *equality*, both from Theorem 6.13 and Theorem 6.16. Hence, it follows that the two characterizations must be equivalent, implying that

$$\frac{1}{n}\sum_{i=1}^{n} \frac{I(\Delta_i; S_i)}{\log 2} = \frac{I(\dot{\mathbf{\Lambda}}; U)}{\log(n+1)}. \tag{6.49}$$

This dual perspective may be beneficial in specific applications, making it possible to choose whichever representation is easier to analyze.

6.7 Bibliographic Remarks and Additional Perspectives

The underlying concept of the CMI framework can be traced back to the work of Audibert (2004) under the name of almost exchangeable priors. Specifically, a function Q on $\mathcal{Z}^{2n}$ is almost exchangeable if, for any permutation π such that $\{\pi(i), \pi(i+n)\} = \{i, i+n\}$ for all $i \in [n]$, it satisfies (Audibert, 2004, Definition 1.1)

$$Q(\tilde{Z}'_1, \dots, \tilde{Z}'_{2n}) = Q(\tilde{Z}'_{\pi(1)}, \dots, \tilde{Z}'_{\pi(2n)}) \tag{6.50}$$

for any $\tilde{\boldsymbol{Z}}' \in \mathcal{Z}^{2n}$. Given a $\tilde{\boldsymbol{Z}}' = (\boldsymbol{Z}, \boldsymbol{Z}') \in \mathcal{Z}^{2n}$, where the first n samples are the training set and the last n are n independent samples (a "ghost sample"), an almost exchangeable prior is an almost exchangeable function of $\tilde{\boldsymbol{Z}}'$. Thus, it has to be invariant to permutations that swap n-separated pairs, *i.e.*, the ith training sample with the ith independent ghost sample. This ensures that the prior does not encode knowledge of which samples are in the training set. Such priors were also used by Catoni (2007) in the transductive learning setting.

This is equivalent to the CMI prior $Q_{W|\tilde{\boldsymbol{Z}}}$, from the independently constructed CMI framework of Steinke and Zakynthinou (2020). Specifically, in both cases, the prior is allowed to depend on the training set and a ghost sample, as long as it cannot tell one from the other. There is a very slight difference in formulation between the two settings: for almost exchangeable priors, $\tilde{\boldsymbol{Z}}'$ is ordered so that the training set comes first, followed by the ghost sample, and the function is required to be invariant to permutations within pairs. For the CMI framework, the n-separated pairs of $\tilde{\boldsymbol{Z}}$ are instead randomly assigned, which directly eliminates the information of which samples are in the training set. Hence, the CMI prior is allowed to depend in an arbitrary way on $\tilde{\boldsymbol{Z}}$. In the end, both formulations lead to equivalent results. We remark once more, though, that the construction of Steinke and Zakynthinou (2020) was formulated independently and with a different motivation.

Many of the results from this chapter were introduced already in the work of Steinke and Zakynthinou (2020), namely Theorems 6.1 to 6.3; the extension to unbounded losses in (6.6); and the concept of e-CMI from Section 6.5. Bounds in terms of a generic convex function, and the specific bound from Theorem 6.4, can be found in (Hellström and Durisi, 2022a). The use of disintegration and the random-subset technique for the CMI framework, as in Theorem 6.5, was introduced by Haghifam *et al.* (2020), later extended in the form of Theorem 6.6 independently by Rodríguez-Gálvez *et al.* (2020) and Zhou *et al.* (2021).

The tail bounds in Theorems 6.7 and 6.9 can be found in Hellström and Durisi (2020a). As mentioned, the PAC-Bayesian bound in Theorem 6.8 is a heavily simplified version of the result of Grünwald *et al.* (2021, Thm. 1). This result, which gives fast-rate CMI-flavored PAC-Bayesian bounds under the Bernstein condition, is significant in how is leads to non-trivial fast-rate bounds in terms of the VC dimension, which was previously shown to be impossible for standard PAC-Bayesian bounds (Livni and Moran, 2017). We will discuss this further in Section 7.3. The single-draw bounds in Theorem 6.10 and Corollary 6.11 can be found in Hellström and Durisi (2020a), and are extensions of bounds from Esposito *et al.* (2021a) to the CMI setting.

As pointed out, the concept of e-CMI was introduced by Steinke and Zakynthinou (2020). This was studied further by Haghifam *et al.* (2022), Harutyunyan *et al.* (2021), Hellström and Durisi (2022a), Rammal *et al.* (2022), and Wang and Mao (2023c), who extended the original bounds in various ways. The form given in Theorem 6.12 is due to Wang and Mao (2023c), as is the result in Theorem 6.13. The extension to leave-one-out CMI, discussed in Section 6.6, was introduced essentially simultaneously by Haghifam *et al.* (2022) and Rammal *et al.* (2022). The result in Theorem 6.15 is due to Rammal *et al.* (2022), while Theorem 6.16 is from Haghifam *et al.* (2022). The information-theoretic approach to generalization was combined with techniques from algorithmic stability by Wang and Mao (2023b), leading to improved bounds for certain stochastic convex optimization problems. Recently, Sachs *et al.* (2023) derived bounds in terms of an algorithm-dependent Rademacher complexity, which is conceptually similar to the CMI framework. Finally, Sefidgaran *et al.* (2023) used related ideas, combined with the information bottleneck and the minimum description length principle, to obtain generalization bounds for representation learning.

Part II

Applications and Additional Topics

7

The Information Complexity of Learning Algorithms

As argued in Section 2.2, one of the benefits of the information-theoretic approach to analyzing generalization is that the resulting bounds depend on both the learning algorithm and the data distribution. This is in contrast to the uniform convergence-flavored bounds of Section 1.3, *i.e.*, bounds that hold uniformly over all data distributions, or even uniformly over all hypotheses. Still, this is not very useful if we cannot compute or bound the information measures that appear in the information-theoretic generalization bounds.

In this chapter, we study these information measures for specific learning algorithms. We begin by looking at the Gibbs posterior, which naturally emerges as the minimizer of some PAC-Bayesian bounds, and whose generalization error can be exactly characterized via a symmetrized relative entropy. After this, we discuss the Gaussian location model, wherein the learner aims to estimate the mean of a Gaussian distribution. This simple setting allows us to exactly evaluate the training and population losses, as well as several information measures, and thus allows us to compare various bounds for a concrete setting. Next, we consider the VC dimension, which plays a fundamental role in uniform convergence-flavored generalization bounds, as well as bounds for compression schemes. It can be shown that, in many cases, such uniform convergence-flavored bounds can (essentially) be recovered from the information-theoretic bounds from the previous chapters. We refer to this property as the *expressiveness* of the bounds—*i.e.*, the extent to which the information-theoretic bounds are able to express results from alternative frameworks. Finally, we discuss connections to algorithmic stability and privacy measures. We postpone applications to neural networks and gradient-based algorithms, such as stochastic gradient descent and stochastic gradient Langevin dynamics, to Chapter 8.

7.1 The Gibbs Posterior

Given a generalization bound, it is tempting to design a learning algorithm so as to minimize it. So far, when presenting information-theoretic bounds, we have considered a specific learning algorithm, characterized in terms of a posterior $P_{W|\boldsymbol{Z}}$. Given this posterior, we mostly focused on the prior given by the marginal distribution P_W, as this typically minimizes the bounds in expectation. However, a slightly different approach is possible, as we exemplified when discussing PAC-Bayesian bounds in Section 5.2. There, we discussed bounds that hold for any prior and posterior. Crucially, the bounds based on the Donsker-Varadhan variational representation of the relative entropy in Theorem 3.17 actually hold simultaneously for *all* posteriors. This is because of the supremum over P in (3.34). This implies that for a fixed prior, we can choose the posterior that minimizes the bound.

Of particular relevance is the Gibbs posterior. Given a prior Q_W, a training loss $L_{\boldsymbol{Z}}(W)$, a parameter λ referred to as the inverse temperature, the Gibbs posterior for any measurable set $\mathcal{E} \subseteq \mathcal{W}$ is given by

$$P^G_{W|\boldsymbol{Z}}(\mathcal{E}|\boldsymbol{Z}) = \frac{\int_{\mathcal{E}} \exp(-\lambda L_{\boldsymbol{Z}}(w))\,\mathrm{d}Q_W(w)}{\int_{\mathcal{W}} \exp(-\lambda L_{\boldsymbol{Z}}(w))\,\mathrm{d}Q_W(w)}. \tag{7.1}$$

The normalization constant in the denominator, referred to as the *partition function*, is a random variable that depends on $\boldsymbol{Z}$. This terminology comes from statistical physics, where the Gibbs posterior also appears under the name of Boltzmann distribution. For later use, it will be convenient to define the *log-partition function*

$$\Psi_\lambda(\boldsymbol{Z}) = \log \int_{\mathcal{W}} \exp(-\lambda L_{\boldsymbol{Z}}(w))\, \mathrm{d}Q_W(w). \tag{7.2}$$

The relevance of the Gibbs posterior is that it is the minimizer of many PAC-Bayesian bounds. Specifically, we have the following result, which is a simple consequence of the Donsker-Varadhan variational representation of the relative entropy applied conditionally on $\boldsymbol{Z}$.

Lemma 7.1. Let the prior Q_W be given. Then, for any $P_{W|\boldsymbol{Z}}$,

$$\mathbb{E}_{P_{W|\boldsymbol{Z}}}[L_{\boldsymbol{Z}}(W)] + \frac{D(P_{W|\boldsymbol{Z}} \,||\, Q_W)}{\lambda} \geq -\frac{1}{\lambda}\Psi_\lambda(\boldsymbol{Z}), \tag{7.3}$$

and equality is achieved uniquely by the Gibbs posterior $P^G_{W|\boldsymbol{Z}}$.

The inverse temperature parameter λ controls the trade-off between the influence of the prior and the influence of the data, and the relative entropy $D(P_{W|\boldsymbol{Z}} \,||\, Q_W)$ acts as a regularizer. On the one hand, when $\lambda \to \infty$, we completely ignore this regularizer and perform unfettered empirical risk minimization. On the other hand, if $\lambda \to 0$, the optimal posterior equals the prior, and we pay no mind to the collected data. In PAC-Bayesian bounds such as (5.14), the inverse temperature is typically chosen to be proportional to n. This leads to a very sensible trade-off: when the amount of data is small, we are not easily convinced to stray far from the prior. However, when the amount of data grows large, we are inclined to place more importance on it, without relying much on the prior.

Lemma 7.1 can be used to obtain bounds on the average generalization error of the Gibbs posterior. To that end, we start with a simple observation based on Corollary 4.2 and the identity

$$\inf_{\lambda>0} \left(a\lambda + \frac{b}{\lambda} \right) = 2\sqrt{ab}. \tag{7.4}$$

Suppose that $\ell(w, Z)$ is σ-subgaussian for all $w \in \mathcal{W}$. Then, for any $P_{W|\boldsymbol{Z}}$ and any $\lambda > 0$,

$$\mathbb{E}[L_{P_Z}(W)] \leq \mathbb{E}[L_{\boldsymbol{Z}}(W)] + \frac{I(W; \boldsymbol{Z})}{\lambda} + \frac{\lambda\sigma^2}{2n}. \tag{7.5}$$

It is tempting to use this inequality to construct a learning algorithm with small expected population loss as follows: fix the inverse temperature $\lambda > 0$ and then choose $P_{W|\boldsymbol{Z}}$ to minimize the right-hand side of (7.5). However, the mutual information $I(W;\boldsymbol{Z})$ depends on both $P_{W|\boldsymbol{Z}}$ and on the marginal distribution P_Z, while the learning algorithm has to be designed without knowledge of P_Z. This can be solved by relaxing the bound using the so-called *golden formula* for the mutual information: for any $Q_W \ll P_W$, we have (Csiszar and Körner, 2011, Eq. (8.7))

$$I(W;\boldsymbol{Z}) = D(P_{W|\boldsymbol{Z}} \| Q_W | P_{\boldsymbol{Z}}) - D(P_W \| Q_W). \tag{7.6}$$

Using this, along with the fact that the relative entropy is nonnegative, we can weaken (7.5) to

$$\begin{aligned} \mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{P_Z}(W)] &\le \mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{\boldsymbol{Z}}(W)] + \frac{D(P_{W|\boldsymbol{Z}} \| Q_W | P_{\boldsymbol{Z}})}{\lambda} + \frac{\lambda\sigma^2}{2n} \qquad (7.7)\\ &= \mathbb{E}_{P_{\boldsymbol{Z}}}\left[\mathbb{E}_{P_{W|\boldsymbol{Z}}}[L_{\boldsymbol{Z}}(W)] + \frac{D(P_{W|\boldsymbol{Z}} \| Q_W)}{\lambda}\right] + \frac{\lambda\sigma^2}{2n}. \qquad (7.8) \end{aligned}$$

Thus, applying Lemma 7.1 conditionally on $\boldsymbol{Z}$, we arrive at the following.

Theorem 7.2. Assume $\ell(w, Z)$ is σ-subgaussian under P_Z for all $w \in \mathcal{W}$. Then, the expected population loss of the Gibbs posterior $P^G_{W|\boldsymbol{Z}}$ at inverse temperature λ satisfies

$$\mathbb{E}[L_{P_Z}(W)] \le -\frac{1}{\lambda}\,\mathbb{E}[\Psi_\lambda(\boldsymbol{Z})] + \frac{\lambda\sigma^2}{2n}. \tag{7.9}$$

Bounds of this sort are common in the PAC-Bayes literature (Catoni, 2007; McAllester, 1998, 1999; Zhang, 2006). To instantiate them in a given setting, we need lower bounds on the log-partition function $\Psi_\lambda(\boldsymbol{Z})$, which are typically derived on a case-by-case basis. As an example, we give the following result, due to Pensia *et al.* (2018).

Theorem 7.3. Assume the following:

1. The hypothesis space $\mathcal{W}$ is the d-dimensional Euclidean space $\mathbb{R}^d$.

2. The loss function $\ell(w, z)$ is differentiable in w, and its gradient $\nabla\ell(w, z)$ with respect to w is Lipschitz-continuous uniformly in z,

that is, there exists a constant $M > 0$, such that for all $w, w' \in \mathcal{W}$

$$\sup_{z \in \mathcal{Z}} \|\nabla \ell(w, z) - \nabla \ell(w', z)\| \le M \|w - w'\| \tag{7.10}$$

where $\| \cdot \|$ denotes the Euclidean (ℓ^2) norm on $\mathbb{R}^d$.

3. For every realization of $\boldsymbol{Z}$, all global minimizers of the training loss $L_{\boldsymbol{Z}}(W)$ lie in the ball of radius R centered at 0.

4. The loss $\ell(w, Z)$ is σ-subgaussian under P_Z for all $w \in \mathcal{W}$.

Let $P^G_{W|\boldsymbol{Z}}$ be the Gibbs posterior with inverse temperature $\lambda > 0$ associated to the Gaussian prior $Q_W = \mathcal{N}(0, \rho^2 I_d)$. Then

$$\begin{aligned} &\mathbb{E}[L_{P_Z}(W)] - \min_{w \in \mathcal{W}} L_{P_Z}(w) \\ &\le \frac{M\pi\rho^2 d}{\lambda} + \frac{1}{2\lambda\rho^2}\left(R + \sqrt{\frac{2\pi\rho^2 d}{\lambda}}\right)^2 + \frac{d}{2\lambda}\log\frac{\lambda}{d} - \frac{1}{\lambda}\log V_d + \frac{\lambda\sigma^2}{2n}, \end{aligned} \tag{7.11}$$

where V_d is the volume of the unit ball in $(\mathbb{R}^d, \| \cdot \|)$.

Proof. Fix $\boldsymbol{Z}$ and let $w^*_{\boldsymbol{Z}}$ be any global minimizer of $L_{\boldsymbol{Z}}(W)$, where $\|w^*_{\boldsymbol{Z}}\| \le R$ by hypothesis. Since the gradient $w \mapsto \nabla \ell(w, \boldsymbol{Z})$ is M-Lipschitz and $\nabla L_{\boldsymbol{Z}}(w^*_{\boldsymbol{Z}}) = 0$, we have

$$L_{\boldsymbol{Z}}(w) - L_{\boldsymbol{Z}}(w^*_{\boldsymbol{Z}}) \le \frac{M}{2}\|w - w^*_{\boldsymbol{Z}}\|^2. \tag{7.12}$$

Therefore,

$$\Psi_\lambda(\boldsymbol{Z}) = -\lambda L_{\boldsymbol{Z}}(w^*_{\boldsymbol{Z}}) + \log \mathbb{E}_{Q_W}\left[\exp\left(-\lambda\left(L_{\boldsymbol{Z}}(W) - L_{\boldsymbol{Z}}(w^*_{\boldsymbol{Z}})\right)\right)\right] \tag{7.13}$$

$$\ge -\lambda L_{\boldsymbol{Z}}(w^*_{\boldsymbol{Z}}) + \log \mathbb{E}_{Q_W}\left[\exp\left(-\frac{\lambda M}{2}\|W - w^*_{\boldsymbol{Z}}\|^2\right)\right], \tag{7.14}$$

so, in order to lower-bound the log-partition function $\Psi_\lambda(\boldsymbol{Z})$, we need to lower-bound the Gaussian integral

$$G = \frac{1}{(2\pi\rho^2)^{d/2}} \int_{\mathbb{R}^d} e^{-\frac{1}{2\rho^2}\|w\|^2} e^{-\frac{\lambda M}{2}\|w - w^*_{\boldsymbol{Z}}\|^2}\, \mathrm{d}w. \tag{7.15}$$

Let $\mathcal{B}$ be the ℓ^2 ball of radius $\varepsilon > 0$ (to be tuned later) centered at $w^*_{\boldsymbol{Z}}$ with volume $\mathsf{Vol}_d(\mathcal{B})$. Then

$$\begin{aligned}
G &\geq \frac{1}{(2\pi\rho^2)^{d/2}} e^{-\frac{\lambda M \varepsilon^2}{2}} \cdot \int_{\mathcal{B}} e^{-\frac{1}{2\rho^2}\|w\|^2} \,\mathrm{d}w \\
&\geq \frac{1}{(2\pi\rho^2)^{d/2}} e^{-\frac{\lambda M \varepsilon^2}{2}} \cdot e^{-\frac{1}{2\rho^2}(\|w^*_{\boldsymbol{Z}}\|+\varepsilon)^2} \mathsf{Vol}_d(\mathcal{B}) \\
&= \frac{1}{(2\pi\rho^2)^{d/2}} e^{-\frac{\lambda M \varepsilon^2}{2}} \cdot e^{-\frac{1}{2\rho^2}(\|w^*_{\boldsymbol{Z}}\|+\varepsilon)^2} \varepsilon^d V_d \\
&\geq \left(\frac{\varepsilon^2}{2\pi\rho^2}\right)^{d/2} \exp\left(-\frac{\lambda M \varepsilon^2}{2} - \frac{1}{2\rho^2}(R+\varepsilon)^2\right) V_d.
\end{aligned}$$

For all $\varepsilon > 0$, this leads to the estimate

$$-\frac{1}{\lambda}\mathbb{E}[\Psi_\lambda(\boldsymbol{Z})] \leq \mathbb{E}\left[\min_{w\in\mathcal{W}} L_{\boldsymbol{Z}}(W)\right] \tag{7.16}$$

$$+\frac{M\varepsilon^2}{2} + \frac{1}{2\lambda\rho^2}(R+\varepsilon)^2 + \frac{d}{2\lambda}\log\left(\frac{2\pi\rho^2}{\varepsilon^2}\right) - \frac{1}{\lambda}\log V_d, . \tag{7.17}$$

Choosing $\varepsilon = \frac{2\pi\rho^2 d}{\lambda}$ and using that

$$\mathbb{E}\left[\min_{w\in\mathcal{W}} L_{\boldsymbol{Z}}(W)\right] = \mathbb{E}[L_{\boldsymbol{Z}}(w^*_{\boldsymbol{Z}})] \leq \min_{w\in\mathcal{W}} L_{P_Z}(w), \tag{7.18}$$

we get (7.11). □

Recently, Aminian *et al.* (2021a) provided an exact information-theoretic characterization of the average generalization error of the Gibbs posterior. Let $P^G_W = \mathbb{E}_{P_{\boldsymbol{Z}}}\left[P^G_{W|\boldsymbol{Z}}\right]$ denote the marginal distribution on W induced by the Gibbs posterior. Then, for the Gibbs posterior, we let the symmetrized KL information between W and $\boldsymbol{Z}$ be given by

$$I_{\mathrm{SKL}}(W;\boldsymbol{Z}) = D(P_{\boldsymbol{Z}}P^G_{W|\boldsymbol{Z}} \,||\, P_{\boldsymbol{Z}}P^G_W) + D(P_{\boldsymbol{Z}}P^G_W \,||\, P_{\boldsymbol{Z}}P^G_{W|\boldsymbol{Z}}). \tag{7.19}$$

This symmetrized relative entropy, where we sum two relative entropies with their arguments swapped, is sometimes referred to as Jeffreys' divergence. Notice that the term $D(P_{\boldsymbol{Z}}P^G_{W|\boldsymbol{Z}} \,||\, P_{\boldsymbol{Z}}P^G_W)$ is the mutual information $I(W;\boldsymbol{Z})$ while the term $D(P_{\boldsymbol{Z}}P^G_W \,||\, P_{\boldsymbol{Z}}P^G_{W|\boldsymbol{Z}})$ is sometimes referred to as the lautum information (Palomar and Verdu, 2008).[1] With

[1] This provides a strong incitement to refer to $I_{\mathrm{SKL}}(\cdot;\cdot)$ as the mutualautum information, but we digress.

this, Aminian *et al.* (2021a) derived the following exact characterization of the average generalization error of the Gibbs posterior.

Theorem 7.4. Given an inverse temperature λ and a prior distribution Q_W, the average generalization error of the Gibbs posterior is given by

$$\mathbb{E}_{P_{\boldsymbol{Z}} P^G_{W|\boldsymbol{Z}}}[L_{P_Z}(W) - L_{\boldsymbol{Z}}(W)] = \frac{I_{\mathrm{SKL}}(W;\boldsymbol{Z})}{\lambda}. \tag{7.20}$$

Proof. Note that $\mathbb{E}_{P_{\boldsymbol{Z}} P^G_{W|\boldsymbol{Z}}}\left[\log P^G_W\right] = \mathbb{E}_{P_{\boldsymbol{Z}} P^G_W}\left[\log P^G_W\right]$. Hence, using (7.19), we can write

$$\begin{aligned} I_{\mathrm{SKL}}(W;\boldsymbol{Z}) &= \mathbb{E}_{P_{\boldsymbol{Z}} P^G_{W|\boldsymbol{Z}}}\left[\log \frac{P^G_{W|\boldsymbol{Z}}}{P^G_W}\right] + \mathbb{E}_{P_{\boldsymbol{Z}} P^G_W}\left[\log \frac{P^G_W}{P^G_{W|\boldsymbol{Z}}}\right] &(7.21)\\ &= \mathbb{E}_{P_{\boldsymbol{Z}} P^G_{W|\boldsymbol{Z}}}\left[\log P^G_{W|\boldsymbol{Z}}\right] - \mathbb{E}_{P_{\boldsymbol{Z}} P^G_W}\left[\log P^G_{W|\boldsymbol{Z}}\right]. &(7.22)\end{aligned}$$

From the definition of the Gibbs posterior, we see that

$$\log P^G_{W|\boldsymbol{Z}}(W|\boldsymbol{Z}) = \log Q_W(W) - \Psi_\lambda(\boldsymbol{Z}) - \lambda L_{\boldsymbol{Z}}(W). \tag{7.23}$$

Since the marginal distributions of W and $\boldsymbol{Z}$ are the same under $P_{\boldsymbol{Z}} P^G_{W|\boldsymbol{Z}}$ and $P_{\boldsymbol{Z}} P^G_W$ we have

$$\mathbb{E}_{P_{\boldsymbol{Z}} P^G_{W|\boldsymbol{Z}}}[\log Q_W(W) - \Psi_\lambda(\boldsymbol{Z})] = \mathbb{E}_{P_{\boldsymbol{Z}} P^G_W}[\log Q_W(W) - \Psi_\lambda(\boldsymbol{Z})]. \tag{7.24}$$

From this, it follows that

$$\begin{aligned} &\mathbb{E}_{P_{\boldsymbol{Z}} P^G_{W|\boldsymbol{Z}}}\left[\log P^G_{W|\boldsymbol{Z}}\right] - \mathbb{E}_{P_{\boldsymbol{Z}} P^G_W}\left[\log P^G_{W|\boldsymbol{Z}}\right] \\ =&\mathbb{E}_{P_{\boldsymbol{Z}} P^G_{W|\boldsymbol{Z}}}[-\lambda L_{\boldsymbol{Z}}(W)] - \mathbb{E}_{P_{\boldsymbol{Z}} P^G_W}[-\lambda L_{\boldsymbol{Z}}(W)] &(7.25)\\ =&\lambda\, \mathbb{E}_{P_{\boldsymbol{Z}} P^G_{W|\boldsymbol{Z}}}[L_{P_Z}(W) - L_{\boldsymbol{Z}}(W)]. &(7.26)\end{aligned}$$

From this, the result follows. □

In order to interpret this result, we need to discuss the extreme cases. First, if $\lambda \to \infty$, it may seem as if the generalization error vanishes. This is the case if $I_{\mathrm{SKL}}(W;\boldsymbol{Z})$ remains finite when we perform exact empirical risk minimization. For this to occur, we need not only that $P^G_{W|\boldsymbol{Z}} \ll P^G_W$, but also that $P^G_W \ll P^G_{W|\boldsymbol{Z}}$. Since the Gibbs posterior with infinite temperature is supported only on empirical risk minimizers, the second criterion can only be fulfilled if the prior is also supported only on empirical risk minimizers. For any non-trivial case, we expect the prior to assign some probability mass to non-minimizers as well, meaning that $I_{\mathrm{SKL}}(W;\boldsymbol{Z})$ would diverge as $\lambda \to \infty$. In a similar vein, when $\lambda \to 0$, the posterior does not change relative to the prior, so $I_{\mathrm{SKL}}(W;\boldsymbol{Z}) \to 0$ as well.

While the Gibbs posterior has many attractive properties theoretically, it is not always straightforward to implement in practice. This is discussed further by, for instance, Alquier *et al.* (2016) and Perlaza *et al.* (2023).

7.2 The Gaussian Location Model

We now turn to a simple learning problem in which many of the quantities in the generalization bounds that we discussed can be evaluated explicitly, allowing us to perform a direct comparison between different bounds for a concrete setting. Specifically, we assume that the data distribution $P_Z = \mathcal{N}(\mu, \sigma^2)$ is a Gaussian distribution with mean μ and variance σ^2, and the training set $\boldsymbol{Z} = (Z_1, \dots, Z_n) \in \mathbb{R}^n$ consists of n independent samples from P_Z. Based on this, the goal is to learn the mean of the Gaussian distribution. Thus, the hypothesis space consists of the real numbers $\mathcal{W} = \mathbb{R}$. A natural choice for the loss function, which we will consider throughout, is the squared loss $\ell(w,z) = (w-z)^2$. We will focus on the empirical risk minimizer obtained by taking the sample average, $W = \frac{1}{n}\sum_{i=1}^n Z_i$.

For this setting, the average generalization error can in fact be computed explicitly as (Bu *et al.*, 2020)

$$\overline{\mathrm{gen}} = \mathbb{E}_{P_{W\boldsymbol{Z}}}\left[\mathbb{E}_{Z' \sim P_Z}\left[(Z'-W)^2\right] - \frac{1}{n}\sum_{i=1}^n (Z_i - W)^2\right] \tag{7.27}$$

$$= \frac{2\sigma^2}{n}. \tag{7.28}$$

We thus have a known baseline with which to compare the generalization bounds that we derived in Chapters 4 and 6, and for this setting, many of them can be computed exactly. It should be noted here that if a bound gives a loose characterization of the generalization error for this specific problem, this is not an indictment of the bound as a whole. Since all of the bounds that we will discuss have been derived for a very general class of learning problems and learning algorithms, it is not unexpected that they will be loose for many specific problems and algorithms. Nevertheless, due to its analytical tractability, this setting serves as an instructive case study. Also note that, as mentioned in Section 7.1, the average generalization error of the Gibbs posterior is exactly characterized by the symmetrized KL information. By evaluating this information-theoretic quantity, one can show that the Gibbs posterior also has a generalization error of order σ^2/n. For more details, see the work of Aminian *et al.* (2022b).

First, we note that the mutual information $I(W;\boldsymbol{Z})$ gives a vacuous bound on the generalization gap. Indeed, since the training data and hypothesis are continuous and we use a deterministic learning algorithm, the mutual information is infinite. However, as noted by Bu *et al.* (2020), this can be rectified by using the individual-sample technique: since the hypothesis is not a deterministic function of any single sample, the individual-sample mutual information is finite. Indeed, it can be computed in closed form as (Bu *et al.*, 2020)

$$I(W;Z_i) = \frac{1}{2}\log\frac{n}{n-1}. \tag{7.29}$$

Inserting this into the generalization bound in Corollary 4.6, we find that

$$\begin{aligned}\overline{\text{gen}} &\le \frac{1}{n}\sum_{i=1}^{n}\sqrt{2\sigma^2 I(W;Z_i)} &(7.30)\\ &= \sigma\sqrt{\log\left(\frac{n}{n-1}\right)} &(7.31)\\ &\le \sigma\sqrt{\frac{1}{n-1}}. &(7.32)\end{aligned}$$

Thus, this gives a bound of order $1/\sqrt{n}$, which is quadratically worse than the true generalization gap.

Next, let us consider the CMI framework. To do this, one needs to go beyond the assumption of a bounded loss that was considered throughout most of Chapter 6. As indicated in (6.6), the main results extend to certain unbounded losses. This includes the squared loss under a Lipschitz condition, provided that the fourth moment of the data is finite (Steinke and Zakynthinou, 2020, Sec. 5.4). This is satisfied for the Gaussian location problem—see the work of Zhou *et al.* (2021) for details. While the CMI yields a finite result, unlike the mutual information, it is significantly looser than the individual-sample mutual information bound. Indeed, we have (Zhou *et al.*, 2021)

$$I(W; \boldsymbol{S}|\tilde{\boldsymbol{Z}}) = \frac{n}{\log_2(e)}. \tag{7.33}$$

The reason for this is that conditioning on the supersample reveals too much information, due to the continuous nature of the output. In fact, if we consider a naïve individual-sample version of the CMI, where we still condition on the full supersample, that is, $I(W; S_i|\tilde{\boldsymbol{Z}})$, we still get a constant—leading to a generalization bound that does not decay with n. Motivated by this, Zhou *et al.* (2021) argue for the individually conditioned CMI, where the conditioning is also on individual pairs of the supersample—as discussed in Theorem 6.6. With this, it can be shown that (Zhou *et al.*, 2021, Lemma. 4)

$$I(W; S_i|Z_i = z_i, Z_{i+n} = z_{i+n}) = \frac{(z_i - z_{i+n})^2}{8\sigma^2(n-1)} + o\left(\frac{1}{n}\right). \tag{7.34}$$

Inserting this into the corresponding generalization bound of Zhou *et al.* (2021), we again get a bound that decays as $1/\sqrt{n}$, but with a slightly improved constant factor.

This raises the question: is it possible to obtain the correct $1/n$-dependence from information-theoretic generalization bounds? The answer turns out to be yes. Through the use of stochastic chaining, as mentioned in Section 4.4, Zhou *et al.* (2022, Sec. 4.1) obtained a generalization bound of $\overline{\text{gen}} \leq 13\sigma^2/n$, thus matching the dependence of the true generalization error but with a larger constant. An alternative approach was taken by Wu *et al.* (2022b), who derived a bound that, on its face, is identical to the individual-sample bound of Bu *et al.* (2020), but with a key modification—instead of assuming the loss to be sub-Gaussian, the *excess risk*, $r(w, Z) = \ell(w, Z) - \ell(w^*, Z)$, is assumed to be sub-Gaussian under P_Z for all $w \in \mathcal{W}$, where w^* is a minimizer of the population loss. For sufficiently large n, the excess risk of the Gaussian location problem with the sample-averaging algorithm actually turns out to be $\sqrt{4\sigma^4/n}$-sub-Gaussian—the sub-Gaussianity parameter decays with n. Evaluating the generalization bound with this yields an $O(1/n)$ rate.

However, it is possible to demonstrate that this fast rate is achievable with arguably simpler techniques. In fact, it turns out that it is possible to derive an information-theoretic generalization bound that is exactly tight for this problem, even up to constants, which was done by Zhou *et al.* (2023a). This is achieved through a variant of the individual-sample approach of Bu *et al.* (2020), with some key modifications: the change of measure is applied to the generalization gap rather than the training loss; disintegration is used; a different prior than the true marginal is used; and the straight-forward sample-averaging algorithm is replaced with a weighted one where Gaussian noise is added (which has the same performance as the sample-averaging algorithm in expectation). This includes many of the techniques that we covered in Chapter 4, applied in a very careful way. If we are satisfied with a bound that is optimal only in an asymptotic sense, the alternative prior and weighted sample-averaging are not needed. The interested reader is referred to the work of Zhou *et al.* (2023a) for the full details.

7.3 The VC Dimension

As discussed in Section 1.3.1, a fundamental quantity that characterizes distribution- and algorithm-independent learnability for binary classification is the VC dimension. While our original motivation for pursuing information-theoretic generalization bounds was to go beyond this style of uniform convergence analysis, an interesting question is whether or not the information-theoretic approach is still expressive enough to capture complexity measures such as the VC dimension. More precisely, we seek to answer the following question: consider a hypothesis class $\mathcal{W}$ with bounded VC dimension d_{VC}. Can we provide a bound on the information measures that appear in our generalization bounds in terms of d_{VC}, and if so, do the resulting bounds coincide with the best available generalization bounds?

To partially answer this question, we will focus on the case of generalization bounds in expectation and consider binary classification with the $0-1$ loss. Throughout, we assume that the instance space $\mathcal{Z}$ factors into a feature space $\mathcal{X}$ and label space $\mathcal{Y} = \{0, 1\}$, and we associate each hypothesis $w \in \mathcal{W}$ with a function $f_w : \mathcal{X} \to \mathcal{Y}$.

7.3.1 Mutual Information

We begin by considering the mutual information between the training data $\boldsymbol{Z}$ and hypothesis W, $I(W; \boldsymbol{Z})$, that appears in, *e.g.*, Corollary 4.2. As an illustrative example of a class with finite VC dimension, we consider threshold classifiers: that is, the set of classifiers is given by $\{f_w(x) = 1\{x \geq w\} \,|\, w \in \mathbb{R}\}$. As this hypothesis class can induce arbitrary labels for a set with a single element, but not a set with two elements (as achieving $f_w(x_1) = 1$ and $f_w(x_2) = 0$ for $x_1 < x_2$ is not possible), its VC dimension is one. Throughout, we shall refer to data distributions for which an element of the hypothesis class achieves zero population loss as *realizable*.

Immediately, we can establish one negative result: the mutual information $I(W; \boldsymbol{Z})$ can be unbounded, even for very reasonable empirical risk minimizers. Consider, for instance, the case of threshold classifiers for a realizable distribution. Let us denote each training sample as $Z_i = (X_i, Y_i)$, which consists of a real number feature X_i and a label $Y_i \in \{0, 1\}$. A reasonable empirical risk minimizer is an algorithm that outputs $f_{\hat{W}}$, where $\hat{W} = \min\{x : (x, 1) \in \boldsymbol{Z}\}$, *i.e.*, the smallest feature labelled 1. Due to the realizability assumption, this must achieve zero training loss. However, since the learning algorithm is a deterministic function of the training set with a continuous output, $I(W; \boldsymbol{Z}) = \infty$.

In order to circumvent this, Xu and Raginsky (2017) considered the following two-stage algorithm. First, split the training set into two halves, so that $\boldsymbol{Z}_a = (Z_1, \ldots, Z_{n/2})$ and $\boldsymbol{Z}_b = (Z_{n/2+1}, \ldots, Z_n)$, where we assume n to be even for simplicity. In the first stage of the algorithm, one constructs an empirical cover of $\mathcal{W}$ on the basis of $\boldsymbol{X}_a = (X_1, \ldots, X_{n/2})$, *i.e.*, a subset $\mathcal{W}_a \subset \mathcal{W}$ such that $\left|\{(f_w(X_1), \ldots, f_w(X_{n/2})) : w \in \mathcal{W}_a\}\right| = |\mathcal{W}_a|$, meaning that each element of $\mathcal{W}_a$ induces a distinct classification, and $\left|\{(f_w(X_1), \ldots, f_w(X_{n/2})) : w \in \mathcal{W}\}\right| = |\mathcal{W}_a|$, meaning that each possible classification using $\mathcal{W}$ is induced by an element of $\mathcal{W}_a$. In the second stage of the algorithm, one selects an empirical risk minimizer for $\boldsymbol{Z}_b$ from the finite $\mathcal{W}_a$. By applying Corollary 4.2 conditional on $\boldsymbol{Z}_a$, evaluating the training loss with respect to $\boldsymbol{Z}_b$, we thus find that

$$\mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{P_Z}(W)-L_{\boldsymbol{Z}_b}(W)]=\mathbb{E}_{P_{\boldsymbol{Z}_a}}\left[\mathbb{E}_{P_{W\boldsymbol{Z}_b|\boldsymbol{Z}_a}}[L_{P_Z}(W)-L_{\boldsymbol{Z}_b}(W)]\right] \tag{7.35}$$

$$\leq \sqrt{\frac{I(W;\boldsymbol{Z}_b|\boldsymbol{Z}_a)}{n}}, \tag{7.36}$$

where we used the fact that the $0-1$ loss is $1/2$-sub-Gaussian. Now, given $\boldsymbol{Z}_a$, W can only take values in the finite set $\mathcal{W}_a$. Furthermore, the cardinality of $\mathcal{W}_a$ can be bounded using the Sauer-Shelah lemma (Lemma 1.3). We thus conclude that

$$I(W;\boldsymbol{Z}_b|\boldsymbol{Z}_a) \leq H(W|\boldsymbol{Z}_a) \leq \log(|\mathcal{W}_a|) \leq d_{\mathrm{VC}} \log\left(\frac{en}{2d_{\mathrm{VC}}}\right), \tag{7.37}$$

where the first step follows from the non-negativity of entropy, the second step from the fact that entropy is maximized by a uniform distribution, and the final step from the Sauer-Shelah lemma. Note that, through these arguments, we have obtained an average version of the standard generalization guarantee in terms of the VC dimension from Theorem 1.4, up to constants and logarithmic dependencies. Still, this applies only to a very particular algorithm, and not the standard empirical risk minimizer. Indeed, Bassily *et al.* (2018) and Nachum *et al.* (2018) showed that for any empirical risk minimizer over a finite input space, there exists a realizable data distribution for which the mutual information $I(W;\boldsymbol{Z})$ scales with the cardinality of the input space. Furthermore, Livni and Moran (2017) demonstrated that for any learning algorithm for threshold classifiers, there exists a realizable distribution for which either the population loss or the mutual information is large (in fact, their result applies more generally to the relative entropy that appears in PAC-Bayesian bounds). On the positive side, Nachum and Yehudayoff (2019) showed that there does exist learning algorithms with bounded mutual information for "most" hypotheses in VC classes.

7.3.2 Conditional Mutual Information

We now turn to the CMI framework of Chapter 6. Specifically, we consider the conditional mutual information between the hypothesis W and the membership vector $\boldsymbol{S}$ given the supersample $\tilde{\boldsymbol{Z}}$, that is, $I(W;\boldsymbol{S}|\tilde{\boldsymbol{Z}})$. As discussed in Chapter 6, bounds in terms of the CMI are tighter (up to constants) than the ones based on the mutual information $I(W;\boldsymbol{Z})$. In contrast to the mutual information, there is a wide class of natural empirical risk minimizers for which the CMI can be shown to be bounded by (approximately) the VC dimension. In particular, this applies to any algorithm satisfying the following consistency property. For simplicity, following Steinke and Zakynthinou (2020), we restrict ourselves to deterministic learning algorithms.

Definition 7.5 (Global consistency property)**.** Let $W(\boldsymbol{z})$ denote the point mass on which $P_{W|\boldsymbol{Z}=\boldsymbol{z}}$ concentrates when trained on $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{Z}^n$. Let $\boldsymbol{z}' = (\boldsymbol{x}', \boldsymbol{y}') \in \mathcal{Z}^m$ with $m \geq n$ be constructed so that (i): for all $i \in [n]$, there is a $j \in [m]$ such that $x_i = x'_j$, and, (ii): for all $i \in [m]$, $f_{W(\boldsymbol{z})}(x_i) = y'_i$. Then, the learning algorithm characterized by $P_{W|\boldsymbol{Z}}$ has the global consistency property if, for any $\boldsymbol{z} \in \mathcal{Z}^n$, $P_{W|\boldsymbol{Z}=\boldsymbol{z}'}$ concentrates on $W(\boldsymbol{z})$.

This property requires that if a training set $\boldsymbol{z}$ is re-labelled to obtain $\boldsymbol{z}'$, which is fully consistent with the output hypothesis $W(\boldsymbol{Z})$ obtained from training on $\boldsymbol{z}$ and possibly expanded with more consistent samples, the output hypothesis obtained from training on $\boldsymbol{z}'$ should still be $W(\boldsymbol{Z})$. Clearly, this property is satisfied for many reasonable empirical risk minimizers.

With this, we can show the following.

Theorem 7.6. Consider the $0-1$ loss and assume that the VC dimension d_{VC} of $\mathcal{W}$ is finite. Assume that the learning algorithm satisfies the global consistency property. Then, if $n > d_{\mathrm{VC}}$,

$$I(W;\boldsymbol{S}|\tilde{\boldsymbol{Z}}) \leq d_{\mathrm{VC}} \log\left(\frac{2en}{d_{\mathrm{VC}}}\right). \tag{7.38}$$

Proof. Let $\tilde{z}_* = \arg\max_{\tilde{z}} I(W;\boldsymbol{S}|\tilde{\boldsymbol{Z}} = \tilde{z})$. Also, let $\hat{\mathcal{W}} \subseteq \mathcal{W}$ denote the set of possible output hypotheses obtainable by varying $\boldsymbol{S}$ given the fixed supersample $\tilde{\boldsymbol{z}}_* = (\tilde{\boldsymbol{x}}_*, \tilde{\boldsymbol{y}}_*)$. Then, we have

$$I(W; \boldsymbol{S}|\tilde{\boldsymbol{Z}}) \leq I(W; \boldsymbol{S}|\tilde{\boldsymbol{Z}} = \tilde{\boldsymbol{z}}_*) \leq \log \left|\hat{\mathcal{W}}\right|. \tag{7.39}$$

Now, by the global consistency property, the output hypothesis $w(\tilde{\boldsymbol{z}}_*(\boldsymbol{s}))$ obtained by running the learning algorithm on the training set $\tilde{\boldsymbol{z}}_*(\boldsymbol{s})$ can also be obtained by running the learning algorithm on the training set $\tilde{\boldsymbol{z}}'_* = (\tilde{\boldsymbol{x}}'_*, \tilde{\boldsymbol{y}}'_*)$, which is constructed so that $\tilde{\boldsymbol{x}}_* = \tilde{\boldsymbol{x}}'_*$ and, for all $i \in [2n]$, $f_{w(\tilde{\boldsymbol{z}}_*(\boldsymbol{s}))}((\tilde{x}_*)_i) = (\tilde{y}'_*)_i$. In words: the output hypothesis $w(\tilde{\boldsymbol{z}}_*(\boldsymbol{s}))$ from the training set $\tilde{\boldsymbol{z}}_*(\boldsymbol{s})$ can be obtained by running the learning algorithm on $\tilde{\boldsymbol{z}}'_*$, which only contains samples that are consistent with $w(\tilde{\boldsymbol{z}}_*(\boldsymbol{s}))$. Hence, the number of distinct possible output hypotheses $\left|\hat{\mathcal{W}}\right|$ is upper-bounded by the number of possible labellings of $\tilde{\boldsymbol{x}}_*$ using hypotheses from $\mathcal{W}$. This, in turn, can be bounded using the Sauer-Shelah lemma (Lemma 1.3). Specifically,

$$I(W; \boldsymbol{S}|\tilde{\boldsymbol{Z}}) \leq \log \left|\hat{\mathcal{W}}\right| \leq d_{\mathrm{VC}} \log\left(\frac{2en}{d_{\mathrm{VC}}}\right). \tag{7.40}$$

□

To complete this argument, it remains to show that there exist deterministic empirical risk minimizers with the global consistency property. Since the argument is quite technical, we will not reproduce it here. The proof can be found in Steinke and Zakynthinou (2020, Lemma 4.15).

Note that this result does not imply that *every* empirical risk minimizer over a hypothesis class with finite VC dimension has bounded CMI. For this, we need to consider further processed versions of the CMI.

7.3.3 Evaluated and Functional CMI

We now turn to the evaluated and functional versions of the CMI, or e-CMI and f-CMI for short. Specifically, recall that the f-CMI is given by the mutual information between the predictions $\mathbf{F}$ (for the supersample $\tilde{\boldsymbol{Z}}$ induced by the hypothesis W) and the membership vector $\boldsymbol{S}$ given $\tilde{\boldsymbol{Z}}$, that is, $I(\mathbf{F}; \boldsymbol{S}|\tilde{\boldsymbol{Z}})$. The e-CMI is obtained by replacing the predictions with the losses $\boldsymbol{\Lambda}$ that they induce, that is, $I(\boldsymbol{\Lambda}; \boldsymbol{S}|\tilde{\boldsymbol{Z}})$.

For binary classification with the $0-1$ loss, there is a bijection between $\mathbf{F}$ and $\mathbf{\Lambda}$ given $\tilde{\boldsymbol{Z}}$: the loss of a prediction is 0 if and only if it matches the corresponding label, otherwise the loss is 1. Thus, for this particular case, $I(\mathbf{\Lambda}; \boldsymbol{S}|\tilde{\boldsymbol{Z}}) = I(\mathbf{F}; \boldsymbol{S}|\tilde{\boldsymbol{Z}})$, although the latter more generally only gives an upper bound. We will thus consider only the f-CMI. In contrast to the CMI, it is possible bound the f-CMI for *every* learning algorithm over a hypothesis class with finite VC dimension. We establish this result in the following theorem.

Theorem 7.7. Consider the $0-1$ loss and assume that the VC dimension d_{VC} of $\mathcal{W}$ is finite. Then, if $n > d_{\mathrm{VC}}$,

$$I(\mathbf{F}; \boldsymbol{S}|\tilde{\boldsymbol{Z}}) \le d_{\mathrm{VC}} \log\left(\frac{2en}{d_{\mathrm{VC}}}\right). \tag{7.41}$$

Proof. Let $\tilde{\boldsymbol{z}}_* = \arg\max_{\tilde{\boldsymbol{z}}} I(F; \boldsymbol{S}|\tilde{\boldsymbol{Z}} = \tilde{\boldsymbol{z}})$. Also, let $\hat{\mathcal{F}} \subseteq \mathcal{Y}^{2\times n}$ denote the set of possible predictions obtainable by varying $\boldsymbol{S}$ given the fixed supersample $\tilde{\boldsymbol{z}}_* = (\tilde{\boldsymbol{x}}_*, \tilde{\boldsymbol{y}}_*)$. Then, we have

$$I(\mathbf{F}; \boldsymbol{S}|\tilde{\boldsymbol{Z}}) \le I(F; \boldsymbol{S}|\tilde{\boldsymbol{Z}} = \tilde{\boldsymbol{z}}_*) \le \log\left|\hat{\mathcal{F}}\right|. \tag{7.42}$$

The number of distinct possible output predictions $\hat{\mathcal{F}}$ is upper-bounded by the number of possible labellings of $\tilde{\boldsymbol{x}}_*$ using hypotheses from $\mathcal{W}$. This can be bounded using the Sauer-Shelah lemma (Lemma 1.3), from which the final result follows. □

Again, we emphasize that this result holds for *every* learning algorithm, even beyond empirical risk minimizers. Furthermore, by using the f-CMI, the proof of this result just involves an application of the Sauer-Shelah lemma. In a sense, this provides an information-theoretic re-interpretation of this classic uniform convergence argument (discussed in Section 1.3.1). Specifically, when the hypothesis class has low complexity as measured by the VC dimension, any learning algorithm for the hypothesis class has low information complexity, as measured by the f-CMI.

While this demonstrates that one can obtain bounds for the f-CMI of any learning algorithm, this does not generally lead to optimal generalization bounds, as they are off by a log-factor (Haghifam *et al.*, 2021, Thm. 4.4).

7.3.4 Leave-One-Out CMI

We conclude the discussion of the VC dimension by describing a bound for learning of VC classes over realizable distributions obtained through the leave-one-out evaluated CMI (loo-e-CMI), due to Haghifam *et al.* (2022). Since the proof of this result is somewhat more involved, we will not give it in full detail, but instead just sketch the arguments.

For the purposes of this discussion, we consider the leave-one-out CMI setting introduced in Section 6.6 with the $0-1$ loss, and assume the data distribution to be realizable. First, we connect the binary loss loo-e-CMI of interpolating learning algorithms and the leave-one-out-error, defined as

$$\hat{R}_{\mathrm{loo}} = \mathbb{E}_{P_U}\left[\mathbb{E}_{P_{\dot{\boldsymbol{\Lambda}}|U\dot{\boldsymbol{Z}}}}\left[\dot{\Lambda}_U\right]\right]. \tag{7.43}$$

In words, given a supersample $\dot{\boldsymbol{Z}}$, $\hat{R}_{\mathrm{loo}}$ is the test loss when leaving out the Uth sample, averaged over U and the randomness of the learning algorithm. Notice that $\hat{R}_{\mathrm{loo}} \in [0,1]$. It can be shown that the loo-e-CMI $I(\dot{\boldsymbol{\Lambda}};U|\dot{\boldsymbol{Z}})$ can be bounded by $H_b(\hat{R}_{\mathrm{loo}}) + \hat{R}_{\mathrm{loo}}\log(n+1)$, where $H_b(\hat{R}_{\mathrm{loo}})$ denotes the binary entropy (*i.e.*, the entropy of a Bernoulli random variable with parameter $\hat{R}_{\mathrm{loo}}$) (Haghifam *et al.*, 2022, Thm. 3.1).

Next, we briefly describe the one-inclusion graph algorithm introduced by Haussler *et al.* (1988). Given $\dot{\boldsymbol{Z}} = (\dot{\boldsymbol{X}}, \dot{\boldsymbol{Y}}) \in \mathcal{Z}^{n+1}$, let $\mathcal{V}$ denote the set of possible labellings of $\dot{\boldsymbol{X}} = (\dot{X}_1, \dots, \dot{X}_{n+1})$ with hypotheses from $\mathcal{W}$. We refer to elements of $\mathcal{V}$ as adjacent if they differ in only one element. We define a probability assignment $P : \mathcal{V} \times \mathcal{V} \to [0,1]$ so that $P(\mathbf{v},\mathbf{w}) = 0$ if $\mathbf{v},\mathbf{w} \in \mathcal{V}$ are not adjacent, and $P(\mathbf{v},\mathbf{w}) + P(\mathbf{w},\mathbf{v}) = 1$ if they are, where P is chosen solely on the basis of $\dot{\boldsymbol{X}}$. Recall that $\boldsymbol{Z}_{\bar{U}}$ denotes the training set, formed by removing the Uth entry of $\dot{\boldsymbol{Z}}$, while Z_U is a test sample. Due to the realizability assumption, either one or two elements of $\mathcal{V}$ are consistent with $\dot{\boldsymbol{Z}}_{\bar{u}}$ for $u \in [n]$. The one-inclusion graph algorithm, given the training set $\dot{\boldsymbol{Z}}_{\bar{u}}$, predicts the label of $\dot{y}_u$ as follows: if only one element $\mathbf{v} \in \mathcal{V}$ is consistent with $\dot{\boldsymbol{Z}}_{\bar{u}}$, it predicts v_u. If two elements $\mathbf{v},\mathbf{w} \in \mathcal{V}$ are consistent with $\dot{\boldsymbol{Z}}_{\bar{u}}$, it predicts v_u with probability $P(\mathbf{v},\mathbf{w})$ and w_u otherwise. Let $\mathbf{v}^*$ denote the vector of correct labels for $\dot{\boldsymbol{X}}$. When using $\dot{\boldsymbol{Z}}_{\bar{u}}$ as training set, the probability of incurring an error on $\dot{\boldsymbol{Z}}_u$ is given by $P(\mathbf{v}',\mathbf{v}^*)$ for $\mathbf{v}'$ such that $v'_u \neq v^*_u$

but all other entries of $\mathbf{v}'$ and $\mathbf{v}^*$ are equal, provided that such a $\mathbf{v}'$ exists in $\mathcal{V}$. Otherwise, it is zero. Therefore, the leave-one-out error is given by

$$\hat{R}_{\text{loo}} = \sum_{\mathbf{v}' \in \mathcal{V}} \frac{P(\mathbf{v}', \mathbf{v}^*)}{n+1}. \tag{7.44}$$

Haussler *et al.* (1988, Lemma 5.2) established that there exists a probability assignment such that $\sum_{\mathbf{v}' \in \mathcal{V}} P(\mathbf{v}', \mathbf{w}) \leq d_{\text{VC}}$ uniformly for $\mathbf{w} \in \mathcal{V}$. By combining this with the bound on $I(\dot{\mathbf{\Lambda}}; U | \dot{\boldsymbol{Z}})$ in terms of $\hat{R}_{\text{loo}}$ provided in the first step, a bound for learning realizable VC classes can be established.

Notably, in the works of Haghifam *et al.* (2021, 2022), the CMI of a learning algorithm is demonstrated to provide a *universal* characterization of realizable generalization in a certain sense: specifically, for every interpolating learning algorithm and data distribution, the population loss vanishes as n goes to infinity *if and only if* the CMI of the learning algorithm grows sub-linearly in n. For the loo-e-CMI, an even stronger characterization can be established, in the sense that the loo-e-CMI also captures the decay rate when the population loss decays polynomially or converges to a positive value. For more details, the reader is referred to Haghifam *et al.* (2021, 2022).

7.4 Compression Schemes

We now consider a class of learning algorithms known as *compression schemes* (Littlestone and Warmuth, 2003). A compression scheme of size k consists of two components: a sequence of maps $\kappa : \mathcal{Z}^n \to \mathcal{Z}^k$ for $n \geq k$, which given an input vector $\boldsymbol{Z}$ of size n outputs a vector $\kappa(\boldsymbol{Z})$ consisting of k elements of $\boldsymbol{Z}$; and a map $\rho : \mathcal{Z}^k \to \mathcal{W}$ that selects a hypothesis based on this compressed training set. By composing these maps, we obtain a learning algorithm for training sets of size $n \geq k$.

As an example, consider threshold classifiers, as introduced in Section 7.3.1, and a learning algorithm that simply sets the threshold W to be the smallest training feature with the label 1, *i.e.*, $W = \min\{x : (x, 1) \in \boldsymbol{Z}\}$ (and $W = \infty$ if there is no sample with the label 1). Clearly, this can be written as the composition of a map κ that outputs $\kappa(\boldsymbol{Z}) = (x_{i^*}, y_{i^*})$, where $i^* = \arg\min_i\{x_i : (x_i, 1) \in \boldsymbol{Z}\}$, and a map

$$\rho(x,y) = \begin{cases} x \text{ if } y = 1 \\ \infty \text{ otherwise.} \end{cases} \tag{7.45}$$

Therefore, it is a compression scheme of size 1.

The mutual information $I(W;\boldsymbol{Z})$ of such algorithms will generally be unbounded, since we are dealing with deterministic algorithms with continuous inputs and outputs. However, for the CMI, the following can be established, as per Steinke and Zakynthinou (2020, Thm 4.2).

Theorem 7.8. Assume that $P_{W|\boldsymbol{Z}}$ is a compression scheme of size k. Then, we have $I(W;\boldsymbol{S}|\tilde{\boldsymbol{Z}}) \leq k\log(2n)$.

Proof. Since W is a function of $\kappa(\boldsymbol{Z}_{\boldsymbol{S}})$,

$$I(W;\boldsymbol{S}|\tilde{\boldsymbol{Z}}) \leq I(\kappa(\boldsymbol{Z}_{\boldsymbol{S}});\boldsymbol{S}|\tilde{\boldsymbol{Z}}) \leq H(\kappa(\boldsymbol{Z}_{\boldsymbol{S}})|\tilde{\boldsymbol{Z}}) \leq k\log(2n). \tag{7.46}$$

Here, the last step follows since, given $\tilde{\boldsymbol{Z}}$, there are at most $\binom{2n}{k} \leq (2n)^k$ possible values of $\kappa(\boldsymbol{Z}_{\boldsymbol{S}})$. This establishes the result. □

Up to constants, this bound cannot be improved for general compression schemes. However, for the important subclass of *stable* compression schemes, the logarithmic dependence on n can be removed. A compression scheme is said to be stable if it is invariant to permutations of its input, and $\kappa(\boldsymbol{Z}) = \kappa(\boldsymbol{Z}')$ if $\kappa(\boldsymbol{Z}) \subseteq \boldsymbol{Z}' \subseteq \boldsymbol{Z}$—that is, if only elements that are not in the compressed set are removed from the training set, this does not change the output. For stable compression schemes, Haghifam *et al.* (2021, Thm. 3.4) showed that $I(W;\boldsymbol{S}|\tilde{\boldsymbol{Z}}) \leq 2k\log(2)$. This result demonstrates that the CMI suffices to obtain generalization bounds for stable compression schemes without a logarithmic dependence on n, which is optimal up to constants (Haghifam *et al.*, 2021, Thm. 3.1).

7.5 Algorithmic Stability

We now turn to algorithmic stability, as discussed in Section 1.4. As mentioned therein, several notions of stability have been discussed in the literature. In this section, following Harutyunyan *et al.* (2021, Thm. 4.2), we will focus on average prediction stability with respect to sample replacement and bound the f-CMI. This notion of stability is comparable to the pointwise hypothesis stability in Bousquet and Elisseeff (2002, Def. 4). Note that Harutyunyan *et al.* (2021) also consider other notions of stability, which we do not cover for brevity. We will discuss further connections between algorithmic stability and information-theoretic and PAC-Bayesian generalization bounds in Section 7.7.

Theorem 7.9. Assume that $\mathcal{Z} = \mathcal{X} \times \mathbb{R}^d$ and $\ell(w, z) = \ell_f(f_w(x), y)$, where each $w \in \mathcal{W}$ induces a function $f_w : \mathcal{X} \to \mathbb{R}^d$. Let $\boldsymbol{Z}_{\boldsymbol{S}}^{(i)}$ equal $\boldsymbol{Z}_{\boldsymbol{S}}$ for all entries except the ith, which we denote by $Z' = (X', Y')$, and assume to be independently drawn from P_Z. Consider a deterministic learning algorithm, and let $f_{W|\boldsymbol{Z}_{\boldsymbol{S}}} : \mathcal{X} \to \mathbb{R}^d$ denote the function that the learning algorithm induces given the training set $\boldsymbol{Z}_{\boldsymbol{S}}$. Assume that the learning algorithm is β-stable, meaning that for all $i \in [n]$,

$$\mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}} P_{Z'}}\left[\left\| f_{W|\boldsymbol{Z}_{\boldsymbol{S}}}(\tilde{X}_{i+S_i n}) - f_{W|\boldsymbol{Z}_{\boldsymbol{S}}^{(i)}}(\tilde{X}_{i+S_i n}) \right\|^2\right] \leq \beta^2. \tag{7.47}$$

Roughly speaking, this means that the prediction that the hypothesis issues for $\tilde{X}_{i+S_i n}$ does not depend too strongly on whether or not this specific sample is included in the training set. Furthermore, suppose that the loss function $\ell_f(\cdot, \cdot)$ is γ-Lipschitz in its first argument. Then, we have that

$$|\overline{\text{gen}}| \leq d^{1/4}\sqrt{8\gamma\beta}. \tag{7.48}$$

Proof. In order to establish this result, we will relate the deterministic algorithm to a stochastic one. Specifically, let

$$f^{\sigma}_{W|\boldsymbol{Z}_{\boldsymbol{S}},N}(x) = f_{W|\boldsymbol{Z}_{\boldsymbol{S}}}(x) + N_\sigma. \tag{7.49}$$

Here, the Gaussian noise $N_\sigma \sim \mathcal{N}(0, \sigma^2 I_d)$, where I_d denotes the d-dimensional identity matrix, is independent for all training sets and inputs. With this, we find that the average generalization gap of the learning algorithm with added noise is

$$\begin{aligned}\overline{\text{gen}}_\sigma =& \bigg| \mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}P_{Z'}} \bigg[\mathbb{E}_{P_{N_\sigma}} \Big[\ell_f(f^\sigma_{W|\boldsymbol{Z_S},N}(X'), Y') \Big] \\ &\qquad - \frac{1}{n} \sum_{i\in[n]} \mathbb{E}_{P_N} \Big[\ell_f(f^\sigma_{W|\boldsymbol{Z_S},N}(X_{i+S_in}), Y_{i+S_in}) \Big] \bigg] \bigg| \\ =& \bigg| \mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}P_{Z'}} \bigg[\ell_f(f_{W|\boldsymbol{Z_S}}(X'), Y') + \mathbb{E}_{P_{N_\sigma}}[\Delta'] \qquad (7.50) \\ &\qquad - \frac{1}{n} \sum_{i\in[n]} \Big(\ell_f(f_{W|\boldsymbol{Z_S}}(X_{i+S_in}), Y_{i+S_in}) + \mathbb{E}_{P_{N_\sigma}}[\Delta_i] \Big) \bigg] \bigg|,\end{aligned}$$

where

$$\Delta' = \ell_f(f^\sigma_{W|\boldsymbol{Z_S},N}(X'), Y') - \ell_f(f_{W|\boldsymbol{Z_S}}(X'), Y'), \tag{7.51}$$

$$\Delta_i = \ell_f(f^\sigma_{W|\boldsymbol{Z_S},N}(X_{i+S_in}), Y_{i+S_in}) - \ell_f(f_{W|\boldsymbol{Z_S}}(X_{i+S_in}), Y_{i+S_in}). \tag{7.52}$$

Due to the Lipschitz assumption, we have $|\Delta'| \leq \gamma \left\| {N_\sigma}' \right\|$, where ${N_\sigma}' \sim \mathcal{N}(0, \sigma^2 I_d)$. Similarly, $|\Delta_i| \leq \gamma \left\| {N_\sigma}' \right\|$. Since $\mathbb{E}[\|{N_\sigma}'\|] \leq 2\sigma\sqrt{d}$, we find that

$$\overline{\text{gen}}_\sigma \geq \overline{\text{gen}} - 2\gamma\sigma\sqrt{d}. \tag{7.53}$$

We now need to bound $\overline{\text{gen}}_\sigma$. Let $\mathbf{F}^\sigma$ denote the vector of predictions on $\tilde{\boldsymbol{X}}$ induced by $f^\sigma_{W|\boldsymbol{Z_S},N}$. By the individual-sample f-CMI version of Theorem 6.12, we have

$$\overline{\text{gen}}_\sigma \leq \frac{1}{n} \sum_{i\in[n]} \sqrt{2I(F^\sigma_i, F^\sigma_{i+n}; S_i | \tilde{\boldsymbol{Z}})} \tag{7.54}$$

$$\leq \frac{1}{n} \sum_{i\in[n]} \sqrt{2I(F^\sigma_i, F^\sigma_{i+n}; S_i | \boldsymbol{S}_{-i}, \tilde{\boldsymbol{Z}})}, \tag{7.55}$$

where $\boldsymbol{S}_{-i}$ is $\boldsymbol{S}$ with the ith entry removed. Here, the last step follows since $\boldsymbol{S}_{-i}$ is independent from S_i. To establish the result, it remains to bound the conditional mutual information in (7.55). Intuitively, computing this quantity involves comparing the conditional joint distribution of $(F^\sigma_i, F^\sigma_{i+n})$ and S_i, given $\boldsymbol{S}_{-i}$ and $\tilde{\boldsymbol{Z}}$, with the products of their conditional marginals. When S_i is drawn independently from all other random variables, there is a 50% chance of drawing the "matching" instance, in which case the two distributions coincide, and a 50% chance of drawing the "opposite" instance, in which case the ith sample of the training set is replaced. Hence, we are comparing two Gaussian distributions with covariance $\sigma^2 I_d$ and means given by the predictions based on the training set corresponding to $\boldsymbol{S}_{-i}$ and either $S_i = 1$ or $S_i = 0$. By the stability assumption, the difference between the means is on average bounded by β^2 (for more details, see the work of Harutyunyan *et al.*, 2021, Prop. 4.2 and Eq. (175)-(179)). Since $D(\mathcal{N}(x_1, \sigma^2 I_d) \,||\, \mathcal{N}(x_2, \sigma^2 I_d)) = \|x_1 - x_2\|^2 / (2\sigma^2)$, we get

$$I(F_i^\sigma, F_{i+n}^\sigma; S_i | \boldsymbol{S}_{-i}, \tilde{\boldsymbol{Z}}) \leq \frac{\beta^2}{2\sigma^2}. \tag{7.56}$$

By combining (7.53), (7.55), and (7.56), setting $\sigma^2 = \beta/(2\gamma\sqrt{d})$ to optimize the bound, we obtain the desired result. □

Thus, for Lipschitz losses, certain notions of algorithmic stability imply bounds on certain information measures for the learning algorithm, allowing us to (essentially) recover known generalization bounds (cf. Section 1.4). The technique used in this proof, where a learning algorithm is compared to a noisy surrogate in order to more easily evaluate the mutual information, is a fruitful approach that has also been used to establish generalization bounds for stochastic gradient descent (Neu *et al.*, 2021).

7.6 Differential Privacy and Related Measures

We now discuss differential privacy, which can be seen as a type of stability measure. As the name suggests, this measure was originally constructed as a guarantee on the privacy of the training data used by a learning algorithm. Specifically, let $\boldsymbol{z}, \boldsymbol{z}' \in \mathcal{Z}^n$ be two training sets that differ in a single element. Then, the algorithm $P_{W|\boldsymbol{Z}}$ is ε-differentially private if, for any measurable set $\mathcal{E} \in \mathcal{W}$ (Dwork *et al.*, 2015)

$$P_{W|\boldsymbol{Z}=\boldsymbol{z}}(\mathcal{E}|\boldsymbol{z}) \leq e^{\varepsilon} P_{W|\boldsymbol{Z}=\boldsymbol{z}'}(\mathcal{E}|\boldsymbol{z}'). \tag{7.57}$$

This is related to so-called ε-MI stability, which requires that for any random $\boldsymbol{Z} \in \mathcal{Z}^n$ (Feldman and Steinke, 2018)

$$\frac{1}{n}\sum_{i=1}^{n} I(W; Z_i|\boldsymbol{Z}_{-i}) \leq \varepsilon, \tag{7.58}$$

where $\boldsymbol{Z}_{-i}$ denotes $\boldsymbol{Z}$ with the ith element removed. As shown by Feldman and Steinke (2018), an algorithm that is $\sqrt{2\varepsilon}$-differentially private is ε-MI stable. If the elements of $\boldsymbol{Z}$ are independent, we have

$$I(W; \boldsymbol{Z}) = \sum_{i=1}^{n} I(W; Z_i|\boldsymbol{Z}_{<i}) \leq \sum_{i=1}^{n} I(W; Z_i|\boldsymbol{Z}_{-i}) \leq \varepsilon n, \tag{7.59}$$

where $\boldsymbol{Z}_{<i} = (Z_1, \dots, Z_{i-1})$ (and $\boldsymbol{Z}_{<1} = \emptyset$). Thus, any ε-MI stable (including any $\sqrt{2\varepsilon}$-differentially private) learning algorithm has mutual information bounded by εn.

We conclude with a brief mention of max information, defined as (Dwork *et al.*, 2015)

$$I_{\max}(W; \boldsymbol{Z}) = \operatorname*{ess\,sup}_{P_{W\boldsymbol{Z}}} \imath(W, \boldsymbol{Z}). \tag{7.60}$$

As established by Esposito *et al.* (2021a, Lemma 12), $\mathcal{L}(\boldsymbol{Z} \to W) \leq I_{\max}(W; \boldsymbol{Z})$. Furthermore, since the α-mutual information is non-decreasing with α (Verdú, 2015), and it coincides with the mutual information for $\alpha = 1$ and the maximal leakage for $\alpha \to \infty$, we have

$$I(W; \boldsymbol{Z}) \leq \mathcal{L}(\boldsymbol{Z} \to W) \leq I_{\max}(W; \boldsymbol{Z}). \tag{7.61}$$

Thus, bounds in terms of max information, as discussed by Dwork *et al.* (2015), can be recovered from bounds in terms of the mutual information and maximal leakage.

7.7 Bibliographic Remarks and Additional Perspectives

In this section, we discuss the relation of the results we presented to the literature, and give a brief overview of results that we did not cover explicitly. For the Gibbs posterior, Theorem 7.3 is largely based on Raginsky *et al.* (2021, Chapter 10), while Theorem 7.4 is due to Aminian *et al.* (2021b).

The Gaussian location model has been studied as an example application of information-theoretic generalization bounds since the work of Bu *et al.* (2019), with later improvements by Wu *et al.* (2022a) and Zhou *et al.* (2021, 2022). An information-theoretic bound that is tight up to constants was provided by Zhou *et al.* (2023a).

For learning with VC classes, Xu and Raginsky (2017) constructed a two-phase learning algorithm with finite mutual information, but this result does not apply to standard empirical risk minimizers. As shown by Bassily *et al.* (2018), Livni and Moran (2017), and Nachum *et al.* (2018), there are certain limitations in obtaining finite PAC-Bayesian and information-theoretic generalization bounds using the standard, non-CMI framework. Recently, Pradeep *et al.* (2022) showed that under the stricter requirement of a finite Littlestone dimension, it can be shown that learnability is possible with finite mutual information, demonstrating a gap compared to just having finite VC dimension. Through the use of the CMI framework, Steinke and Zakynthinou (2020) obtained Theorem 7.6 for all empirical risk minimizers satisfying the consistency property of Definition 7.5. As shown by Harutyunyan *et al.* (2021), the use of functional CMI enables Theorem 7.7, which applies to any learning algorithm. An extension to the Natarajan dimension, which is an analogue of the VC dimension for the multiclass setting, was provided by Hellström and Durisi (2022a). Finally, the leave-one-out CMI framework enables optimal bounds for VC classes in certain situations, as shown by Haghifam *et al.* (2022) and discussed in Section 7.3.4. Further discussion of the expressiveness of information-theoretic generalization bounds can be found in the work of Haghifam *et al.* (2021). Notably, generalization bounds in terms of the VC dimension obtained from PAC-Bayesian bounds were originally derived in the work of Catoni (2004a, Corollary 2.4). The derivation is very similar to the CMI case, and based on the formalism of exchangeable priors. This was extended to almost exchangeable priors by Audibert (2004) and Catoni (2007). Recently, a further extension that allows for bounds with fast rates under a Bernstein condition was provided by Grünwald *et al.* (2021). Furthermore, Grünwald and Mehta (2019) also explored connections between PAC-Bayesian bounds and the Rademacher complexity.

For compression schemes, Steinke and Zakynthinou (2020) obtained the result of Theorem 7.8. This was improved by a logarithmic factor for stable compression schemes by Haghifam *et al.* (2021, Theorem 3.1). Catoni (2004a, Sec. 3) studied the use of exchangeable priors to obtain bounds for compression schemes.

The result in Theorem 7.9 is due to Harutyunyan *et al.* (2021), who also established results for other notions of algorithmic stability. Bounds based on average stability, with connections to information-theoretic generalization bounds, were also established by Banerjee *et al.* (2022). PAC-Bayesian generalization bounds in terms of stability have been established by, for instance, London (2017), London *et al.* (2014), Rivasplata *et al.* (2018), Sun *et al.* (2022), and Zhou *et al.* (2023b).

The discussion of privacy measures, such as the differential privacy of Dwork *et al.* (2015), in Section 7.6 is largely based on results from Feldman and Steinke (2018), with additional results due to Esposito *et al.* (2021a). For further discussion of these and other privacy measures, see for instance the work of Esposito *et al.* (2021a), Hellström and Durisi (2020a), Oneto *et al.* (2020), Rodríguez-Gálvez *et al.* (2021a), and Steinke and Zakynthinou (2020).

8

Neural Networks and Iterative Algorithms

In this chapter, we apply the bounds from Chapters 4 to 6 to learning algorithms that are *iterative* in nature, in the sense that they proceed by updating a hypothesis step-by-step with the aim to converge to a final output hypothesis with good properties. A key example of such an algorithm is the ubiquitous *gradient descent*, which updates the current hypothesis by adding the negative gradient of the training loss, scaled by a parameter called the learning rate. Of particular importance in modern machine learning are neural networks, which are typically trained using variants of (stochastic) gradient descent. However, the framework of iterative learning algorithms applies to a much broader class of learning algorithms.

In Section 8.1, we discuss iterative, noisy algorithms in general, before specializing to the case of stochastic gradient Langevin dynamics (SGLD). SGLD is a variant of stochastic gradient descent (SGD) with added Gaussian noise, which makes it particularly well-suited to analysis via information-theoretic bounds. In Section 8.2, we discuss the application of generalization bounds from Chapters 4 to 6 to neural networks. Clearly, some bounds cannot be computed for practical scenarios—for instance, the mutual information depends on the unknown data distribution, and some information metrics can be prohibitively expensive to estimate due to high dimensionality or the lack of closed-form expressions. For many bounds, however, it is possible to obtain informative values, for instance by using Monte Carlo estimates. We will mainly focus on methods for numerically evaluating the bounds, and discuss training algorithms inspired by them. We will also provide pointers to methods for obtaining generalization bounds in closed form.

8.1 Noisy Iterative Algorithms and SGLD

Here, we consider iterative learning algorithms of the following general form. The hypothesis space $\mathcal{W}$ is the d-dimensional Euclidean space $\mathbb{R}^d$. Given the training data $\boldsymbol{Z} = (Z_1, \dots, Z_n)$, we generate the hypothesis W as follows:

$$\begin{aligned} W &= f(V_1, \dots, V_T) \\ V_t &= g(V_{t-1}) - \eta_t F(V_{t-1}, Z_{J_t}) + \xi_t, \qquad t = 1, \dots, T \end{aligned} \tag{8.1}$$

where V_0 is a random initial condition independent of everything else; $T \in \mathbb{N}$ is a fixed number of iterations; $J_1, \dots, J_t$ is a sequence of random elements of $[n] = \{1, \dots, n\}$; $\xi_t \sim \mathcal{N}(0, \rho_t^2 I_d)$ is a sequence of independent Gaussian random vectors which are also independent of everything else; and finally, $f(\cdot), g(\cdot), F(\cdot, \cdot)$ are deterministic mappings. We will use the shorthand $\boldsymbol{V} = (V_0, \dots, V_T)$.

The analysis relies on the following regularity assumptions:

1. The following holds for the algorithm's *sampling strategy*, *i.e.*, the conditional probability law of $\boldsymbol{J} = (J_1, \dots, J_T)$ given $(\boldsymbol{Z}, \boldsymbol{V})$: for each $t \in [T-1]$,

$$P_{J_{t+1}|J_1,\dots,J_t,\boldsymbol{V},\boldsymbol{Z}} = P_{J_{t+1}|J_1,\dots,J_t,\boldsymbol{Z}}. \tag{8.2}$$

   That is, the index of the sample in round $t+1$ does not depend on the iterates $V_1, \dots, V_t$, given the previous choices $J_1, \dots, J_t$ and the data $\boldsymbol{Z}$.

2. The update function $F(\cdot, \cdot)$ is bounded:

$$\sup_{v \in \mathbb{R}^d} \sup_{z \in \mathcal{Z}} \|F(v, z)\| \le L < \infty. \tag{8.3}$$

To control the generalization error, we will upper-bound the mutual information $I(W;\boldsymbol{Z})$. Let $\boldsymbol{Z}^{\boldsymbol{J}} = (Z_{J_1}, \ldots, Z_{J_T})$ denote the random T-tuple of the training instances "visited" by the algorithm and observe that $\boldsymbol{Z}$ and $\boldsymbol{V}$ are conditionally independent given $\boldsymbol{Z}^{\boldsymbol{J}}$. Using this fact together with the data processing inequality and the chain rule, we have the following:

$$\begin{aligned} I(W;\boldsymbol{Z}) &= I(f(\boldsymbol{V});\boldsymbol{Z}) && (8.4)\\ &\le I(\boldsymbol{V};\boldsymbol{Z}) && (8.5)\\ &\le I(\boldsymbol{V};\boldsymbol{Z}^{\boldsymbol{J}}) && (8.6)\\ &= \sum_{t=1}^{T} I(V_t;\boldsymbol{Z}^{\boldsymbol{J}}|V^{t-1}). && (8.7) \end{aligned}$$

Each term in (8.7) admits a simple expression involving only random variables from two successive time steps, as we show in the following lemma.

Lemma 8.1. Under the conditional independence assumption on the sampling strategy in (8.2),

$$I(V_t;\boldsymbol{Z}^{\boldsymbol{J}}|V^{t-1}) = I(V_t;Z_{J_t}|V_{t-1}). \tag{8.8}$$

Proof. First, we express $I(V_t;\boldsymbol{Z}^{\boldsymbol{J}}|V^{t-1})$ as

$$I(V_t;\boldsymbol{Z}^{\boldsymbol{J}}|V^{t-1}) = h(V_t|V^{t-1}) - h(V_t|V^{t-1},\boldsymbol{Z}^{\boldsymbol{J}}), \tag{8.9}$$

where $h(\cdot|\cdot)$ is the conditional differential entropy (Definition 3.4). From the update rule for V_t in (8.1) and the assumption on $\{\xi_t\}_{t\in[T]}$, it follows that V_t is conditionally independent from $(V^{t-2},\boldsymbol{Z}^{\boldsymbol{J}\setminus\{J_t\}})$ given (V_{t-1},Z_{J_t}). Using this, we conclude that

$$\begin{aligned} h(V_t|V^{t-1},\boldsymbol{Z}^{\boldsymbol{J}}) &= h(V_t|V_{t-1},Z_{J_t},V^{t-2},\boldsymbol{Z}^{\boldsymbol{J}\setminus\{J_t\}})\\ &= h(V_t|V_{t-1},Z_{J_t}). \end{aligned}$$

By the same token, $h(V_t|V^{t-1}) = h(V_t|V_{t-1})$. Using these expressions in (8.9), we obtain the desired result. □

The following lemma provides an easy-to-compute upper bound on $I(V_t;Z_{J_t}|V_{t-1})$.

Lemma 8.2. For every $t \in [T]$,

$$I(V_t; Z_{J_t}|V_{t-1}) \leq \frac{d}{2} \log\left(1 + \frac{\eta_t^2 L^2}{d\rho_t^2}\right) \leq \frac{\eta_t^2 L^2}{2\rho_t^2}. \tag{8.10}$$

Proof. Given $V_{t-1} = v_{t-1}$, we have

$$V_t = g(v_{t-1}) - \eta_t F(v_{t-1}, Z_{J_t}) + \xi_t, \tag{8.11}$$

where Z_{J_t} and ξ_t are independent. Consequently, by the shift-invariance property of differential entropy,

$$\begin{aligned} h(V_t|V_{t-1} = v_{t-1}) &= h(V_t - g(v_{t-1})|V_{t-1} = v_{t-1}) &(8.12)\\ &= h(-\eta_t F(v_{t-1}, Z_{J_t}) + \xi_t|V_{t-1} = v_{t-1}). &(8.13)\end{aligned}$$

Now, recall that for any d-dimensional random vector U with finite second moment, *i.e.*, $\mathbb{E}\left[\|U\|^2\right] < \infty$, we have (Polyanskiy and Wu, 2022, Thm. 2.7)

$$h(U) \leq \frac{d}{2} \log\left(\frac{2\pi e\, \mathbb{E}\left[\|U\|^2\right]}{d}\right). \tag{8.14}$$

Since Z_{J_t} and ξ_t are independent and ξ_t has zero mean, we obtain

$$\begin{aligned} &\mathbb{E}\Big[\| - \eta_t F(v_{t-1}, Z_{J_t}) + \xi_t\|^2 \,|\, V_{t-1} = v_{t-1}\Big]\\ &= \eta_t^2\, \mathbb{E}\Big[\|F(v_{t-1}, Z_{J_t})\|^2 \,|\, V_{t-1} = v_{t-1}\Big] + \mathbb{E}\Big[\|\xi_t\|^2\Big]\\ &\leq \eta_t^2 L^2 + \rho_t^2 d, &(8.15)\end{aligned}$$

where we have also used the uniform boundedness assumption on $F(\cdot, \cdot)$. Consequently,

$$h(V_t|V_{t-1}) \leq \frac{d}{2} \log\left(\frac{2\pi e(\eta_t^2 L^2 + \rho_t^2 d)}{d}\right). \tag{8.16}$$

By the same reasoning,

$$\begin{aligned} h(V_t|V_{t-1}, Z_{J_t}) &= h(g(V_{t-1}) - \eta_t F(V_{t-1}, Z_{J_t}) + \xi_t|V_{t-1}, Z_{J_t}) &(8.17)\\ &= h(\xi_t|V_{t-1}, Z_{J_t}) &(8.18)\\ &= h(\xi_t) &(8.19)\\ &= \frac{d}{2} \log(2\pi e \rho_t^2), &(8.20)\end{aligned}$$

where we have used the fact that ξ_t is independent of the pair (V_{t-1}, Z_{J_t}). Hence,

$$\begin{aligned} I(V_t; Z_{J_t}|V_{t-1}) &= h(V_t|V_{t-1}) - h(V_t|V_{t-1}, Z_{J_t}) && (8.21) \\ &\le \frac{d}{2}\log\left(1 + \frac{\eta_t^2 L^2}{\rho_t^2 d}\right) && (8.22) \\ &\le \frac{\eta_t^2 L^2}{2\rho_t^2}, && (8.23) \end{aligned}$$

where the last step follows from the inequality $\log x \le x - 1$. □

Combining Lemmas 8.1 and 8.2 and the mutual information generalization bound in Corollary 4.2, we get the following result, due to Pensia *et al.* (2018).

Theorem 8.3. Suppose that $\ell(w, Z)$ is σ^2-subgaussian for every $w \in \mathcal{W}$ under P_Z. Then, under the assumptions on the sampling strategy and on F stated in (8.1) and (8.2), we have

$$\mathbb{E}_{P_{W\boldsymbol{Z}}}[\mathrm{gen}(W, \boldsymbol{Z})] \le \sqrt{\frac{\sigma^2}{n}\sum_{t=1}^{T}\frac{\eta_t^2 L^2}{\rho_t^2}}. \qquad (8.24)$$

We now specialize the result in (8.24) to the case of SGLD. Specifically, we assume that the loss $\ell(w, z)$ is differentiable as a function of w for every z, and take

$$\begin{aligned} V_0 &= 0 \\ V_t &= V_{t-1} - \eta_t \nabla\ell(V_{t-1}, Z_{J_t}) + \xi_t, \qquad t = 1, \dots, T \qquad (8.25) \\ W &= V_T \end{aligned}$$

where $J_1, \dots, J_T$ are i.i.d. samples from the uniform distribution on $[n]$ (in each iteration, we sample with replacement from the n-tuple $\boldsymbol{Z}$); $\eta_1, \dots, \eta_T$ are positive step sizes; and $\xi_t \sim \mathcal{N}(0, \rho_t^2 I_d)$, with $\rho_t^2 = \frac{\eta_t}{\beta}$ for some $\beta > 0$. The resulting SGLD algorithm is a special case of (8.1) with $g(v) = v$, $F(v, z) = \nabla\ell(v, z)$, and $f(v_1, \dots, v_T) = v_T$. Thus, W is the last iterate V_T, although other choices are possible, such as $f(v_1, \dots, v_T) = \frac{1}{T}\sum_{t=1}^{T} v_t$ (trajectory averaging). There exists a large literature on generalization bounds in expectation for SGLD; here we provide one such result due to Pensia *et al.* (2018), obtained under the (restrictive) assumption of a Lipschitz-continuous loss.

Theorem 8.4. Suppose that the loss function $w \mapsto \ell(w, z)$ is L-Lipschitz uniformly in z:

$$\sup_{z \in \mathcal{Z}} |\ell(w, z) - \ell(w', z)| \le L\|w - w'\|. \tag{8.26}$$

Assume that the SGLD algorithm in (8.25) (with an arbitrary postprocessing step) runs for $T = nk$ steps, where k is a positive integer, and let $\eta_t = \frac{1}{t}$. Then

$$\mathbb{E}_{P_{W\boldsymbol{Z}}}[\mathrm{gen}(W, \boldsymbol{Z})] \le \sqrt{\frac{\beta\sigma^2 L^2}{n} \sum_{t=1}^{nk} \frac{1}{t}} \tag{8.27}$$

$$\le \sqrt{\frac{\beta\sigma^2 L^2}{n}(\log n + \log k + 1)}. \tag{8.28}$$

Proof. By the Lipschitz assumption on ℓ, its gradient $\nabla\ell(\cdot, \cdot)$ is bounded by L in ℓ^2 norm. The result then follows from Theorem 8.3. □

8.2 Numerical Bounds for Neural Networks

In recent years, many practical successes in machine learning have relied on neural networks (NNs). Although a comprehensive discussion of NNs is beyond the scope of this monograph, we will provide a very brief description of NNs and introduce some notation. Further details can be found in, for instance, Murphy (2022, Chapter III). While a whole host of different NN architectures have been developed for specific application areas, we will focus solely on so-called feedforward NNs. We proceed by defining a single layer, from which NNs can be formed through composition. Each layer consists of two components: an affine transformation and an activation function. Denote the input to the lth layer as $x_{l-1} \in \mathbb{R}^{d_{l-1}}$. The weights of the lth layer are denoted by $A_l \in \mathbb{R}^{d_l \times d_{l-1}}$, while the bias vector is $b_l \in \mathbb{R}^{d_l}$. We refer to d_l as the width of the layer. Then, the pre-activation output is given by $a_l = A_l x_{l-1} + b_l$, which is simply an affine transformation of the input. In order to allow the network to express non-linear functions, we also use an activation function $\phi_l : \mathbb{R} \to \mathbb{R}$. Then, the final output from the layer is given by $x_l = \phi_l(a_l)$, where the activation function is applied elementwise to the pre-activation vector a_l. Since NNs are typically trained through a gradient-based algorithm, this activation function is often required to be (almost everywhere) differentiable. An NN $f_W(\cdot)$ of depth L consists of L such layers, where we let $W \in \mathbb{R}^p$, with $p = \sum_{l=1}^{L}(d_{l-1}+1)d_l$, denote the concatenation of all weights and biases expressed in vector form. We will typically also denote the output as $\hat{y} = x_L \in \mathbb{R}^{d_L}$ and the input as $x = x_0 \in \mathbb{R}^{d_0}$. Thus, the final output is $\hat{y} = f_W(x) = \phi_L(a_L)$.

For a given sample $z = (x, y)$, the loss is given by $\ell(W, z) = \ell_f(\hat{y}, y)$. Given the training set $\boldsymbol{Z}$, we assume that the NN is trained as follows: first, the weights and biases of the network are initialized as W_0. At each time step t, they are then updated as

$$\begin{aligned} W_t &= W_{t-1} - \eta \nabla_W L_{\boldsymbol{Z}}(W) && (8.29) \\ &= W_{t-1} - \eta \sum_{i=1}^{n} \nabla_W \ell_f(\hat{y}_i, y_i) && (8.30) \\ &= W_{t-1} - \eta \sum_{i=1}^{n} \nabla_W f_W(x_i) \frac{\mathrm{d}\ell_f(\hat{y}_i, y_i)}{\mathrm{d}\hat{y}_i}. && (8.31) \end{aligned}$$

Here, $\eta > 0$ is the learning rate. The exact form of this update depends on the specific activation function under consideration, and can be computed for each parameter of the network through the chain rule. This process may, for instance, continue for a fixed number of steps or until a certain target loss, either evaluated on the training set or on a held-out validation set, is reached. One common variant of (8.31) is SGD, where the training loss gradient is not evaluated with respect to the entire training set at each time step. Instead, a “mini batch” of $K < n$ samples is selected at each time step, and the weight update is computed with respect to these samples. This has several benefits, such as speeding up computation and reducing memory requirements.

Typically, NNs operate in the so-called *overparameterized* regime. This means that p, which is determined by the widths and depth of the network, is greater than what would be needed in order to interpolate the n training samples in $\boldsymbol{Z}$ after gradient descent training. In many practically relevant scenarios, p is many orders of magnitude greater than n. In fact, NNs often have the capacity to interpolate the training data even with randomly assigned labels. This indicates that they do not operate in a regime where notions like the VC dimension are relevant (Zhang *et al.*, 2021). Still, when trained using data with the correct labels, NNs display impressive generalization performance. So, in the regime that is relevant in practice, NNs generalize well when trained with true labels, but generalize poorly when trained with random labels. This appears to indicate that any generalization guarantee that is uniform over all data distributions is doomed to be vacuous, since such a guarantee would need to hold for both scenarios. This provides a motivation for considering PAC-Bayesian and information-theoretic bounds, as these can incorporate data-distribution dependence. We now discuss various ways to evaluate information-theoretic and PAC-Bayesian bounds for NNs.

8.2.1 Weights with Gaussian Noise

One issue with applying many standard PAC-Bayesian and information-theoretic generalization bounds, as repeatedly discussed, is that they are often vacuous for deterministic learning algorithms. For instance, training an NN using gradient descent with a fixed initialization and stopping criterion would yield infinite mutual information between the training data and the parameters of the NN. Now, typically, there are sources of stochasticity in NN training. First, the initialization is often not fixed, but instead drawn from some distribution. Second, training is usually based on SGD, or one of its variants, rather than deterministic gradient descent. However, characterizing information-theoretic quantities in the presence of these sources of stochasticity is not entirely straightforward. Furthermore, one would still expect the bulk of generalization performance to be present even for deterministic gradient descent—while the stochasticity of SGD, for instance, may provide a marginal benefit, it is unlikely to make the difference between very poor and very good generalization. This was empirically demonstrated by Geiping *et al.* (2022).

An alternative approach builds on the popular hypothesis that the generalization capabilities of an NN are related to the *flatness* of the loss function in the vicinity of its global minima. If the training loss of the NN is not significantly affected when its parameters are perturbed, this indicates some kind of robustness that could lead to good generalization. This is intimately related to the concept of margins, which has previously been successfully used to analyze the performance of support vector machines (Cristianini and Shawe-Taylor, 2000). It is with this motivation that Langford and Caruana (2001) considered stochastic NNs, for which the parameters are randomly drawn from a particular distribution each time the NN is used. The distribution of each parameter is set as an independent Gaussian distribution, whose mean coincided with the underlying deterministic NN and with variance selected to be as large as possible without degrading the training loss by more than a given threshold. Exploiting this randomization, they were able to evaluate PAC-Bayesian generalization bounds, which can be related to the performance of the underlying deterministic NN through parameters such as the margin and Lipschitz properties of the NN. In order to be able to select reasonable parameters for the prior, Langford and Caruana (2001) considered a suitable dyadic grid of candidate values, applying a union bound over these to obtain bounds that hold simultaneously for all candidates on the entire grid. This led to bounds that are nonvacuous, and significantly better than known generalization bounds for deterministic networks—although the NNs that were considered by Langford and Caruana (2001) were naturally significantly less complex than what has been used in recent years.

This approach was adapted to more modern settings by Dziugaite and Roy (2017). While Langford and Caruana (2001) performed a sensitivity analysis for each parameter separately, this is not tenable for large NNs. Instead, given a trained NN, Dziugaite and Roy (2017) selected the weight distributions by directly optimizing a PAC-Bayesian bound, using Corollary 5.4 as a starting point. By using the relaxation obtained via Pinsker's inequality, replacing the training loss with a convex surrogate, fixing the prior to be a Gaussian distribution centered on the underlying deterministic network, and restricting the posterior to be an isotropic Gaussian, they obtained a training objective that can be optimized via gradient-based methods. The underlying motivation for why this procedure is successful is, as already indicated, the hypothesized flatness of the loss landscape around minimizers of the training loss. While certain measures of flatness have been criticized as insufficient to explain generalization, since they can be arbitrarily altered through reparameterizations that do not affect the neural network itself (Dinh *et al.*, 2017), measuring flatness through the relative entropy avoids such drawbacks. Indeed, the relative entropy is invariant under parameter transformations.

This idea was further developed by Dziugaite *et al.* (2021), who pointed out the crucial role that data-dependent priors, discussed in Section 5.2.3, can play in the tightness of PAC-Bayesian bounds, as observed earlier by, *e.g.*, Ambroladze *et al.* (2006) and Mhammedi *et al.* (2019). In fact, as demonstrated in Dziugaite *et al.* (2021, Lemma 3.3), there exist learning settings for which data-dependent priors are necessary in order to obtain a nonvacuous PAC-Bayesian bound.

Motivated by this, Dziugaite *et al.* (2021) proceed to evaluate such data-dependent priors for NNs. Roughly speaking, a fraction α of the training set, $\boldsymbol{Z}_P$, is used to train an NN upon which the prior is based, while the full training set $\boldsymbol{Z}$ is used to train another NN that corresponds to the posterior. In order to obtain a tighter characterization, this is done in such a way that both NNs process the same samples in the initial epochs, since these will have the largest impact on the final weights. Experiments are also performed where the prior is further informed by a ghost sample, which is not used for selecting the posterior, in order to approximate an oracle prior. The use of data-dependent priors leads to tighter bounds than just the use of a ghost sample. Crucially, unlike the aforementioned results, this leads to nonvacuous bounds when the posterior is chosen through a standard SGD-based procedure (with added noise). However, an even tighter bound can be obtained by optimizing the PAC-Bayesian bound via SGD, as shown in Dziugaite *et al.* (2021, Fig. 5). Even tighter results, where bounds with data-dependent priors were directly optimized, were obtained by Pérez-Ortiz *et al.* (2021), who argued that this could potentially be used for self-certified learning, where no separate test set is needed to certify the performance of the learned hypothesis. Still, the utility of these data-dependent priors is not entirely clear. As argued by Lotfi *et al.* (2022, Fig. 1(a)), similar or better bounds can be obtained by simply letting the posterior equal the data-dependent prior, and using the remaining data to obtain an unbiased estimate of the population loss.

8.2.2 Using the CMI Framework

As discussed in Section 6.3, the CMI framework of Chapter 6 can be viewed as an alternative path to data-dependent priors. This was exploited in Hellström and Durisi (2021a,b), wherein an approach similar to that of Dziugaite *et al.* (2021) and Dziugaite and Roy (2017) was used, in that Gaussian distributions centered on the outputs of SGD are set as the posterior and prior. Specifically, given a supersample $\tilde{\boldsymbol{Z}}$ of training samples, half of them are selected to form the training set $\boldsymbol{Z}_{\boldsymbol{S}}$. The mean of the posterior is then found by running SGD for a fixed set of iterations on $\boldsymbol{Z}_{\boldsymbol{S}}$. Next, the true marginal distribution $P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}}$ in Theorem 6.7 is replaced by an auxiliary $Q_{W|\tilde{\boldsymbol{Z}}}$, the mean of which is obtained by averaging the output of SGD trained on a number of samples of $\boldsymbol{Z}_{\boldsymbol{S}}$ with a fixed $\tilde{\boldsymbol{Z}}$. For both the posterior and prior, the variance is set to be as large as possible while not degrading the training loss of the randomized NN too much—similar to Langford and Caruana (2001), but with a uniform choice for all parameters. While this yields similar numerical bounds as Dziugaite *et al.* (2021), there is one notable drawback—the bound cannot be directly optimized, as this would introduce a direct dependence of the posterior on $\boldsymbol{Z}_{\bar{\boldsymbol{S}}}$. This would violate the required conditional independence between $\tilde{\boldsymbol{Z}}$ and W given $\boldsymbol{Z}_{\boldsymbol{S}}$.

All these bounds apply to stochastic networks, with noise added to the parameters, and not to the underlying, deterministic ones that are typically used in practice. While the CMI bounds are finite without this added noise, as guaranteed by the CMI framework, they are typically vacuous. This can be avoided through the use of evaluated or functional CMI (e-CMI or f-CMI). Motivated by the aim of obtaining information-theoretic generalization bounds that depend on the predictions induced by a learning algorithm, instead of the hypothesis itself, Harutyunyan *et al.* (2021) derived several bounds in terms of the f-CMI. To illustrate the benefits of this shift, consider the case of binary classification.

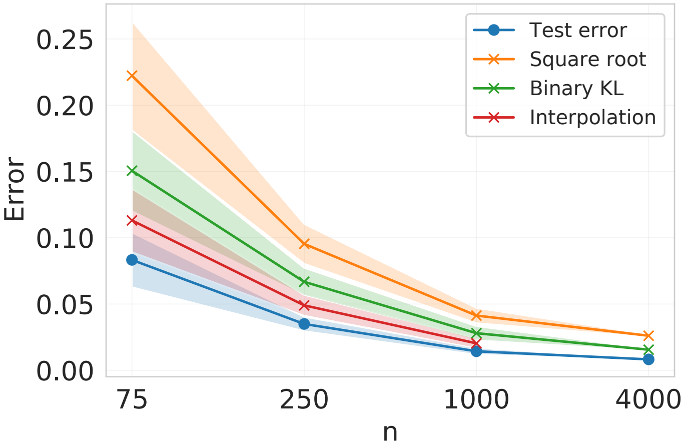

Figure 8.1: Numerical evaluation for a CNN trained on a binary version of MNIST (Hellström and Durisi, 2022a, Fig. 2(a)).

Then, the f-CMI $I(\mathbf{F}; \boldsymbol{S}|\tilde{\boldsymbol{Z}})$ measures the mutual information between the predictions F and the membership vector $\boldsymbol{S}$—two discrete random variables—given the supersample $\tilde{\boldsymbol{Z}}$. Furthermore, for individual-sample f-CMI bounds, $I(F_i, F_{i+n}; S_i|\tilde{\boldsymbol{Z}})$ measures mutual information between binary random variables. This dramatically expands the set of possible scenarios where the information measure, and thus the bound itself, can be small even for deterministic learning algorithms, while being easy to evaluate numerically. Specifically, Harutyunyan *et al.* (2021) evaluated an average, disintegrated, individual-sample f-CMI bound through Monte Carlo estimation, and obtained nearly accurate estimates of the test error for deterministic NNs with relatively small training set sizes. These numerical evaluations were extended to tighter generalization bounds and e-CMI by Hellström and Durisi (2022a). In subsequent work, Wang and Mao (2023c) obtained further improvements through the use of ld-MI.

For a concrete example, consider Fig. 8.1 (Hellström and Durisi, 2022a, Fig. 2(a)). The setting under consideration is binary classification for a version of the MNIST data set, which consists of 32×32 images of handwritten digits. Specifically, the data set is restricted to the digits

4 and 9, and a CNN trained with Adam (a variant of SGD) is used. The plot shows the test error, *i.e.*, the test loss using the $0-1$ loss, along with several upper bounds. Specifically, these are samplewise, disintegrated e-CMI versions of the square-root bound in (6.1), the binary KL bound in (6.9), and the interpolation bound in (6.8). To be explicit, the bounds are, recalling the notation of Section 6.5,

$$L \leq \hat{L} + \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\tilde{\boldsymbol{Z}}}\left[\sqrt{2I^{\tilde{\boldsymbol{Z}}}(\Lambda_i; \boldsymbol{S}_i)}\right] \tag{8.32}$$

$$L \leq \mathbb{E}_{\tilde{\boldsymbol{Z}}}\left[d_2^{-1}\left(\mathbb{E}_{P_{W\boldsymbol{S}|\tilde{\boldsymbol{Z}}}}[L_{\boldsymbol{Z}_{\boldsymbol{S}}}(W)]\,, \frac{1}{n}\sum_{i=1}^{n} I^{\tilde{\boldsymbol{Z}}}(\Lambda_i; \boldsymbol{S}_i)\right)\right]. \tag{8.33}$$

$$L \leq \sum_{i=1}^{n} \frac{I(\Lambda_i; \boldsymbol{S}_i|\tilde{\boldsymbol{Z}})}{n\log(2)}. \tag{8.34}$$

where $d_2^{-1}(q,c) = \sup\{p \in [0,1] : d(q\,||\,\frac{q+p}{2}) \leq c\}$. The disintegrated samplewise e-CMI $I^{\tilde{\boldsymbol{Z}}}(\Lambda_i; \boldsymbol{S}_i)$ is evaluated via sampling: for each $n \in \{75, 250, 1000, 4000\}$, a supersample of $2n$ samples is drawn from the full data set. Half of these are selected to obtain the n training samples, and the network is then trained and evaluated. This is repeated several times to build an empirical distribution of the relevant random variables, which is used to compute the mutual information term via a plug-in estimator. The results show that, whenever it is applicable, the interpolating bound (8.34) is tightest. For $n = 4000$, not all training losses were zero, precluding its use. Thus, the binary KL bound of (8.33) is tightest of the applicable bounds. For all values, it improves on the square-root bound (8.32). Thus, these results demonstrate that the bounds can be estimated and are numerically fairly accurate. For more details and results for other settings, the reader is referred to, for instance, the work of Harutyunyan *et al.* (2021), Hellström and Durisi (2022a), and Wang and Mao (2023c).

Note that, in contrast to the aforementioned bounds for stochastic NNs, these bounds hold only in expectation. While corresponding results can be obtained in probability, this would limit the possibility of using the individual-sample technique, potentially degrading the bounds significantly.

8.2.3 Compression-Based Bounds

An alternative approach to obtaining numerically nonvacuous generalization bounds for NNs is through the lens of *compression* (Arora *et al.*, 2018; Bu *et al.*, 2021). This approach builds on the observation that, often, well-performing NNs can be significantly compressed without noticably affecting their performance. While generalization bounds for the original NN may be far from accurate, applying the same bound to a compressed NN can yield much better results. While these bounds still do not explain the generalization capabilities of the original NN, they can provide guarantees for the compressed counterparts.

This approach was used by Zhou *et al.* (2019), who obtained nonvacuous generalization bounds for NNs by combining off-the-shelf compression algorithms and PAC-Bayesian bounds. The idea is essentially to set the posterior in the PAC-Bayesian bound to be a point mass centered on the output of the combined NN training and compression algorithm, and combine this with a suitably chosen prior on the set of possible hypotheses following the compression step. The specific compression algorithm considered by Zhou *et al.* (2019) is weight pruning, whereby a large number of parameters are set to zero in a way that aims to minimize adversely affecting predictive performance (Han *et al.*, 2016). Finally, in order to further exploit the flatness of the loss surface, Gaussian noise is added to the non-zero weights, similar to the approach taken by Dziugaite and Roy (2017).

This approach was extended in several ways by Lotfi *et al.* (2022), who aimed to leverage these bounds to shed light on various factors behind generalization in NNs. First, they perform training only in a carefully constructed random linear subspace of the parameters, constraining the space of possible hypotheses and thus enabling smaller compressed sizes. Instead of pruning, Lotfi *et al.* (2022) use trainable quantization, whereby the quantization levels and the weights themselves can be learned simultaneously. Furthermore, whereas Zhou *et al.* (2019) considered a prior based on a uniform distribution, Lotfi *et al.* (2022) replaced it with a so-called universal prior, which places greater weight on more compressed hypotheses. This leads to nonvacuous bounds, which can be further tightened through the use of data-dependent priors in the style of Ambroladze *et al.* (2006) and Dziugaite *et al.* (2021). However, Lotfi *et al.* (2022) argue that while this leads to numerically accurate bounds, it does not explain generalization for the full learning procedure: such bounds only compare the posterior to the data-dependent prior, but the question of why the prior is good is left unanswered. Finally, numerical experiments by Lotfi *et al.* (2022) indicate that one possible explanation for why techniques such as transfer learning and the use of symmetries improve generalization is that they improve compressibility.

8.3 Bibliographic Remarks and Additional Perspectives

The results in Section 8.1 are based on the work of Pensia *et al.* (2018). Additionally, information-theoretic bounds for SGLD have also been derived by, for instance, Bu *et al.* (2020), Futami and Fujisawa (2023), Haghifam *et al.* (2020), Issa *et al.* (2023), Li *et al.* (2020), Mou *et al.* (2018), Negrea *et al.* (2019), Wang *et al.* (2021a), and Wang *et al.* (2021b, 2023). By relating the parameter trajectory of SGLD to the corresponding noise-free trajectory of SGD, Neu *et al.* (2021) and Wang and Mao (2022) obtained bounds for SGD. However, as demonstrated by Haghifam *et al.* (2023), current information-theoretic approaches are not sufficient to obtain minimax optimal rates for stochastic convex optimization problems. This was rectified to some extent by Wang and Mao (2023b), who combined the information-theoretic approach with techniques from algorithmic stability.

In addition to the results for NNs that we have discussed so far, several alternative approaches to obtain generalization bounds for neural networks have been explored in the literature, both within the scope of information-theoretic and PAC-Bayesian bounds and beyond it. While a comprehensive overview of all such work is beyond the scope of this monograph, we will mention some of the approaches here. For instance, bounds have been derived based on the norms of the weights of the NN (Bartlett *et al.*, 2017; Neyshabur *et al.*, 2015). A PAC-Bayesian view on this approach was taken by Neyshabur *et al.* (2018), who used the robustness of NNs to parameter perturbations in order to obtain a derandomized bound in terms of a relative entropy that can be evaluated explicitly. Bartlett and Mendelson (2002) derived norm-based bounds for NNs starting from the Rademacher complexity. The connection between PAC-Bayesian bounds and flatness has also been explored by, *e.g.*, Foret *et al.* (2021) and Tsuzuku *et al.* (2020). Several works have derived generalization bounds for NNs trained via SGLD (Bu *et al.*, 2020; Haghifam *et al.*, 2021), and other noisy versions of SGD (Banerjee *et al.*, 2022). Pitas (2020) explored the use of Gaussian posteriors in PAC-Bayesian bounds for NNs, while Dziugaite and Roy (2018a) established a connection to entropy-SGD.

In the limit of infinite width, and under certain conditions on their initialization, NNs can be described as a Gaussian process (Neal, 1994), a correspondence referred to as the NN Gaussian process (NNGP—Lee *et al.*, 2018). For certain loss functions and suitably scaled learning rates, the evolution of the infinitely wide NN during training is also tractable, and is described by the neural tangent kernel (NTK) (Jacot *et al.*, 2018). Pérez *et al.* (2019) combined PAC-Bayesian bounds with the NNGP correspondence to argue that the functions learned by NN tend to be simple in a sense that leads to generalization, and support their arguments by numerically estimating the relevant quantities. Bernstein and Yue (2021) took a similar approach, but derived analytical upper bounds that lead to nonvacuous generalization guarantees. Shwartz-Ziv and Alemi (2020) used the NTK formalism to analytically study many information metrics for NNs, such as $I(W; \boldsymbol{Z})$. Clerico *et al.* (2023), Clerico and Guedj (2024), and Huang *et al.* (2023) extended the NTK formalism to networks trained by optimizing PAC-Bayesian bounds, while Wang *et al.* (2022) explored connections to the information bottleneck.

Viallard *et al.* (2019) used the PAC-Bayesian framework to analyze a particular two-phase procedure to train NNs. Rivasplata *et al.* (2019) considered a broad family of methods for training stochastic NNs by minimizing PAC-Bayesian bounds. Letarte *et al.* (2019) considered NNs with binary activation functions, and used PAC-Bayesian bounds to both formulate a framework for training and to obtain nonvacuous generalization guarantees. Biggs and Guedj (2021) considered ensembling over stochastic NNs, obtaining differentiable PAC-Bayes objectives, while Biggs and Guedj (2022a) derived a de-randomized PAC-Bayesian bound for shallow NNs, using data-dependent priors to get nonvacuous generalization bounds. Zantedeschi *et al.* (2021) used PAC-Bayesian bounds to learn stochastic majority votes, while Nagarajan and Kolter (2019) obtained de-randomized PAC-Bayes bounds via noise-resilience. Tinsi and Dalalyan (2022) obtained tractable bounds for certain aggregated shallow NNs, using a PAC-Bayesian bound with Gaussian priors as the starting point, while Clerico *et al.* (2022a) derived a training algorithm for stochastic NNs without the need for a surrogate loss. Jin *et al.* (2022) discussed how the use of dropout affects PAC-Bayesian generalization bound through the concept of weight expansion. Liao *et al.* (2021) used PAC-Bayes to derive generalization bounds for graph NNs, while Viallard *et al.* (2021) and Xiao *et al.* (2023) derived bounds for adversarial robustness.

Comprehensive surveys of various complexity measures and their connection to generalization can be found in, for instance, the works of Dziugaite *et al.* (2020), Jiang *et al.* (2020), and Neyshabur *et al.* (2017).

9

Alternative Learning Models

So far, we have considered a generic learning model in which the learner has access to n (typically i.i.d.) data points from a fixed data distribution, and the goal is to achieve a small loss on new samples from the same distribution. While this learning model covers many learning settings of interest, it is not all-encompassing. In this chapter, we consider learning problems that do not fit neatly into the generic setting we discussed so far. We will not analyze any of these settings in depth. Our aim is merely to illustrate the wide applicability of the information-theoretic and PAC-Bayesian approaches to generalization.

First, we discuss the setting of meta learning, wherein the learner observes training data from several related tasks, and the goal is to learn how to perform well on a new task. Next, we consider transfer learning, wherein the distribution of the training data is not the same as the distribution of the test data. This is closely related to domain adaptation and out-of-distribution generalization. Following this, we present an information-theoretic generalization bound for federated learning, where a set of distributed nodes separately observe training samples, on the basis of which a composite hypothesis is formed under certain communication constraints. Finally, we look at reinforcement learning, wherein the learner collects observations by interacting with an environment. Specifically, it observes states, takes actions according to a policy, and receives rewards, with the goal of learning a policy that yields high rewards. We conclude by briefly discussing the application of information-theoretic and PAC-Bayesian generalization bounds to online learning, active learning, and density estimation.

9.1 Meta Learning

In typical supervised learning, each learning task is considered in a vacuum: the learner has access to n training samples from the task, and this is all it has to go by. In reality, this is usually not the case: different tasks of interest may have many commonalities. For instance, any computer vision task is based on processing of visual data, which may be similar across many different tasks.

This idea is captured by the framework of meta learning (Baxter, 2000; Caruana, 1997; Thrun and Pratt, 1998). In this setting, we assume that there exists a task space $\mathcal{T}$, paired with a task distribution $P_{\mathcal{T}}$. For each task $\tau \in \mathcal{T}$, there is a corresponding in-task data distribution P_Z^{τ}. In order to form the meta-training set $\hat{\boldsymbol{Z}} \in \mathcal{Z}^{m\times n}$, m tasks are drawn from $P_{\mathcal{T}}$, and for each of these, n samples are drawn from the corresponding P_Z^{τ}. Thus, for each $i \in [m]$, τ_i is drawn independently from $P_{\mathcal{T}}$, and for each $j \in [n]$, $\hat{Z}_{i,j}$ is drawn independently from $P_Z^{\tau_i}$. On this basis, the meta learner aims to find a hyperparameter (or meta hypothesis) $U \in \mathcal{U}$ on the basis of the meta-learning algorithm $P_{U|\hat{\boldsymbol{Z}}}$. This hyperparameter will serve as an additional input to a base learner, allowing it to use information from the meta-training set for new tasks. Specifically, for $\boldsymbol{Z} \in \mathcal{Z}^n$, the base learner is characterized by the conditional distribution $P_{W|\boldsymbol{Z}U}$. The performance of the meta learner is evaluated through the test loss of the base learner on a test task. Specifically, let τ be drawn from $P_{\mathcal{T}}$, independently from $\hat{Z}$, let the "test-training set" $\boldsymbol{Z}^{\tau}$ consist of n i.i.d. samples from P_Z^{τ}, and let the "test-test sample" $Z^{\tau} \sim P_Z^{\tau}$. Then, the average meta-test loss is defined as

$$L = \mathbb{E}_{P_{\hat{\boldsymbol{Z}}} P_{U|\hat{\boldsymbol{Z}}} P_{\boldsymbol{Z}^{\tau}} P_{W|\boldsymbol{Z}^{\tau}U} P_{Z^{\tau}}} \left[\ell(W, Z^{\tau})\right] = \mathbb{E}_{P_W P_{Z^{\tau}}} \left[\ell(W, Z^{\tau})\right]. \qquad (9.1)$$

While the meta learner does not have access to L, it can compute the meta-training loss, defined as

$$\hat{L} = \mathbb{E}_{P_{\hat{Z}} P_{U|\hat{\boldsymbol{Z}}}} \left[\frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{P_{W_i|\hat{\boldsymbol{Z}}_{i,:} U}} \left[\frac{1}{n} \sum_{j=1}^{n} \ell(W_i, \hat{Z}_{i,j}) \right] \right]. \tag{9.2}$$

Here, $\hat{\boldsymbol{Z}}_{i,:} = (\hat{Z}_{i,1}, \dots, \hat{Z}_{i,n})$ denotes the training set for the ith task and W_i is the corresponding hypothesis of the base algorithm. For simplicity, we only focus on generalization bounds in expectation. We can extend all of these results to obtain PAC-Bayesian and single-draw counterparts, by following the approach detailed in Chapter 5.

In the standard learning setting, a key step was to perform a change of measure to handle the dependence between the training data and the hypothesis. In the meta-learning setting, there is an additional dependence between the training data and the hyperparameter. One way to handle this additional dependence is to use a two-step approach, wherein an auxiliary loss is introduced as an intermediate step between the meta-training and meta-population loss. This allows us to obtain generalization bounds by applying two changes of measure, separately: one to relate the meta-training loss to the auxiliary loss, and one to relate the auxiliary loss to the meta-population loss. This allows us to apply standard generalization bounds on the intra-task and inter-task levels separately. However, tighter bounds can be obtained by dealing with them simultaneously. This joint approach leads to the following generalization bound for meta learning, due to Chen *et al.* (2021).

Theorem 9.1. Assume that the loss is σ-sub-Gaussian. Let $\hat{W} = (W_1, \dots, W_m)$ denote the output hypotheses of the base learners for the m training tasks. Then,

$$\left| L - \hat{L} \right| \le \sqrt{\frac{2\sigma^2 I(U, \hat{W}; \hat{\boldsymbol{Z}})}{nm}}. \tag{9.3}$$

Proof. The proof essentially follows immediately by the same approach as was used in the proof of Corollary 4.2, once we make the following observation: the average loss on the meta-training set under the joint distribution of U, $\hat{W}$, and $\hat{\boldsymbol{Z}}$ equals $\hat{L}$. If we instead draw $(U, \hat{W})$ independent from $\hat{\boldsymbol{Z}}$, it equals L. We begin by re-writing the training loss as

$$\hat{L} = \mathbb{E}_{P_{\hat{\boldsymbol{Z}}} P_{U|\hat{\boldsymbol{Z}}}} \left[\frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{P_{W_i|\hat{\boldsymbol{Z}}_{i,:}U}} \left[\frac{1}{n} \sum_{j=1}^{n} \ell(W_i, \hat{Z}_{i,j}) \right] \right] \tag{9.4}$$

$$= \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{P_{W_i|\hat{\boldsymbol{Z}}_{i,:}U} P_{\hat{\boldsymbol{Z}}} P_{U|\hat{\boldsymbol{Z}}}} \left[\frac{1}{n} \sum_{j=1}^{n} \ell(W_i, \hat{Z}_{i,j}) \right]. \tag{9.5}$$

Furthermore, since the tasks and samples are i.i.d.,

$$\frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{P_{W_i|U} P_{\hat{\boldsymbol{Z}}} P_U} \left[\frac{1}{n} \sum_{j=1}^{n} \ell(W_i, \hat{Z}_{i,j}) \right] = \mathbb{E}_{P_W P_{Z^\tau}} [\ell(W, Z^\tau)] = L. \tag{9.6}$$

We conclude the proof by changing measure from $P_{U\hat{W}\hat{Z}}$ to $P_{U\hat{W}} P_{\hat{Z}}$ and using sub-Gaussian concentration. □

The effects of the environment level and in-task level in Theorem 9.1 can be disentangled through the use of the chain rule:

$$\sqrt{\frac{2\sigma^2 I(U, \hat{W}; \hat{\boldsymbol{Z}})}{nm}} = \sqrt{\frac{2\sigma^2 (I(U; \hat{\boldsymbol{Z}}) + I(\hat{W}; \hat{Z}|U))}{nm}} \tag{9.7}$$

$$\leq \sqrt{\frac{2\sigma^2 I(U; \hat{\boldsymbol{Z}})}{nm}} + \sqrt{\frac{2\sigma^2 I(W_1; \hat{\boldsymbol{Z}}_{1,:}|U)}{n}}. \tag{9.8}$$

In the second step, we used the fact that $I(\hat{W}; \hat{Z}|U)$ can be separated as m mutual information terms, one for each task, with the same underlying distributions.

The bound in Theorem 9.1 can be tightened through the use of alternative changes of measure and concentration methods, disintegration, and the individual-sample technique. We will not discuss this explicitly, but instead provide pointers for such extensions and to additional results. PAC-Bayesian bounds for meta learning have been derived, often with a focus on algorithms that minimize these bounds to improve generalization, by, *e.g.*, Amit and Meir (2018), Pentina and Lampert (2014), Rezazadeh (2022), and Rothfuss *et al.* (2021). Information-theoretic bounds were provided by Jose and Simeone (2021a) and Jose *et al.* (2022b), who used a two-step derivation, and Chen *et al.* (2021) who used the one-step derivation above. A CMI formulation of meta learning was introduced by Rezazadeh *et al.* (2021), which was later extended to incorporate one-step derivations, disintegration, and alternative comparator functions by Hellström and Durisi (2022b). Finally, Jose and Simeone (2021c) derived generalization bound that explicitly incorporate task similarity, as measured through, for instance, the relative entropy.

9.2 Out-of-Distribution Generalization and Domain Adaptation

In the standard learning setting, the population loss is defined with respect to the same distribution from which the training set was drawn. While this is a natural assumption to make from a theoretical standpoint, there are many situations in which one expects a distribution shift when deploying a model. There are also scenarios where there is an abundance of data from a surrogate distribution, while there is a lack of data from the actual distribution of interest. This motivates theoretical settings where the population loss is defined with respect to a target distribution, which may differ from the source distribution used to generate the training data.

For the purposes of this discussion, we assume that the sample space factors into a feature space and a label space as $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. The overarching framework, where the only assumption is that the training data is drawn from a source distribution P_Z but we evaluate the model on a target distribution P_Z^T, is usually referred to as *out-of-distribution* (OOD) generalization (Liu *et al.*, 2021a). When the marginal distribution on $\mathcal{X}$ induced by P_Z differs from the one induced by P_Z^T, but the conditional distributions of the label given the features are identical, we refer to this as *domain adaptation* (Kouw and Loog, 2019; Redko *et al.*, 2022). Finally, whenever the learner has access to (partial) samples from the target distribution, we refer to this as *transfer learning*, categorized as *unsupervised* if the learner only has access to unlabelled target features and *supervised* if it has access to full target samples (Weiss *et al.*, 2016). While the definitions of OOD generalization and domain adaptation provided above are fairly established, the term transfer learning is overloaded, and is sometimes used to refer to OOD generalization more broadly, or even to certain variations of meta learning.

For simplicity, we will only consider bounds in expectation. As usual, we denote the training set as $\boldsymbol{Z}$, drawn from $P_{\boldsymbol{Z}} = P_Z^n$, and the output hypothesis from the stochastic algorithm $P_{W|\boldsymbol{Z}}$ as W. Similarly, the average training and population loss with respect to the source distribution are still given by

$$\hat{L} = \mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{\boldsymbol{Z}}(W)], \qquad L = \mathbb{E}_{P_{W\boldsymbol{Z}}}[\mathbb{E}_{P_Z}[\ell(W,Z)]]. \tag{9.9}$$

However, the performance metric that we actually wish to minimize is the average *target* population loss, given by

$$L^T = \mathbb{E}_{P_{W\boldsymbol{Z}}}\left[\mathbb{E}_{P_Z^T}\left[\ell(W,Z^T)\right]\right]. \tag{9.10}$$

9.2.1 Generic OOD Generalization Bounds

Our first approach to obtaining OOD generalization bounds is natural. Since we have already established bounds for the population loss under the source distribution, but are now interested in bounds under the target distribution, we can just apply a change of measure. By a direct application of the Donsker-Varadhan variational representation of the relative entropy, we obtain the following (Wang and Mao, 2023a).

Proposition 9.2. Assume that the loss function is σ-sub-Gaussian under P_Z almost surely under P_W and that $P_Z^T \ll P_Z$. Then,

$$\left|L^T - L\right| \leq \sqrt{2\sigma^2 D(P_Z^T \,||\, P_Z)}. \tag{9.11}$$

Proof. By the Donsker-Varadhan variational representation of the relative entropy in Theorem 3.17, we have for any $\lambda \in \mathbb{R}$

$$D(P_Z^T \,||\, P_Z) \geq \mathbb{E}_{P_Z^T}\left[\lambda\, \mathbb{E}_{P_{W\boldsymbol{Z}}}\left[\ell(W,Z^T)\right]\right] - \log \mathbb{E}_{P_Z}\left[e^{\lambda\, \mathbb{E}_{P_{W\boldsymbol{Z}}}[\ell(W,Z)]}\right]. \tag{9.12}$$

Due to the sub-Gaussianity assumption, we have

$$\begin{aligned} &\log \mathbb{E}_{P_Z}\left[e^{\lambda\, \mathbb{E}_{P_{W\boldsymbol{Z}}}[\ell(W,Z)]}\right] \\ &= \log \mathbb{E}_{P_Z}\left[e^{\lambda(\mathbb{E}_{P_{W\boldsymbol{Z}}}[\ell(W,Z)] - \mathbb{E}_{P_Z}[\mathbb{E}_{P_{W\boldsymbol{Z}}}[\ell(W,Z)]] + \mathbb{E}_{P_Z}[\mathbb{E}_{P_{W\boldsymbol{Z}}}[\ell(W,Z)]])}\right] \qquad (9.13) \\ &\geq \lambda\, \mathbb{E}_{P_Z}[\mathbb{E}_{P_{W\boldsymbol{Z}}}[\ell(W,Z)]] + \frac{\lambda^2\sigma^2}{2}. \qquad (9.14) \end{aligned}$$

By combining these steps and optimizing over λ for the two cases $\lambda > 0$ and $\lambda < 0$, we obtain the final result. □

Proposition 9.2 allows us to turn any generalization bound for standard learning into an OOD generalization bound via the triangle inequality, at the cost of a term depending on $D(P_Z^T \,||\, P_Z)$. This result confirms the intuition that OOD generalization works well if the target and source distribution are similar, with the added specificity that similarity in terms of relative entropy is sufficient. One drawback of the relative entropy is that it requires absolute continuity for finiteness. This can be alleviated to some extent: the role of the source distribution P_Z and target distribution P_Z^T in the derivation above can be swapped, leading to a bound in terms of $D(P_Z \,||\, P_Z^T)$. For this to work, we instead need to assume that the loss function is σ-sub-Gaussian under P_Z^T almost surely under P_W and that $P_Z \ll P_Z^T$.

Unfortunately, there are scenarios where neither of these conditions are satisfied—for instance, if the two distributions have disjoint supports. This motivates bounds in terms of other information measures, such as the Wasserstein distance. The following result follows directly from the Kantorovich-Rubinstein duality.

Proposition 9.3. Assume that the loss is 1-Lipschitz. Then,

$$\left| L^T - L \right| \leq \mathbb{W}_1(P_Z, P_Z^T). \tag{9.15}$$

The benefit of this result is that, unlike for the relative entropy, it remains finite even for the case where the source and target distributions have disjoint support.

9.2.2 Unsupervised Transfer Learning

In the previous section, we derived generic bounds in which minimal assumptions were made on the distributions and task, and the learning algorithm did not have access to any samples from the target distribution. While this led to explicit bounds in terms of discrepancy measures between the source and target distribution, the utility is limited—we cannot minimize these discrepancy measures, and in fact, we do not have any access to the source and target distributions.

In order to gain algorithmic insights, we will now consider unsupervised transfer learning. More precisely, we assume that the sample space factors into a feature space and label space as $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Hence, the target distribution also factors as $P_Z^T = P_X^T P_{Y|X}^T$. Furthermore, we assume that the hypothesis W implements a function $f_W : \mathcal{X} \to \mathcal{Y}$, and that its loss depends on the true label and the corresponding prediction as $\ell(W, Z) = \ell_f(f_W(X), Y)$. In addition to the training set $\boldsymbol{Z}$ drawn from $P_{\boldsymbol{Z}}$, the learning algorithm now also has access to a set of unlabelled features $\boldsymbol{X}^T = (X_1^T, \dots, X_m^T)$, with each element drawn independently from P_X^T. The learning algorithm is now characterized by the conditional distribution $P_{W|\boldsymbol{Z}\boldsymbol{X}^T}$, and the training loss and target population loss are thus given by

$$\hat{L} = \mathbb{E}_{P_{W\boldsymbol{Z}\boldsymbol{X}^T}}\left[L_{\boldsymbol{Z}}(W)\right], \qquad L^T = \mathbb{E}_{P_{W\boldsymbol{Z}\boldsymbol{X}^T}}\left[\mathbb{E}_{P_Z^T}\left[\ell(W, Z^T)\right]\right]. \tag{9.16}$$

Following Wang and Mao (2023a), we can derive bounds on L^T directly from $\hat{L}$, *i.e.*, without going through the source-distribution population loss.

Theorem 9.4. Assume that the loss function is σ-sub-Gaussian under P_Z almost surely under P_W and that $P_Z^T \ll P_Z$. Then,

$$\left|L^T - \hat{L}\right| \le \mathbb{E}_{P_{\boldsymbol{X}^T}}\left[\sqrt{\frac{2\sigma^2 I^{\boldsymbol{X}^T}(W; \boldsymbol{Z})}{n} + 2\sigma^2 D(P_Z^T \,||\, P_Z)}\right]. \tag{9.17}$$

Proof. We begin by considering a specific $\boldsymbol{X}^T$. Then, by the same argument as used in Proposition 9.2, for all $\lambda \in \mathbb{R}$

$$\begin{aligned} &D(P_{WZ_i|X_j^T} \,||\, P_{W|X_j^T} P_Z^T) \\ &\ge \mathbb{E}_{P_{WZ_i|X_j^T}}[\lambda\ell(W, Z_i)] - \mathbb{E}_{P_{W|X_j^T} P_Z^T}\left[\lambda\ell(W, Z^T)\right] - \frac{\sigma^2\lambda^2}{2n}. \end{aligned} \tag{9.18}$$

Now, note that

$$D(P_{WZ_i|X_j^T} \,||\, P_{W|X_j^T} P_Z^T) = I^{X_j^T}(W; Z_i) + D(P_Z \,||\, P_Z^T). \tag{9.19}$$

Hence, by optimizing over λ as before, we get

$$\begin{aligned} &\left|\mathbb{E}_{P_{WZ_i|X_j^T}}[\lambda\ell(W, Z_i)] - \mathbb{E}_{P_{W|X_j^T} P_Z^T}\left[\lambda\ell(W, Z^T)\right]\right| \\ &\le \sqrt{2\sigma^2 I^{X_j^T}(W; Z_i) + D(P_Z \,||\, P_Z^T)} \end{aligned} \tag{9.20}$$

The stated result now follows by decomposing $\left|L^T - \hat{L}\right|$, applying (9.20) termwise, and performing a full-sample relaxation. □

The role of $\boldsymbol{X}^T$ in the disintegrated mutual information here is not entirely clear. Indeed, if we use Jensen's inequality to move the expectation inside the square root, we get

$$\mathbb{E}_{P_{\boldsymbol{X}^T}}\left[\sqrt{I^{\boldsymbol{X}^T}(W;\boldsymbol{Z})}\right] \leq \sqrt{I(W;\boldsymbol{Z}|\boldsymbol{X}^T)}. \tag{9.21}$$

This conditional mutual information is lower-bounded as $I(W;\boldsymbol{Z}|\boldsymbol{X}^T) \geq I(W;\boldsymbol{Z})$. If we had not fixed $\boldsymbol{X}^T$ at the beginning of the derivation, and had instead just averaged it out, we would have obtained a generalization bound in terms of $I(W;\boldsymbol{Z})$, where the role of $\boldsymbol{X}^T$ is ignored, as was done by Jose and Simeone (2021d). However, the relation between $\mathbb{E}_{P_{\boldsymbol{X}^T}}\left[\sqrt{I^{\boldsymbol{X}^T}(W;\boldsymbol{Z})}\right]$ and $I(W;\boldsymbol{Z})$ is not clear. Indeed, the unlabelled target features could potentially be used to decrease the information measure that appears in the bound, as discussed by Wang and Mao (2023a).

Still, this does not address the term $D(P_Z \,||\, P_Z^T)$ in Theorem 9.4. This term can be controlled to some extent when the function implemented by the learning algorithm can be expressed as a composition $f_W = g_W \circ h_W$, where $h_W : \mathcal{X} \to \mathcal{R}$ is a mapping to a *representation* space $\mathcal{R}$ and $g_W : \mathcal{R} \to \mathcal{Y}$ is the final mapping to the prediction. Here, $f_W(\cdot)$ can for instance be an N-layer neural network, where $h_W(\cdot)$ consists of the first $N-k$ layers and $g_W(\cdot)$ consists of the remaining k layers, for some $k \in [N]$. For this setting, we can try to align the distributions on the representation induced by the source and target distributions.

For the purposes of this discussion, we will look at the relative entropy $D(P_Z^T \,||\, P_Z)$, but similar techniques can be applied to, *e.g.*, the Wasserstein distance. First, consider a fixed function $h : \mathcal{X} \to \mathcal{R}$, and let $P_{h_W}^T$ denote the pushforward of P_X^T with respect to h—that is, the distribution on $\mathcal{R}$ induced by h acting on P_X^T—and similarly for P_{h_W}. Furthermore, let $P_{Y|h_W}^T$ and $P_{Y|h_W}$ denote the conditional target and source distributions for the label, given the representation. Then, for a fixed W, we have

$$L^T(W) = \mathbb{E}_{P_Z^T}[\ell(W, Z)] = \mathbb{E}_{P_{h_W}^T P_{Y|h_W}^T}[\ell(g_W(h_W(X)))] , \tag{9.22}$$

$$L(W) = \mathbb{E}_{P_Z}[\ell(W, Z)] = \mathbb{E}_{P_{h_W} P_{Y|h_W}}[\ell(g_W(h_W(X)))] . \tag{9.23}$$

Therefore, by repeating the argument of Proposition 9.2 with this re-formulation at the start, we obtain

$$\left|L^T - L\right| \le \mathbb{E}_{P_W}\left[\sqrt{2\sigma^2 D(P_{h_W}^T P_{Y|h_W}^T \,||\, P_{h_W} P_{Y|h_W})}\right] . \tag{9.24}$$

The result in Theorem 9.4 can be adapted similarly. Next, note that the relative entropy can be decomposed as

$$D(P_{h_W}^T P_{Y|h_W}^T || P_{h_W} P_{Y|h_W}) = D(P_{h_W}^T || P_{h_W}) + D(P_{Y|h_W}^T || P_{Y|h_W}). \tag{9.25}$$

Consequently, we have two components of the discrepancy measure: the representation discrepancy $D(P_{h_W}^T || P_{h_W})$ and the conditional discrepancy $D(P_{Y|h_W}^T || P_{Y|h_W})$. The representation discrepancy is something that we actually *can* aim to minimize by suitably designing our learning algorithm. While we do not have access to the underlying feature distribution for neither the source nor the target, we have empirical estimates based on the source features in $\boldsymbol{Z}$ and the unlabelled target features $\boldsymbol{X}^T$. Thus, as part of choosing W, we can aim to minimize the discrepancy between the pushforward of these empirical source and target feature distributions with respect to h_W.

Now, the relative entropy between the two conditional distributions is not under our control in the same sense, but there are situations where its contribution can be minor. For the setting of domain adaptation, this term will be zero, as we assume that the conditional distribution on the label given the features is identical for the source and target distributions. This implies that the corresponding pushforward measures are also equal. Under some additional assumptions, this relative entropy can also be replaced by a term that is small for settings of practical relevance. Specifically, as shown by Wang and Mao (2023a, Thm. 4.2), if we assume that the loss is symmetric and satisfies the triangle inequality, we find that, for any fixed W,

$$L^T(W) - L(W) \le \sqrt{2\sigma^2 D(P_X^T \,||\, P_X)} + \min_{w^* \in \mathcal{W}} \{L^T(w^*) + L(w^*)\}. \tag{9.26}$$

Thus, the relative entropy between the conditional distributions can be replaced by the smallest possible sum of source and target population losses. If transfer learning is to be successful in the sense that we should be able to find a hypothesis that works well for both the source and the target distributions—even given oracle knowledge of the true distributions—this quantity has to be small.

We conclude this section by presenting a generalization bound for *supervised* transfer learning, where the learning algorithm has access to labelled data from the target distribution. This bound is in terms of the f-mutual information and uses total variation as discrepancy measure, and is due to Wu *et al.* (2022a). We shall assume that, in addition to the source training set $\boldsymbol{Z}$, the learning algorithm also has access to a set of m labelled examples from the target distribution $\boldsymbol{Z}^T = (Z_1^T, \dots, Z_m^T)$, with all elements drawn independently from P_Z^T. Thus, the learning algorithm is characterized by a conditional distribution $P_{W|\boldsymbol{Z}\boldsymbol{Z}^T}$. We define the weighted training loss as

$$\hat{L} = \mathbb{E}_{P_{W\boldsymbol{Z}\boldsymbol{Z}^T}}\left[\frac{\alpha}{m}\sum_{i=1}^{m}\ell(W, Z_i^T)\right] + \mathbb{E}_{P_{W\boldsymbol{Z}\boldsymbol{Z}^T}}\left[\frac{1-\alpha}{n}\sum_{i=1}^{n}\ell(W, Z_i)\right] \tag{9.27}$$

$$= \frac{\alpha}{m}\sum_{i=1}^{m}\mathbb{E}_{P_{WZ_i^T}}\left[\ell(W, Z_i^T)\right] + \frac{1-\alpha}{n}\sum_{i=1}^{n}\mathbb{E}_{P_{WZ_i}}\left[\ell(W, Z_i)\right]. \tag{9.28}$$

Here, the parameter $\alpha \in [0, 1]$ determines the relative emphasis that we place on the data from the target distribution. When $\alpha = 1$, it reduces to the standard training loss for supervised learning. When $\alpha = 0$, we are instead back to a generic OOD setting with no target data to learn from.

Theorem 9.5. Assume that, for any $w \in \mathcal{W}$, the loss is bounded by σ in L_∞-norm, *i.e.*,

$$|\ell(w, Z)|_\infty = \inf\{s : P_Z^T(\ell(w, Z) > s) = 0\} \leq \sigma. \tag{9.29}$$

Then, we have

$$\left|L^T - \hat{L}\right| \le \frac{2\alpha\sigma}{m} \sum_{i\in[m]} \mathrm{TV}(P_{WZ_i}, P_W P_{Z_i^T}) \\ + \frac{2(1-\alpha)\sigma}{n} \sum_{i\in[n]} \Big(\mathrm{TV}(P_{WZ_i}, P_W P_{Z_i}) + \mathrm{TV}(P_Z, P_Z^T)\Big). \quad (9.30)$$

Proof. First, we decompose the generalization gap as

$$\begin{aligned}\left|L^T - \hat{L}\right| &= \left|L^T - \frac{\alpha}{m}\sum_{i=1}^m \mathbb{E}_{P_{WZ_i^T}}\left[\ell(W, Z_i^T)\right] - \frac{1-\alpha}{n}\sum_{i=1}^n \mathbb{E}_{P_{WZ_i}}[\ell(W, Z_i)]\right| \\ &\le \frac{\alpha}{m}\sum_{i=1}^m \left|\mathbb{E}_{P_W P_{Z_i^T}}\left[\ell(W, Z^T)\right] - \mathbb{E}_{P_{WZ_i^T}}\left[\ell(W, Z_i^T)\right]\right| \qquad (9.31) \\ &\quad + \frac{1-\alpha}{n}\sum_{i=1}^n \left|\mathbb{E}_{P_W P_{Z_i^T}}\left[\ell(W, Z^T)\right] - \mathbb{E}_{P_{WZ_i}}[\ell(W, Z_i)]\right|.\end{aligned}$$

Now, the terms in the first sum are individual-sample generalization gaps. By applying Theorem 4.4 to each term, we can bound them as

$$\mathbb{E}_{P_W P_{Z_i^T}}\left[\ell(W, Z^T)\right] - \mathbb{E}_{P_{WZ_i^T}}\left[\ell(W, Z_i^T)\right] \le \mathrm{TV}(P_{WZ_i^T}, P_W P_{Z_i^T}). \quad (9.32)$$

Proceeding similarly with the second sum, we can bound each term as

$$\mathbb{E}_{P_W P_{Z_i^T}}\left[\ell(W, Z^T)\right] - \mathbb{E}_{P_{WZ_i}}[\ell(W, Z_i)] \le \mathrm{TV}(P_{WZ_i}, P_W P_{Z_i^T}). \quad (9.33)$$

To isolate the effect of the distribution shift, we can decompose this last upper bound as

$$\begin{aligned}\mathrm{TV}(P_{WZ_i}, P_W P_{Z_i^T}) &= \frac{1}{2}\int_{\mathcal{W}\times\mathcal{Z}} \left|\mathrm{d}P_{WZ_i} - \mathrm{d}P_W P_{Z_i^T}\right| \qquad (9.34) \\ &\le \frac{1}{2}\int_{\mathcal{W}\times\mathcal{Z}} \left|\mathrm{d}P_{WZ_i} - \mathrm{d}P_W P_{Z_i}\right| \qquad (9.35) \\ &\quad + \frac{1}{2}\int_{\mathcal{W}\times\mathcal{Z}} \left|\mathrm{d}P_W P_{Z_i} - \mathrm{d}P_W P_{Z_i^T}\right| \\ &= \mathrm{TV}(P_{WZ_i}, P_W P_{Z_i}) + \mathrm{TV}(P_Z, P_Z^T). \qquad (9.36)\end{aligned}$$

We obtain the desired result by substituting (9.32), (9.33) and (9.36) into (9.31). □

While we only covered bounds in expectation, many of these results can be extended to PAC-Bayesian and single-draw variants. Further discussion regarding many of these topics, as well as practical algorithms based on these bounds, are provided by Aminian *et al.* (2022a), Wang and Mao (2023a), and Wu *et al.* (2022a).

9.3 Federated Learning

Federated learning is a framework for describing distributed learning, for instance in mobile networks (Kairouz *et al.*, 2021). Specifically, we assume that there are K separate nodes, all having access to their own training set $\boldsymbol{Z}_k = (Z_{k,1}, \dots, Z_{k,n})$ of size n, for each $k \in [K]$. We assume that $Z_{k,i} \sim P_Z$ for all $(k,i) \in [K] \times [n]$, and denote the collection of all training sets as $\boldsymbol{Z} = (\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_K)$. Each node uses a learning algorithm $P_{W_k|\boldsymbol{Z}_k}$ to generate the hypothesis W_k on the basis of $\boldsymbol{Z}_k$. These local models are then combined to form the final model W through an aggregation algorithm $P_{W|W_1,\dots,W_k}$. A common choice is to use averaging, so that $W = \frac{1}{K}\sum_{k=1}^{K} W_k$. Composing the local learning algorithms and the aggregation algorithm induces a conditional distribution on W given the full training set $\boldsymbol{Z}$, denoted as $P_{W|\boldsymbol{Z}}$. As usual, our aim is to bound the population loss $L_{P_Z}(W)$.

One way to obtain generalization bounds is simply to consider $P_{W|\boldsymbol{Z}}$ as a learning algorithm acting on nK samples, and use a generalization bound for standard supervised learning. Alternatively, assuming that the aggregation algorithm performs averaging and that the loss is convex, we have

$$L_{P_Z}(W) = \mathbb{E}_{P_Z}\left[\ell\left(\frac{1}{K}\sum_{k=1}^{K} W_k, Z\right)\right] \tag{9.37}$$

$$\leq \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{P_Z}[\ell(W_k, Z)]\,. \tag{9.38}$$

This allows us to apply a standard generalization bound for each node. Neither of these approaches, as noted by Barnes *et al.* (2022), exploits the specific structure of federated learning, except potentially implicitly through the information measures that appear in the bounds. We will therefore focus here on the result in Barnes *et al.* (2022, Thm. 4), in which an explicit improved dependence on the number of nodes K is achieved.

To this end, we need to assume that the loss can be described as a *Bregman divergence*. Specifically, for a continuously differentiable and strictly convex function $f : \mathbb{R}^m \to \mathbb{R}$, the Bregman divergence between two points $p, q \in \mathbb{R}^m$ is defined as

$$\mathcal{B}_f(p,q) = f(p) - f(q) - \langle \nabla f(q), p - q \rangle, \tag{9.39}$$

where $\langle \cdot, \cdot \rangle$ is the inner product. Notably, this includes the squared loss, which is obtained by setting $f(\cdot)$ to be the squared two-norm. With this, the following can be established.

Theorem 9.6. Assume that the loss function is a Bregman divergence $\ell(w,z) = \mathcal{B}_f(w,z)$. Furthermore, assume that $\ell(w,Z)$ is σ-sub-Gaussian under P_Z for all $w \in \mathcal{W}$. Then, if $W = \frac{1}{K}\sum_{k=1}^{K} W_k$,

$$\mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{P_Z}(W) - L_{\boldsymbol{Z}}(W)] \leq \frac{1}{K^2}\sum_{k\in[K]}\sqrt{\frac{I(W_k;\boldsymbol{Z}_k)}{n}}. \tag{9.40}$$

Proof. Let $\boldsymbol{Z}' = (\boldsymbol{Z}'_1, \dots, \boldsymbol{Z}'_K)$ be an independent copy of $\boldsymbol{Z}$, and let $\boldsymbol{Z}^{(k,i)}$ equal $\boldsymbol{Z}$ for all elements except $Z^{(k,i)}_{k,i} = Z'_{k,i}$. Then, we have (Shalev-Shwartz *et al.*, 2010, Lemma 11)

$$\begin{aligned}\mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{P_Z}(W)] &= \frac{1}{nK}\sum_{k,i}\mathbb{E}_{P_{W\boldsymbol{Z}}P_{\boldsymbol{Z}'}}\Big[\ell(W,Z'_{k,i})\Big] \qquad (9.41)\\ &= \frac{1}{nK}\sum_{k,i}\mathbb{E}_{P_{W\boldsymbol{Z}}P_{\boldsymbol{Z}'}}\Big[f(W) - f(Z'_{k,i}) - \langle \nabla f(Z'_{k,i}), W - Z'_{k,i}\Big],\end{aligned}$$

since $Z'_{k,i}$ is independent from W. Here, the summation indices implicitly run over $k \in [K]$ and $i \in [n]$. Let $W^{k,i}$ be drawn according to $P_{W^{k,i}|\boldsymbol{Z}^{(k,i)}}$. Then,

$$\begin{aligned}\mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{\boldsymbol{Z}}(W)] &= \frac{1}{nK}\sum_{k,i}\mathbb{E}_{P_{W^{k,i}\boldsymbol{Z}\boldsymbol{Z}'}}\Big[\ell(W^{k,i},Z'_{k,i})\Big] \qquad (9.42)\\ &= \frac{1}{nK}\sum_{k,i}\mathbb{E}_{P_{W^{k,i}\boldsymbol{Z}\boldsymbol{Z}'}}\Big[f(W^{k,i}) - f(Z'_{k,i}) \qquad (9.43)\\ &\qquad - \langle \nabla f(Z'_{k,i}), W^{k,i} - Z'_{k,i}\Big],\end{aligned}$$

since $Z'_{k,i}$ is in the training set of $W^{k,i}$. It follows that

$$\begin{aligned}&\mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{P_Z}(W) - L_{\boldsymbol{Z}}(W)]\\ &\qquad = \frac{1}{nK}\sum_{k,i}\mathbb{E}_{P_{WW^{k,i}\boldsymbol{Z}\boldsymbol{Z}'}}\left[\langle \nabla f(Z'_{k,i}), W^{k,i} - W\rangle\right]. \end{aligned} \tag{9.44}$$

Here, we used that $\mathbb{E}_{P_W}[f(W)] = \mathbb{E}_{P_{W^{k,i}}}\left[f(W^{k,i})\right]$ since W and $W^{k,i}$ have the same marginal distributions. The key observation that leads to the improved dependence on K, compared to an approach using (9.38), is that W and $W^{k,i}$ are the average of K sub-models, but they differ only in the kth sub-model. Hence, $W^{k,i} - W = \frac{1}{K}(W_k^i - W_k)$, where W_k^i denotes the kth submodel trained on $\boldsymbol{Z}_k^{(i)}$. Therefore,

$$\begin{aligned}&\mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{P_Z}(W) - L_{\boldsymbol{Z}}(W)]\\ &\qquad = \frac{1}{nK^2}\sum_{k,i}\mathbb{E}_{P_{WW^{k,i}\boldsymbol{Z}\boldsymbol{Z}'}}\left[\langle \nabla f(Z'_{k,i}), W_k^i - W_k\rangle\right]. \end{aligned} \tag{9.45}$$

Hence, we can conclude that

$$\mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{P_Z}(W) - L_{\boldsymbol{Z}}(W)] = \frac{1}{K^2}\sum_{k\in[K]}\mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{P_Z}(W_k) - L_{\boldsymbol{Z}_k}(W_k)]\,. \tag{9.46}$$

We obtain the desired result by applying Corollary 4.2. □

If $z = (x, y)$, this result also holds if $\ell(w, (x, y)) = \mathcal{B}_f(\langle w, x\rangle, y)$, with a nearly identical proof. Intuitively, the improved dependence on K arises because the dependence of the final hypothesis W on any individual sample is dampened by $1/K$ due to the averaging. Naturally, this result can be extended to incorporate disintegration, the individual-sample technique, or by using other generalization bounds than Corollary 4.2 in the proof. For further discussion and extensions of these bounds, see for instance the work of Barnes *et al.* (2022) and Yagli *et al.* (2020).

9.4 Reinforcement Learning

So far, we have assumed that the training data is independent from the learning algorithm. In this section, we instead look at reinforcement learning, wherein the learner collects observations by taking observation-dependent actions in an environment. Specifically, in Section 9.4.1, we present extensions of PAC-Bayesian bounds from i.i.d. data to martingales, which allows us to capture some of the interactions that occur in reinforcement learning. Then, in Section 9.4.2, we discuss information-theoretic bounds for Markov decision processes (MDP), which constitute an important class of reinforcement learning problems.

9.4.1 PAC-Bayesian Bounds for Martingales

We begin by presenting a PAC-Bayesian bound for martingales (described in Section 3.3.4) due to Seldin *et al.* (2012b). This can be used to apply generalization bounds like those in Section 5.2 developed for i.i.d. training samples to various types of interactive settings.

Theorem 9.7. Let M_i for $i \in [n]$ be a martingale sequence of random functions $M_i : \mathcal{W} \to [-1,1]$ such that $\mathbb{E}[M_{i+1}(w)|\boldsymbol{M}_{\leq i}(w)] = 0$ for all $w \in \mathcal{W}$, where $\boldsymbol{M}_{\leq i}(w) = (M_1(w), \dots, M_i(w))$. Suppose that the randomness of each M_i is captured by a random variable Z_i, and let $\bar{M}_t = \sum_{i=1}^t M_i$ and $\boldsymbol{Z} = (Z_1, \dots, Z_n)$. Fix a prior distribution Q_W on $\mathcal{W}$ and a $\delta \in (0,1)$. Then, for every distribution $P_{W|\boldsymbol{Z}}$ on $\mathcal{W}$, with probability at least $1-\delta$ over $P_{\boldsymbol{Z}}$,

$$\left|\mathbb{E}_{P_{W|\boldsymbol{Z}}}\left[\frac{\bar{M}_n(W)}{n}\right]\right| \leq \sqrt{\frac{D(P_{W|\boldsymbol{Z}}\,||\,Q_W) + \log\frac{4en}{\delta}}{2n}}. \tag{9.47}$$

Proof. By the Donsker-Varadhan variational representation of the relative entropy, we have, for a fixed $\lambda > 0$,

$$\mathbb{E}_{P_{W|\boldsymbol{Z}}}\left[\frac{\lambda\bar{M}_n(W)}{n}\right] \leq D(P_{W|\boldsymbol{Z}}\,||\,Q_W) + \log\mathbb{E}_{Q_W}\left[e^{\frac{\lambda\bar{M}_n(W)}{n}}\right]. \tag{9.48}$$

By Markov's inequality, we have with probability at least $1-\delta$

$$\log\mathbb{E}_{Q_W}\left[e^{\frac{\lambda\bar{M}_n(W)}{n}}\right] \leq \log\mathbb{E}_{Q_W P_Z}\left[\frac{1}{\delta}e^{\frac{\lambda\bar{M}_n(W)}{n}}\right] \tag{9.49}$$

$$\leq \log\frac{1}{\delta} + \frac{\lambda^2}{8n}, \tag{9.50}$$

where the last step is due to Theorem 3.34. After repeating this argument for $-\bar{M}_n$ and using the union bound, we find that with probability at least $1-\delta$,

$$\left|\mathbb{E}_{P_{W|\boldsymbol{Z}}}\left[\frac{\bar{M}_n(W)}{n}\right]\right| \leq \frac{D(P_{W|\boldsymbol{Z}} \,||\, Q_W) + \log\frac{2}{\delta}}{\lambda} + \frac{\lambda}{8n}. \tag{9.51}$$

To complete the proof, we need to select λ. We will do this by optimizing the bound over a grid of candidate values, using a union bound to ensure that the result is valid for all possible values.[1] First, note that if $D(P_{W|\boldsymbol{Z}} \,||\, Q_W) > 2n$, the right-hand side of (9.51) is lower-bounded by 1 for all λ, meaning that the resulting bound is vacuous (since $\bar{M}_n(W) \leq n$). Hence, the result in (9.47) holds trivially in this case. Thus, we only consider $D(P_{W|\boldsymbol{Z}} \,||\, Q_W) \leq 2n$. Specifically, assume that $D(P_{W|\boldsymbol{Z}} \,||\, Q_W) \in [k-1, k]$ for $k \in [2n]$. Then, by (9.51), we have

$$\left|\mathbb{E}_{P_{W|\boldsymbol{Z}}}\left[\frac{\bar{M}_n(W)}{n}\right]\right| \leq \frac{k + \log\frac{2}{\delta}}{\lambda} + \frac{\lambda}{8n}. \tag{9.52}$$

For a fixed k, this is minimized by $\lambda = 2\sqrt{2n(k + \log\frac{2}{\delta})}$, which gives

$$\left|\mathbb{E}_{P_{W|\boldsymbol{Z}}}\left[\frac{\bar{M}_n(W)}{n}\right]\right| \leq \sqrt{\frac{k + \log\frac{2}{\delta}}{2n}}. \tag{9.53}$$

By the union bound, this holds simultaneously for $k \in [2n]$ with probability at least $1 - 2n\delta$. Hence, by substituting δ with $\delta/(2n)$, noting that $k \leq D(P_{W|\boldsymbol{Z}} \,||\, Q_W) + 1$,

$$\left|\mathbb{E}_{P_{W|\boldsymbol{Z}}}\left[\frac{\bar{M}_n(W)}{n}\right]\right| \leq \sqrt{\frac{D(P_{W|\boldsymbol{Z}} \,||\, Q_W) + 1 + \log\frac{4n}{\delta}}{2n}} \tag{9.54}$$

with probability at least $1-\delta$. From this, the desired result follows. □

By suitably selecting $\bar{M}_i$—for instance, as the difference between the loss for a training instance and its expectation—this bound can be instantiated for various settings with martingale data, extending

[1] In the original proof, Seldin *et al.* (2012b) use a dyadic grid and a weighted union bound over an infinite range. We restrict ourselves to a finite range, similar to Rodríguez-Gálvez *et al.* (2023), in order to simplify the proof.

the applicability of PAC-Bayesian bounds beyond i.i.d. data. For instance, Seldin *et al.* (2011, 2012a) apply these bounds to the case of multiarmed bandits. It is worth noting that Seldin *et al.* (2012b) derive additional bounds using martingale versions of the concentration for binary relative entropy in Theorem 3.29 as well as Bernstein's inequality.

9.4.2 Markov Decision Processes

In reinforcement learning, the learner is viewed as an "agent" that interacts with an environment and takes actions according to a strategy, also known as policy, obtaining rewards on this basis. The goal of this is to learn a good policy for how to select actions depending on the state of the environment. A defining characteristic of reinforcement learning is that the environment is only partially observed through the agent's interaction with it. A specific example of this is the setting of contextual bandits, where the PAC-Bayesian bounds for martingales can be applied, as demonstrated by Seldin *et al.* (2011). Here, following Gouverneur *et al.* (2022), we will focus on Bayesian regret in an MDP, presenting a bound that extend the result obtained by Xu and Raginsky (2022) for supervised learning.

In order to formally describe an MDP, we need the following definitions. We let $\mathcal{S}$ denote a set of states, let $\mathcal{A}$ denote a set of actions, and let $\mathcal{Y}$ denote a set of outcomes. At each time $t \in [T]$, the learner observes the state $S_t \in \mathcal{S}$ and takes an action $A_t \in \mathcal{A}$, after which the environment produces an outcome $Y_t \in \mathcal{Y}$. This leads to the reward $R_t = r(Y_t, A_t) \in \mathbb{R}$. The environment is characterized by a random variable $\theta \in \Theta$, drawn according to P_θ. More specifically, it consists of a transition kernel $P_{S_{t+1}|S_t,A_t,\theta}$, an outcome kernel $P_{Y_t|S_t,\theta}$, an initial state distribution $P_{S|\theta}$, from which S_1 is drawn, and the reward function $r : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$. The stochastic mapping from the state S_t and action A_t to the reward R_t is characterized by the kernel $P_{R_t|S_t,A_t,\theta}$. The goal is to learn a policy $\varphi = \{\varphi_t : \mathcal{S} \times (\mathcal{S}, \mathcal{A}, \mathbb{R})^t \to \mathcal{A}\}_{t\in[T]}$, which selects an action A_t on the basis of S_t and the observed history $H_{<t} = (H_1, \ldots, H_{t-1})$, where $H_t = (S_t, A_t, R_t)$. Specifically, the policy should be chosen to obtain a high cumulative expected re-

ward $r_c(\varphi)$, defined as

$$r_c(\varphi) = \mathbb{E}\left[\sum_{t\in[T]} r(Y_t, \varphi_t(S_t, H_{\le t}))\right]. \tag{9.55}$$

We refer to the maximal expected cumulative reward as the Bayesian cumulative reward, and denote it by $R_c = \sup_\varphi r_c(\varphi)$, where the supremum is taken over all policies that lead to a finite expectation in (9.55). We will compare this to the maximal expected cumulative reward that can be obtained by an oracle that has knowledge of θ. Specifically, we consider decision rules $\psi = \{\psi_t : \mathcal{S}\times\Theta \to \mathcal{A}\}_{t\in[T]}$ and define the oracle Bayesian cumulative reward as

$$R_B^o = \sup_\psi \mathbb{E}\left[\sum_{t\in[T]} r(Y_t, \psi_t(S_t, \theta))\right]. \tag{9.56}$$

We let $\psi^* = \{\psi_t^*\}_{t\in[T]}$ denote the policy that achieves the supremum in (9.56), and assume that it exists. With this, we are ready to define the key quantity that we wish to bound: the minimum Bayesian regret (MBR) given by

$$\text{MBR} = R_B^o - R_c. \tag{9.57}$$

This quantity is the difference between the reward that is obtainable based only on observing the system through interactions and the one that is obtainable when the underlying system parameters are known.

In order to bound the MBR, we will consider a specific learning algorithm, related to Thompson sampling (Russo and Van Roy, 2016; Thompson, 1933). One approach to selecting ϕ_t is to use $H_{\le t}$ to compute an estimate $\hat{\theta}_t$ through a kernel $P_{\hat{\theta}_t|H_{\le t}}$, and then select an action on the basis of $(S_t, \hat{\theta}_t)$. Since this is a special instance of a learning algorithm, the resulting cumulative expected reward cannot be greater than the

Bayesian cumulative reward.

$$R_c = \sup_{\varphi} \mathbb{E}\left[\sum_{t\in[T]} r(Y_t, \varphi_t(S_t, H_{\le t}))\right] \tag{9.58}$$

$$\ge \sup_{\psi} \mathbb{E}\left[\sum_{t\in[T]} r(Y_t, \psi_t(S_t, \hat{\theta}_t))\right] \tag{9.59}$$

$$\ge \mathbb{E}\left[\sum_{t\in[T]} r(Y_t, \psi_t^*(S_t, \hat{\theta}_t))\right]. \tag{9.60}$$

We now introduce Y_t^* and S_t^* as the outcomes and states that are obtained through ψ^* acting on the MDP with the true θ as input. Similarly, we let $\hat{Y}_t$, $\hat{S}_t$, and $\hat{H}_t$ denote the outcomes, states, and histories that are obtained through ψ^* acting on the MDP with the estimated $\{\hat{\theta}_t\}_{t\in[T]}$ as input. Now, by expanding the expression above, we find that the MBR can be bounded as

$$\mathrm{MBR} \le R_B^o - \mathbb{E}\left[\sum_{t\in[T]} r(Y_t, \psi_t^*(S_t, \hat{\theta}_t))\right] \tag{9.61}$$

$$= \sum_{t\in[T]} \mathbb{E}_{P_{\theta\hat{\theta}_t\hat{H}_{\le t}}}\left[\mathbb{E}_{P_{Y_t^* S_t^* \hat{Y}_t \hat{S}_t|\theta\hat{\theta}_t\hat{H}_{\le t}}}\left[r(Y_t^*, \psi_t^*(S_t^*, \theta)) - r(\hat{Y}_t, \psi_t^*(\hat{S}_t, \hat{\theta}_t))\right]\right].$$

Now, observe that the following Markov chain holds:

$$Y_t^*, S_t^*) - \theta - (\hat{Y}_t, \hat{S}_t) - \hat{H}_{\le t} - \hat{\theta}_t. \tag{9.62}$$

From this, it follows that for each $t \in [T]$, the first term of the inner expectation is distributed according to $P_{Y_t^*, S_t^*|\theta}$, while the second is distributed according to $P_{\hat{Y}_t, \hat{S}_t|H_{\le t}}$. Therefore, we can use change of measure techniques to relate the two terms, by following the same arguments as in Chapter 4 (and in particular, Section 4.2). This leads to the following result (Gouverneur *et al.*, 2022, Prop. 1).

Theorem 9.8. Assume that, for all $t \in [T]$, $r(\hat{Y}_t, \psi_t^*(\hat{S}_t, \theta))$ is σ_t^2-sub-Gaussian under $P_{\hat{Y}_t, \hat{S}_t|\hat{H}_{\le t}}$ for all $\theta \in \Theta$. Then,

$$\mathrm{MBR} \le \sum_{t\in[T]} \mathbb{E}_{P_{\theta\hat{H}_{\le t}}}\left[\sqrt{2\sigma_t^2 D(P_{Y_t^*, S_t^*|\theta} \,||\, P_{\hat{Y}_t, \hat{S}_t|\hat{H}_{\le t}})}\right]. \tag{9.63}$$

More discussion of these results, including applications to special cases and results in terms of the Wasserstein distance, can be found in the work of Gouverneur *et al.* (2022).

9.5 Bibliographic Remarks and Additional Perspectives

The result in Theorem 9.1 is due to Chen *et al.* (2021). Information-theoretic generalization bounds for meta learning can also be found in the work of Jose and Simeone (2021a) and Jose *et al.* (2022b), and were extended to the case of e-CMI in Hellström and Durisi (2022b). Additional works that provide PAC-Bayesian and information-theoretic generalization bounds for meta learning include, *e.g.*, Amit and Meir (2018), Farid and Majumdar (2021), Flynn *et al.* (2022), Jose *et al.* (2022a), Liu *et al.* (2021b), Meunier and Alquier (2021), Pentina and Lampert (2014), Rezazadeh (2022), Riou *et al.* (2023), and Rothfuss *et al.* (2021). The bounds for OOD generalization in Propositions 9.2 and 9.3 and Theorem 9.4 are due to Wang and Mao (2023a), while Theorem 9.5 is due to Wu *et al.* (2022a). Jose *et al.* (2022b) considered a combination of transfer learning and meta learning, while Jose and Simeone (2023) analyzed transfer learning for quantum classifiers. Additional results for transfer learning and domain adaptation can be found in the works of Achille *et al.* (2021), Aminian *et al.* (2022b), Bu *et al.* (2022), Germain *et al.* (2016b), and Jose and Simeone (2021c). Relatedly, He *et al.* (2022) derived bounds for iterative semi-supervised learning. Theorem 9.6 is due to Barnes *et al.* (2022), with earlier work by Yagli *et al.* (2020). Sefidgaran *et al.* (2022a) derived generalization bounds for distributed learning using rate-distortion techniques. The extension of PAC-Bayesian bounds to martingales in Theorem 9.7 is due to Seldin *et al.* (2012b); Seldin *et al.* (2011) applied these to contextual bandits. Theorem 9.8 is due to Gouverneur *et al.* (2022). Additional PAC-Bayesian results for reinforcement learning can be found in the work of Fard and Pineau (2010) and Wang *et al.* (2019b).

We conclude by mentioning alternative learning models, and their connections to PAC-Bayesian and information-theoretic generalization bounds. Seeger (2002) applied PAC-Bayesian bounds to Gaussian process classification, while Shawe-Taylor and Hardoon (2009) considered the problem of maximum entropy classification. Unsupervised learning models, such as various types of clustering, were studied by, *e.g.*, Higgs and Shawe-Taylor (2010), Li *et al.* (2018), and Seldin and Tishby (2010). Alquier and Lounici (2011) considered the sparse regression model in high dimension, while Guedj and Robbiano (2018) derived PAC-Bayesian bounds for the bipartite ranking problem in high dimension. Ralaivola *et al.* (2010) derived bounds for non-i.i.d. data, with applications to certain ranking statistics, while Li *et al.* (2013) extended PAC-Bayesian bounds to the nonadditive ranking risk. Jose and Simeone (2021b) used PAC-Bayesian bounds to analyze machine unlearning, where a learning algorithm has to "forget" specific samples. Online learning, where the learner has to sequentially select hypotheses to minimize losses set by a potentially adversarial environment (a recent introduction is provided by Orabona, 2023), is intimately related to PAC-Bayesian and information-theoretic bounds. In particular, there is a formal relationship between the Gibbs posterior and the exponential weights algorithm. PAC-Bayesian bounds for a version of online learning were studied by Haddouche and Guedj (2022). Recently, Lugosi and Neu (2022, 2023) established a method for converting regret bounds from online learning to PAC-Bayesian and information-theoretic bounds, allowing them to (essentially) recover established results and derive new ones. Finally, Sharma *et al.* (2023) exploited PAC-Bayesian generalization bounds in the context of inductive conformal prediction, allowing the calibration data set to be used for learning the hypothesis and score function.

10

Concluding Remarks

In this monograph, we provided a broad overview of information-theoretic and PAC-Bayesian generalization bounds. We highlighted the connection between these fields; presented a wide array of bounds for different settings in terms of different information measures; detailed analytical applications of the bounds to specific learning algorithms; discussed recent applications to iterative methods and neural networks; and covered extensions to alternative settings. We hope that this exposition demonstrates the versatility and potential of the information-theoretic approach to generalization results.

Still, there are many unanswered questions and directions to explore. On the one hand, as shown by Haghifam *et al.* (2021, 2023), there are certain settings for which the information-theoretic approaches discussed in this monograph yield provably suboptimal bounds. On the other hand, there are bounds in terms of the evaluated mutual information that equal the population loss for interpolating settings (Haghifam *et al.*, 2022; Wang and Mao, 2023c), as discussed in Section 6.5, and by appropriately adapting standard information-theoretic bounds, optimal characterizations of the generalization gap in the Gaussian location model can be derived (Zhou *et al.*, 2023a). This raises the question of which settings the information-theoretic approach to generalization is suitable for, and whether or not it can be extended further through new ideas, or whether alternative approaches are necessary.

As discussed in Section 8.2, information-theoretic and PAC-Bayesian bounds have been shown to be numerically accurate for certain settings with neural networks. However, the utility and interpretation of these results is not entirely clear. Dziugaite and Roy (2017) connect their bound to the flatness of the loss landscape; Harutyunyan *et al.* (2021) draw parallels to stability; and Lotfi *et al.* (2022) point towards compressibility, exploring its relation to, *e.g.*, equivariance and transfer learning. Pinning down these connections more precisely, and developing the bounds to such an extent that they can guide model selection *a priori*, are intriguing avenues to explore.

Regarding the structure of the bounds themselves, Foong *et al.* (2021) and Hellström and Guedj (2024) explore the question of what the tightest attainable bound is. For instance, what is the best comparator function to use in Proposition 5.2? Can the $\log \sqrt{n}$ dependence in Corollary 5.4 be removed? Another question is whether the most suitable information measure to use for a given setting can be determined. As discussed throughout, the specific information measure that arises in a bound is just a consequence of the change of measure technique that is used in its derivation.

Finally, there are several interesting extensions to other settings and connections to other approaches that can be explored. While we covered some topics in Chapter 9, the relation to, for instance, active learning, wherein the information carried by a sample is a central quantity (Settles, 2012), and online learning, the analysis of which shares many tools with the information-theoretic approach (Orabona, 2023), is a promising direction. For instance, recently, Lugosi and Neu (2023) showed that any regret bound for online learning implies a corresponding generalization bound for statistical learning.

While this discussion is far from comprehensive, addressing these questions and exploring the aforementioned connections may be a fruitful path forward. We hope that this monograph will be valuable in pursuing these goals.

Acknowledgements

The authors thank Yunwen Lei, Alex Olshevsky, Olivier Wintenberger, the anonymous referees, and the associate editor for providing valuable comments on an earlier version of this work.

F.H. and G.D. acknowledge support by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

G.D. also acknowledges support by the Swedish Foundation for Strategic Research under grant number FUS21-0004.

B.G. acknowledges partial support by the U.S. Army Research Laboratory and the U.S. Army Research Office, and by the U.K. Ministry of Defence and the U.K. Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/R013616/1. B.G. acknowledges partial support from the French National Agency for Research, through grants ANR-18-CE40-0016-01 and ANR-18-CE23-0015-02, and through the programme "France 2030" and PEPR IA on grant SHARP ANR-23-PEIA-0008.

M.R. acknowledges partial support by the U.S. National Science Foundation (NSF) through the Illinois Institute for Data Science and Dynamical Systems (iDS^2), an NSF HDR TRIPODS Institute, under Award CCF-193498.

References

Achille, A., Paolini, G., Mbeng, G., and Soatto, S. (2021). “The information complexity of learning tasks, their structure and their distance”. *Information and Inference: A Journal of the IMA*. 10(1): 51–72. DOI: 10.1093/imaiai/iaaa033 (cit. on p. 183).

Achille, A. and Soatto, S. (2018). “Emergence of Invariance and Disentanglement in Deep Representations”. *Journal of Machine Learning Research (JMLR)*. 19(Sept.): 1–34. DOI: 10.1109/ITA.2018.8503149 (cit. on p. 29).

Akaike, H. (1974). “A new look at the statistical model identification”. *IEEE Trans. Autom. Control*. 19(6): 716–723. DOI: 10.1109/TAC.1974.1100705 (cit. on p. 5).

Alabdulmohsin, I. (2020). “Towards a Unified Theory of Learning and Information”. *Entropy*. 22(4). DOI: 10.3390/e22040438 (cit. on p. 71).

Alquier, P. (2006). “Transductive and inductive adaptative inference for regression and density estimation”. *PhD thesis*. University of Paris (cit. on p. 92).

Alquier, P. (2008). “PAC-Bayesian bounds for randomized empirical risk minimizers”. *Mathematical Methods of Statistics*. 17(4): 279–304. DOI: 10.3103/S1066530708040017 (cit. on p. 92).

Alquier, P. (2024). "User-friendly Introduction to PAC-Bayes Bounds". *Foundations and Trends® in Machine Learning*. 17(2): 174–303. ISSN: 1935-8237. DOI: 10.1561/2200000100 (cit. on pp. 5, 42, 76, 81, 89, 91, 92).

Alquier, P. and Biau, G. (2013). "Sparse single-index model". *Journal of Machine Learning Research (JMLR)*. 14(1): 243–280 (cit. on pp. 9, 78, 91).

Alquier, P. and Guedj, B. (2017). "An oracle inequality for quasi-Bayesian nonnegative matrix factorization". *Mathematical Methods of Statistics*. 26(1): 55–67. DOI: 10.3103/S1066530717010045 (cit. on p. 91).

Alquier, P. and Guedj, B. (2018). "Simpler PAC-Bayesian bounds for hostile data". *Machine Learning*. 107(5): 887–902. DOI: 10.1007/s10994-017-5690-0 (cit. on pp. 69, 76, 81, 82, 90).

Alquier, P. and Lounici, K. (2011). "PAC-Bayesian bounds for sparse regression estimation with exponential weights". *Electronic Journal of Statistics*. 5(Mar.): 127–145. DOI: 10.1214/11-EJS601 (cit. on pp. 91, 92, 184).

Alquier, P., Ridgway, J., and Chopin, N. (2016). "On the properties of variational approximations of Gibbs posteriors". *Journal of Machine Learning Research (JMLR)*. 17(236): 1–41 (cit. on p. 128).

Ambroladze, A., Parrado-Hernandez, E., and Shawe-Taylor, J. (2006). "Tighter PAC-Bayes Bounds." In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Vancouver, Canada. DOI: 10.7551/mitpress/7503.003.0007 (cit. on pp. 83, 90, 155, 160).

Aminian, G., Abroshan, M., Khalili, M. M., Toni, L., and Rodrigues, M. R. D. (2022a). "An Information-theoretical Approach to Semi-supervised Learning under Covariate-shift". In: *Proc. Artif. Intell. Statist. (AISTATS)*. Virtual conference (cit. on pp. 70, 71, 175).

Aminian, G., Bu, Y., Toni, L., Rodrigues, M. R. D., and Wornell, G. (2021a). "An Exact Characterization of the Generalization Error for the Gibbs Algorithm". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Virtual Conference (cit. on pp. 126, 127).

Aminian, G., Bu, Y., Wornell, G. W., and Rodrigues, M. R. D. (2022b). "Tighter Expected Generalization Error Bounds via Convexity of Information Measures". In: *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*. Espoo, Finland. DOI: 10.1109/ISIT50566.2022.9834474 (cit. on pp. 61, 67, 129, 183).

Aminian, G., Toni, L., and Rodrigues, M. R. D. (2020). "Jensen-Shannon Information Based Characterization of the Generalization Error of Learning Algorithms". In: *Proc. IEEE Inf. Theory Workshop (ITW)*. Riva del Garda, Italy. DOI: 10.1109/ITW46852.2021.9457642 (cit. on p. 71).

Aminian, G., Toni, L., and Rodrigues, M. R. D. (2021b). "Information-Theoretic Bounds on the Moments of the Generalization Error of Learning Algorithms". In: *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*. Melbourne, Australia. DOI: 10.1109/ISIT45174.2021.9518043 (cit. on pp. 71, 143).

Amit, R. and Meir, R. (2018). "Meta-Learning by Adjusting Priors Based on Extended PAC-Bayes Theory". In: *Proc. Int. Conf. Mach. Learning (ICML)*. Stockholm, Sweden (cit. on pp. 166, 183).

Amit, R., Epstein, B., Moran, S., and Meir, R. (2022). "Integral Probability Metrics PAC-Bayes Bounds". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. New Orleans, LA, USA (cit. on p. 93).

Arora, S., Ge, R., Neyshabur, B., and Zhang, Y. (2018). "Stronger generalization bounds for deep nets via a compression approach". In: *Proc. Int. Conf. Mach. Learning (ICML)*. Stockholm, Sweden (cit. on p. 159).

Asadi, A. R., Abbe, E., and Verdú, S. (2018). "Chaining Mutual Information and Tightening Generalization Bounds". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Montreal, Canada (cit. on pp. 63, 64, 70).

Asadi, A. R. and Abbe, E. (2020). "Chaining Meets Chain Rule: Multilevel Entropic Regularization and Training of Neural Networks". *Journal of Machine Learning Research (JMLR)*. 21(1) (cit. on p. 92).

Audibert, J.-Y. (2004). “A better variance control for PAC-Bayesian classification”. URL: certis.enpc.fr/~audibert/Mes%20articles/PhDthesis.pdf (cit. on pp. 27, 94, 102, 118, 144).

Audibert, J.-Y. and Bousquet, O. (2007). “Combining PAC-Bayesian and Generic Chaining Bounds”. *Journal of Machine Learning Research (JMLR)*. 8(32): 863–889 (cit. on pp. 70, 92).

Banerjee, A. (2006). “On Bayesian Bounds”. In: *Proc. Int. Conf. Mach. Learning (ICML)*. Pittsburgh, PE, USA. DOI: 10.1145/1143844.1143855 (cit. on p. 42).

Banerjee, A., Chen, T., Li, X., and Zhou, Y. (2022). “Stability Based Generalization Bounds for Exponential Family Langevin Dynamics”. In: *Proc. Int. Conf. Mach. Learning (ICML)*. Baltimore, MD, USA (cit. on pp. 145, 161).

Banerjee, P. K. and Montufar, G. (2021). “Information Complexity and Generalization Bounds”. In: *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*. Melbourne, Australia. DOI: 10.1109/ISIT45174.2021.9517960 (cit. on pp. 5, 9, 88).

Barnes, L. P., Dytso, A., and Poor, H. V. (2022). “Improved Information Theoretic Generalization Bounds for Distributed and Federated Learning”. In: *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*. Espoo, Finland. DOI: 10.1109/ISIT50566.2022.9834700 (cit. on pp. 175–177, 183).

Barron, A., Rissanen, J., and Yu, B. (1998). “The minimum description length principle in coding and modeling”. *IEEE Trans. Info. Theory*. 44(6): 2743–2760. DOI: 10.1109/18.720554 (cit. on p. 5).

Barron, A. and Cover, T. (1991). “Minimum complexity density estimation”. *IEEE Trans. Info. Theory*. 37(4): 1034–1054. DOI: 10.1109/18.86996 (cit. on p. 5).

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. (2017). “Spectrally-normalized margin bounds for neural networks”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Long Beach, CA, USA (cit. on p. 160).

Bartlett, P. L. and Mendelson, S. (2001). “Rademacher and Gaussian Complexities: Risk Bounds and Structural Results”. In: *Proc. Euro. Conf. Comput. Learn. Theory (EuroCOLT)*. Amsterdam, The Netherlands (cit. on pp. 11, 14).

Bartlett, P. L. and Mendelson, S. (2002). “Rademacher and Gaussian Complexities: Risk Bounds and Structural Results”. *Journal of Machine Learning Research (JMLR)*. 3(Nov.): 463–482. ISSN: 1532-4435. DOI: 10.1007/3-540-44581-1_15 (cit. on pp. 11, 14, 160).

Bassily, R., Moran, S., Nachum, I., Shafer, J., and Yehudayoff, A. (2018). “Learners That Use Little Information”. *Journal of Machine Learning Research (JMLR)*. 83(Apr.): 25–55 (cit. on pp. 133, 144).

Bassily, R., Nissim, K., Smith, A., Steinke, T., Stemmer, U., and Ullman, J. (2016). “Algorithmic Stability for Adaptive Data Analysis”. In: vol. 50. No. 3. DOI: 10.1145/2897518.2897566 (cit. on p. 91).

Baxter, J. (2000). “A Model of Inductive Bias Learning”. *J. Artif. Int. Res.* 12(1): 149–198. DOI: 10.1613/jair.731 (cit. on p. 164).

Bégin, L., Germain, P., Laviolette, F., and Roy, J.-F. (2014). “PAC-Bayesian Theory for Transductive Learning”. In: *Proc. Artif. Intell. Statist. (AISTATS)*. Reykjavik, Iceland (cit. on pp. 76, 89).

Bégin, L., Germain, P., Laviolette, F., and Roy, J.-F. (2016). “PAC-Bayesian Bounds based on the Rényi Divergence”. In: *Proc. Artif. Intell. Statist. (AISTATS)*. Cadiz, Spain (cit. on pp. 69, 81, 82, 90).

Bernstein, J. and Yue, Y. (2021). “Computing the Information Content of Trained Neural Networks”. In: *Workshop on the Theory of Overparameterized Machine Learning* (cit. on p. 161).

Biggs, F. and Guedj, B. (2021). “Differentiable PAC–Bayes Objectives with Partially Aggregated Neural Networks”. *Entropy*. 23(10). DOI: 10.3390/e23101280 (cit. on p. 161).

Biggs, F. and Guedj, B. (2022a). “Non-Vacuous Generalisation Bounds for Shallow Neural Networks”. In: *Proc. Int. Conf. Mach. Learn. (ICML)*. Baltimore, MD (cit. on p. 161).

Biggs, F. and Guedj, B. (2022b). “On Margins and Derandomisation in PAC-Bayes”. In: *Proc. Artif. Intell. Statist. (AISTATS)*. Virtual Conference (cit. on p. 92).

Biggs, F. and Guedj, B. (2023). “Tighter PAC-Bayes Generalisation Bounds by Leveraging Example Difficulty”. In: *Proc. Artif. Intell. Statist. (AISTATS)*. Valencia, Spain (cit. on p. 92).

Biggs, F., Zantedeschi, V., and Guedj, B. (2022). “On Margins and Generalisation for Voting Classifiers”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. New Orleans, LA, USA (cit. on p. 91).

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. (1989). "Learnability and the Vapnik-Chervonenkis Dimension". *J. ACM.* 36(4): 929–965. DOI: 10.1145/76359.76371 (cit. on p. 4).

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. (1987). "Occam's Razor". *Information Processing Letters.* 24(6): 377–380. DOI: https://doi.org/10.1016/0020-0190(87)90114-1 (cit. on p. 4).

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities. A nonasymptotic theory of independence.* Oxford, United Kingdom: Oxford University Press (cit. on p. 46).

Bousquet, O. and Elisseeff, A. (2002). "Stability and Generalization". *Journal of Machine Learning Research (JMLR).* 2(Mar.): 499–526 (cit. on pp. 16, 140).

Bretagnolle, J. and Huber, C. (1978). "Estimation des densités : risque minimax". fre. *Séminaire de probabilités de Strasbourg.* 12: 342–363 (cit. on p. 36).

Bu, Y., Zou, S., and Veeravalli, V. V. (2020). "Tightening Mutual Information-Based Bounds on Generalization Error". *IEEE J. Sel. Areas Inf. Theory.* 1(1): 121–130. DOI: 10.1109/ISIT.2019.8849590 (cit. on pp. 58, 59, 70, 94, 128–131, 160, 161).

Bu, Y., Aminian, G., Toni, L., Wornell, G. W., and Rodrigues, M. R. D. (2022). "Characterizing and Understanding the Generalization Error of Transfer Learning with Gibbs Algorithm". In: *Proc. Artif. Intell. Statist. (AISTATS).* Virtual conference (cit. on p. 183).

Bu, Y., Gao, W., Zou, S., and Veeravalli, V. V. (2021). "Population Risk Improvement with Model Compression: An Information-Theoretic Approach". *Entropy.* 23(10). DOI: 10.3390/e23101255 (cit. on p. 159).

Bu, Y., Zou, S., and Veeravalli, V. V. (2019). "Tightening Mutual Information Based Bounds on Generalization Error". In: *Proc. IEEE Int. Symp. Inf. Theory (ISIT).* Paris, France. DOI: 10.1109/ISIT.2019.8849590 (cit. on pp. 70, 144).

Canonne, C. L. (2022). "A short note on an inequality between KL and TV". *arXiv.* DOI: 10.48550/arxiv.2202.07198 (cit. on p. 35).

Caruana, R. (1997). "Multitask Learning". *Mach. Learn.* 28(1): 41–75. DOI: 10.1007/978-1-4615-5529-2_5 (cit. on p. 164).

Catoni, O. (2007). *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning.* Vol. 56. IMS Lecture Notes Monogr. Ser. 1–163 (cit. on pp. 4, 5, 9, 27, 75, 77, 79, 85, 89–92, 94, 102, 118, 124, 144).

Catoni, O. (2004a). “A PAC-Bayesian approach to adaptive classification”. URL: yaroslavvb.com/papers/notes/catoni-pac.pdf (cit. on pp. 144, 145).

Catoni, O. (2004b). *Statistical Learning Theory and Stochastic Optimization.* Ed. by J. Picard. *Lecture Notes in Mathematics: Saint-Flour Summer School on Probability Theory XXXI 2001.* DOI: 10.1007/b99352 (cit. on p. 92).

Catoni, O. and Giulini, I. (2018). “Dimension-free PAC-Bayesian bounds for the estimation of the mean of a random vector”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS). (Almost) 50 Shades of Bayesian Learning: PAC-Bayesian trends and insights* (cit. on p. 92).

Chen, Q., Shui, C., and Marchand, M. (2021). “Generalization Bounds For Meta-Learning: An Information-Theoretic Analysis”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS).* Virtual Conference (cit. on pp. 165, 166, 183).

Chérief-Abdellatif, B.-E., Shi, Y., Doucet, A., and Guedj, B. (2022). “On PAC-Bayesian reconstruction guarantees for VAEs”. In: *Proc. Artif. Intell. Statist. (AISTATS).* Virtual conference (cit. on p. 93).

Chu, Y. and Raginsky, M. (2023). “A unified framework for information-theoretic generalization bounds”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS).* New Orleans, LO, USA (cit. on p. 72).

Clerico, E., Deligiannidis, G., and Doucet, A. (2022a). “Conditionally Gaussian PAC-Bayes”. In: *Proc. Artif. Intell. Statist. (AISTATS).* Virtual conference (cit. on p. 162).

Clerico, E., Deligiannidis, G., and Doucet, A. (2023). “Wide stochastic networks: Gaussian limit and PAC-Bayesian training”. In: *Proc. Conf. Alg. Learn. Theory (ALT).* Singapore (cit. on p. 161).

Clerico, E. and Guedj, B. (2024). “A note on regularised NTK dynamics with an application to PAC-Bayesian training”. *Transactions on Machine Learning Research (TMLR).* Apr. ISSN: 2835-8856 (cit. on p. 161).

Clerico, E., Shidani, A., Deligiannidis, G., and Doucet, A. (2022b). “Chained generalisation bounds”. In: *Proc. Conf. Learn. Theory (COLT)*. Boulder, CO, USA (cit. on pp. 71, 92).

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. USA: Wiley-Interscience (cit. on pp. 31, 64, 113).

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press. DOI: 10.1017/CBO9780511801389 (cit. on p. 154).

Csiszar, I. (1975). “I-Divergence Geometry of Probability Distributions and Minimization Problems”. *The Annals of Probability*. 3(1): 146–158. DOI: 10.1214/aop/1176996454 (cit. on p. 22).

Csiszar, I. and Körner, J. (2011). *Information Theory: Coding Theorems for Discrete Memoryless Systems*. 2nd. Cambridge, U.K.: Cambridge Univ. Press. DOI: 10.1017/CBO9780511921889 (cit. on pp. 54, 124).

Dalalyan, A. S. and Salmon, J. (2012). “Sharp oracle inequalities for aggregation of affine estimators”. *The Annals of Statistics*. 40(4): 2327–2355. DOI: 10.1214/12-AOS1038 (cit. on pp. 9, 91, 92).

Dalalyan, A. S. and Tsybakov, A. B. (2007). “Aggregation by exponential weighting and sharp oracle inequalities”. In: *Proc. Conf. Learn. Theory (COLT)*. DOI: 10.1007/978-3-540-72927-3_9 (cit. on p. 91).

Dalalyan, A. S. and Tsybakov, A. B. (2008). “Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity”. *Machine Learning*. 72(Aug.): 39–61. DOI: 10.1007/s10994-008-5051-0 (cit. on pp. 91, 92).

Dalalyan, A. S. and Tsybakov, A. B. (2012). “Sparse regression learning by aggregation and Langevin Monte-Carlo”. *J. Comput. System Sci.* 78: 1423–1443. DOI: 10.1016/j.jcss.2011.12.023 (cit. on pp. 91, 92).

Devroye, L. and Wagner, T. (1979). “Distribution-free performance bounds for potential function rules”. *IEEE Trans. Inf. Theory*. 25(5): 601–604. DOI: 10.1109/TIT.1979.1056087 (cit. on p. 15).

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. (2017). “Sharp Minima Can Generalize For Deep Nets”. In: *Proc. Int. Conf. Mach. Learning (ICML)*. Sydney, Australia (cit. on p. 155).

Dogan, M. B. and Gastpar, M. (2021). "Lower Bounds on the Expected Excess Risk Using Mutual Information". In: *Proc. IEEE Inf. Theory Workshop (ITW)*. Kanazawa, Japan. DOI: 10.1109/ITW48936.2021.9611483 (cit. on p. 72).

Donsker, M. D. and Varadhan, S. R. S. (1975). "Asymptotic evaluation of certain Markov process expectations for large time, I". *Comm. Pure Appl. Math.* 28(1): 1–47. DOI: 10.1002/cpa.3160280102 (cit. on pp. 22, 42).

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. (2015). "Generalization in Adaptive Data Analysis and Holdout Reuse". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Montreal, Canada (cit. on pp. 142, 143, 145).

Dziugaite, G. K., Hsu, K., Gharbieh, W., and Roy, D. M. (2021). "On the role of data in PAC-Bayes bounds". In: *Proc. Artif. Intell. Statist. (AISTATS)*. San Diego, CA, USA (cit. on pp. 83, 90, 155, 156, 160).

Dziugaite, G. K. and Roy, D. M. (2017). "Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data". In: *Proc. Conf. Uncertainty in Artif. Intell. (UAI)*. Sydney, Australia (cit. on pp. 154, 156, 159, 186).

Dziugaite, G. K. and Roy, D. M. (2018a). "Entropy-SGD optimizes the prior of a PAC-Bayes bound: Generalization properties of Entropy-SGD and data-dependent priors". In: *Proc. Int. Conf. Mach. Learning (ICML)*. Stockholm, Sweden (cit. on p. 161).

Dziugaite, G. K., Drouin, A., Neal, B., Rajkumar, N., Caballero, E., Wang, L., Mitliagkas, I., and Roy, D. M. (2020). "In Search of Robust Measures of Generalization". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Vancouver, Canada (cit. on p. 162).

Dziugaite, G. K. and Roy, D. M. (2018b). "Data-dependent PAC-Bayes priors via differential privacy". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Montreal, Canada (cit. on pp. 84, 90).

Edgeworth, F. Y. (1908). "On the Probable Errors of Frequency-Constants". *Journal of the Royal Statistical Society*. 71(2): 381–397 (cit. on p. 5).

Esposito, A. R. and Gastpar, M. (2022). “From Generalisation Error to Transportation-cost Inequalities and Back”. In: *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*. Espoo, Finland. DOI: 10.1109/ISIT50566.2022.9834354 (cit. on p. 71).

Esposito, A. R., Gastpar, M., and Issa, I. (2021a). “Generalization Error Bounds via Rényi-, f-Divergences and Maximal Leakage”. *IEEE Trans. Inf. Theory*. 67(8): 4986–5004. DOI: 10.1109/TIT.2021.3085190 (cit. on pp. 86, 91, 119, 143, 145).

Esposito, A. R. and Mondelli, M. (2023). “Concentration without Independence via Information Measures”. In: *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*. Taipei, Taiwan. DOI: 10.1109/ISIT54713.2023.10206899 (cit. on p. 46).

Esposito, A. R., Wu, D., and Gastpar, M. (2021b). “On conditional Sibson’s α-Mutual Information”. In: *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*. Melbourne, Australia. DOI: 10.1109/ISIT45174.2021.9517944 (cit. on p. 38).

Fard, M. M. and Pineau, J. (2010). “PAC-Bayesian Model Selection for Reinforcement Learning”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Vancouver, Canada (cit. on p. 183).

Farid, A. and Majumdar, A. (2021). “Generalization Bounds for Meta-Learning via PAC-Bayes and Uniform Stability”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Virtual Conference (cit. on p. 183).

Feldman, V. and Steinke, T. (2018). “Calibrating Noise to Variance in Adaptive Data Analysis”. In: *Proc. Conf. Learning Theory (COLT)*. Stockholm, Sweden (cit. on pp. 143, 145).

Fenchel, W. (1949). “On Conjugate Convex Functions”. *Canadian Journal of Mathematics*. 1(1): 73–77. DOI: 10.1007/978-3-0348-0439-4_7 (cit. on p. 43).

Fisher, R. A. and Russell, E. J. (1922). “On the mathematical foundations of theoretical statistics”. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*. 222(594-604): 309–368. DOI: 10.1098/rsta.1922.0009 (cit. on p. 5).

Flynn, H., Reeb, D., Kandemir, M., and Peters, J. (2022). "PAC-Bayesian Lifelong Learning for Multi-Armed Bandits". *Data Min. Knowl. Discov.* 36(2): 841–876 (cit. on p. 183).

Foong, A. Y. K., Bruinsma, W. P., Burt, D. R., and Turner, R. E. (2021). "How Tight Can PAC-Bayes be in the Small Data Regime?" In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS).* Virtual Conference (cit. on pp. 90, 186).

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. (2021). "Sharpness-aware Minimization for Efficiently Improving Generalization". In: *Proc. Int. Conf. Learn. Representations (ICLR).* Vienna, Austria (cit. on p. 161).

Futami, F. and Fujisawa, M. (2023). "Time-Independent Information-Theoretic Generalization Bounds for SGLD". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS).* New Orleans, LO, USA (cit. on p. 160).

Geiger, B. C. (2021). "On Information Plane Analyses of Neural Network Classifiers——A Review". *IEEE Trans. Neural Networks Learning Systems.* 33(June): 7039–7051. DOI: 10.1109/TNNLS.2021.3089037 (cit. on p. 29).

Geiping, J., Goldblum, M., Pope, P. E., Moeller, M., and Goldstein, T. (2022). "Stochastic Training is Not Necessary for Generalization". In: *Proc. Int. Conf. Learn. Representations (ICLR).* Virtual Conference (cit. on p. 153).

Germain, P., Bach, F., Lacoste, A., and Lacoste-Julien, S. (2016a). "PAC-Bayesian Theory Meets Bayesian Inference". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS).* Barcelona, Spain (cit. on p. 92).

Germain, P., Habrard, A., Laviolette, F., and Morvant, E. (2016b). "A New PAC-Bayesian Perspective on Domain Adaptation". In: *Proc. Int. Conf. Mach. Learning (ICML).* New York, NY, USA (cit. on p. 183).

Germain, P., Lacasse, A., Laviolette, F., March, M., and Roy, J.-F. (2015). "Risk Bounds for the Majority Vote: From a PAC-Bayesian Analysis to a Learning Algorithm". *Journal of Machine Learning Research (JMLR).* 16(26): 787–860 (cit. on p. 91).

Germain, P., Lacasse, A., Laviolette, F., and Marchand, M. (2009a). "PAC-Bayesian Learning of Linear Classifiers". In: *Proc. Int. Conf. Mach. Learning (ICML)*. Montreal, Canada. DOI: 10.1145/1553374.1553419 (cit. on pp. 69, 76, 89).

Germain, P., Lacasse, A., Marchand, M., Shanian, S., and Laviolette, F. (2009b). "From PAC-Bayes Bounds to KL Regularization". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Vancouver, Canada (cit. on p. 92).

Gine, E. and Zinn, J. (1984). "Some Limit Theorems for Empirical Processes". *The Annals of Probability*. 12(4): 929–989. DOI: 10.1214/aop/1176993138 (cit. on pp. 11, 14).

Goldfeld, Z. and Polyanskiy, Y. (2020). "The Information Bottleneck Problem and its Applications in Machine Learning". *IEEE J. Sel. Areas Inf. Theory*. 1(1): 19–38. DOI: 10.1109/JSAIT.2020.2991561 (cit. on p. 29).

Gouverneur, A., Rodríguez-Gálvez, B., Oechtering, T. J., and Skoglund, M. (2022). "An Information-Theoretic Analysis of Bayesian Reinforcement Learning". In: *Allerton Conf. Communication, Control, Computing (Allerton)*. Monticello, IL, USA. DOI: 10.1109/Allerton49937.2022.9929353 (cit. on pp. 180, 182, 183).

Goyal, A., Morvant, E., Germain, P., and Amini, M. (2017). "PAC-Bayesian Analysis for a Two-Step Hierarchical Multiview Learning Approach". In: *Proc. Mach. Learn. Knowl. Discovery in Databases - Eur. Conf., ECML PKDD, Part II*. Vol. 10535. *Lecture Notes in Computer Science*. Skopje, Macedonia: Springer. 205–221. DOI: 10.1007/978-3-319-71246-8_13 (cit. on p. 54).

Grünwald, P. and Mehta, N. A. (2020). "Fast Rates for General Unbounded Loss Functions: From ERM to Generalized Bayes". *Journal of Machine Learning Research (JMLR)*. 21(Mar.): 1–80 (cit. on pp. 14, 53, 89).

Grünwald, P., Steinke, T., and Zakynthinou, L. (2021). "PAC-Bayes, MAC-Bayes and Conditional Mutual Information: Fast rate bounds that handle general VC classes". In: *Proc. Conf. Learn. Theory (COLT)*. Boulder, CO, USA (cit. on pp. 5, 9, 102, 105, 106, 119, 144).

Grünwald, P. (2007). *The minimum description length principle.* MIT press (cit. on p. 5).

Grünwald, P. and Mehta, N. A. (2019). “A tight excess risk bound via a unified PAC-Bayesian–Rademacher–Shtarkov–MDL complexity”. In: *Proc. Conf. Alg. Learn. Theory (ALT).* Chicago, IL, USA (cit. on p. 144).

Grünwald, P., Pérez-Ortiz, M. F., and Mhammedi, Z. (2023). “Exponential Stochastic Inequality”. *arXiv.* May. DOI: 10.48550/arxiv.2304.14217 (cit. on pp. 53, 75, 89).

Guedj, B. (2019). “A primer on PAC-Bayesian learning”. *Proc. 2nd Congress Société Mathématique de France*: 391–414. DOI: 10.48550/arxiv.1901.05353 (cit. on pp. 76, 91).

Guedj, B. and Alquier, P. (2013). “PAC-Bayesian estimation and prediction in sparse additive models”. *Electronic Journal of Statistics.* 7(Jan.): 264–291. DOI: 10.1214/13-EJS771 (cit. on pp. 9, 78, 91).

Guedj, B. and Robbiano, S. (2018). “PAC-Bayesian high dimensional bipartite ranking”. *Journal of Statistical Planning and Inference.* 196(Aug.): 70–86. DOI: 10.1016/j.jspi.2017.10.010 (cit. on p. 184).

Haddouche, M. and Guedj, B. (2022). “Online PAC-Bayes Learning”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS).* New Orleans, LA, USA (cit. on p. 184).

Haddouche, M. and Guedj, B. (2023a). “PAC-Bayes Generalisation Bounds for Heavy-Tailed Losses through Supermartingales”. *Transactions on Machine Learning Research (TMLR).* Apr. ISSN: 2835-8856 (cit. on p. 93).

Haddouche, M. and Guedj, B. (2023b). “Wasserstein PAC-Bayes Learning: Exploiting Optimisation Guarantees to Explain Generalisation”. *arXiv.* DOI: 10.48550/arXiv.2304.07048 (cit. on p. 93).

Haddouche, M., Guedj, B., Rivasplata, O., and Shawe-Taylor, J. (2021). “PAC-Bayes Unleashed: Generalisation Bounds with Unbounded Losses”. *Entropy.* 23(10). DOI: 10.3390/e23101330 (cit. on p. 93).

Hafez-Kolahi, H., Golgooni, Z., Kasaei, S., and Soleymani, M. (2020). “Conditioning and Processing: Techniques to Improve Information-Theoretic Generalization Bounds”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS).* Vol. 33. Vancouver, Canada (cit. on p. 71).

Hafez-Kolahi, H., Moniri, B., and Kasaei, S. (2023). “Information-Theoretic Analysis of Minimax Excess Risk”. *IEEE Trans. Inf. Theory*. 69(7): 4659–4674. DOI: 10.1109/TIT.2023.3249636 (cit. on p. 72).

Hafez-Kolahi, H., Moniri, B., Kasaei, S., and Baghshah, M. S. (2021). “Rate-Distortion Analysis of Minimum Excess Risk in Bayesian Learning”. In: *Proc. Int. Conf. Mach. Learning (ICML)*. Virtual conference (cit. on p. 72).

Haghifam, M., Negrea, J., Khisti, A., Roy, D. M., and Dziugaite, G. K. (2020). “Sharpened Generalization Bounds based on Conditional Mutual Information and an Application to Noisy, Iterative Algorithms”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Vancouver, Canada (cit. on pp. 70, 98, 101, 119, 160).

Haghifam, M., Dziugaite, G. K., Moran, S., and Roy, D. M. (2021). “Towards a Unified Information-Theoretic Framework for Generalization”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Virtual Conference (cit. on pp. 136, 138, 139, 144, 145, 161, 185).

Haghifam, M., Moran, S., Roy, D. M., and Dziugiate, G. K. (2022). “Understanding Generalization via Leave-One-Out Conditional Mutual Information”. In: *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*. Espoo, Finland. DOI: 10.1109/ISIT50566.2022.9834400 (cit. on pp. 110, 114–116, 119, 137, 138, 144, 185).

Haghifam, M., Rodríguez-Gálvez, B., Thobaben, R., Skoglund, M., Roy, D. M., and Dziugaite, G. K. (2023). “Limitations of Information-Theoretic Generalization Bounds for Gradient Descent Methods in Stochastic Convex Optimization”. In: *Proc. Conf. Alg. Learn. Theory (ALT)*. Singapore (cit. on pp. 160, 185).

Han, S., Mao, H., and Dally, W. J. (2016). “Deep Compression: Compressing Deep Neural Network with Pruning, Trained Quantization and Huffman Coding”. In: *Proc. Int. Conf. Learn. Representations (ICLR)*. San Juan, Puerto Rico (cit. on p. 159).

Harutyunyan, H., Raginsky, M., Steeg, G. V., and Galstyan, A. (2021). “Information-theoretic generalization bounds for black-box learning algorithms”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Virtual Conference (cit. on pp. 60, 70, 105, 112, 119, 139, 140, 142, 144, 145, 156–158, 186).

Harutyunyan, H., Steeg, G. V., and Galstyan, A. (2022). “Formal limitations of sample-wise information-theoretic generalization bounds”. In: *Proc. IEEE Inf. Theory Workshop (ITW)*. Mumbai, India. DOI: 10.1109/ITW54588.2022.9965850 (cit. on pp. 61, 70, 80).

Haussler, D., Littlestone, N., and Warmuth, M. (1988). “Predicting (0, 1)-functions on randomly drawn points”. In: *Annual Symposium on Foundations of Computer Science*. White Plains, NY, USA. DOI: 10.1006/inco.1994.1097 (cit. on pp. 137, 138).

He, H., Yan, H., and Tan, V. Y. F. (2022). “Information-Theoretic Characterization of the Generalization Error for Iterative Semi-Supervised Learning”. *Journal of Machine Learning Research (JMLR)*. 23(287): 1–52 (cit. on p. 183).

Hellström, F. and Durisi, G. (2020a). “Generalization Bounds via Information Density and Conditional Information Density”. *IEEE J. Sel. Areas Inf. Theory*. 1(3): 824–839. DOI: 10.1109/JSAIT.2020.3040992 (cit. on pp. 90, 91, 119, 145).

Hellström, F. and Durisi, G. (2020b). “Generalization Error Bounds via mth Central Moments of the Information Density”. In: *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*. Los Angeles, CA, USA. DOI: 10.1109/ISIT44484.2020.9174475 (cit. on pp. 95, 103, 109).

Hellström, F. and Durisi, G. (2021a). “Data-dependent PAC-Bayesian bounds in the random-subset setting with applications to neural networks”. In: *Proc. Int. Conf. Mach. Learn. (ICML). Workshop on Inf.-Theoretic Methods Rigorous, Responsible, and Reliable Mach. Learn. (ITR3)*. Virtual conference (cit. on pp. 70, 91, 156).

Hellström, F. and Durisi, G. (2021b). “Fast-Rate Loss Bounds via Conditional Information Measures with Applications to Neural Networks”. In: *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*. Melbourne, Australia. DOI: 10.1109/ISIT45174.2021.9517731 (cit. on pp. 91, 156).

Hellström, F. and Durisi, G. (2022a). “A New Family of Generalization Bounds Using Samplewise Evaluated CMI”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. New Orleans, LA, USA (cit. on pp. 70, 118, 119, 144, 157, 158).

Hellström, F. and Durisi, G. (2022b). "Evaluated CMI Bounds for Meta Learning: Tightness and Expressiveness". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. New Orleans, LA, USA (cit. on pp. 167, 183).

Hellström, F. and Guedj, B. (2024). "Comparing Comparators in Generalization Bounds". In: *Proc. Artif. Intell. Statist. (AISTATS)*. Valencia, Spain (cit. on pp. 90, 186).

Herbrich, R. and Graepel, T. (2002). "A PAC-Bayesian margin bound for linear classifiers". *IEEE Trans. Inf. Theory*. 48(12): 3140–3150. DOI: 10.1109/TIT.2002.805090 (cit. on p. 92).

Higgs, M. and Shawe-Taylor, J. (2010). "A PAC-Bayes Bound for Tailored Density Estimation". In: *Proc. Conf. Alg. Learn. Theory (ALT)*. Canberra, Australia. DOI: 10.1007/978-3-642-16108-7_15 (cit. on p. 184).

Holland, M. (2019). "PAC-Bayes under potentially heavy tails". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Vancouver, Canada (cit. on p. 92).

Huang, W., Liu, C., Chen, Y., Xu, R. Y. D., Zhang, M., and Weng, T.-W. (2023). "Analyzing Deep PAC-Bayesian Learning with Neural Tangent Kernel: Convergence, Analytic Generalization Bound, and Efficient Hyperparameter Selection". *Transactions on Machine Learning Research (TMLR)*. May. ISSN: 2835-8856 (cit. on p. 161).

Issa, I., Kamath, S., and Wagner, A. B. (2020). "An operational approach to information leakage". *IEEE Trans. Inf. Theory*. 66(3): 1625–1657. DOI: 10.1109/TIT.2019.2962804 (cit. on pp. 37, 108).

Issa, I., Esposito, A. R., and Gastpar, M. (2023). "Generalization Error Bounds for Noisy, Iterative Algorithms via Maximal Leakage". In: *Proc. Conf. Learn. Theory (COLT)*. Bangalore, India (cit. on p. 160).

Jacot, A., Gabriel, F., and Hongler, C. (2018). "Neural Tangent Kernel: Convergence and Generalization in Neural Networks". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Montreal, Canada (cit. on p. 161).

Jang, K., Jun, K.-S., Kuzborskij, I., and Orabona, F. (2023). "Tighter PAC-Bayes Bounds Through Coin-Betting". In: *Proc. Conf. Learn. Theory (COLT)*. Bangalore, India (cit. on p. 90).

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. (2020). “Fantastic Generalization Measures and Where to Find Them”. In: *Proc. Int. Conf. Learn. Representations (ICLR)*. Addis Ababa, Ethiopia (cit. on p. 162).

Jiao, J., Han, Y., and Weissman, T. (2017). “Dependence measures bounding the exploration bias for general measurements”. In: *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*. Aachen, Germany. DOI: 10.1109/ISIT.2017.8006774 (cit. on pp. 57, 58, 70).

Jin, G., Yi, X., Yang, P., Zhang, L., Schewe, S., and Huang, X. (2022). “Weight Expansion: A New Perspective on Dropout and Generalization”. *Transactions on Machine Learning Research (TMLR)*. Sept. ISSN: 2835-8856 (cit. on p. 162).

Jose, S. T., Park, S., and Simeone, O. (2022a). “Information-Theoretic Analysis of Epistemic Uncertainty in Bayesian Meta-learning”. In: *Proc. Artif. Intell. Statist. (AISTATS)*. Virtual conference (cit. on p. 183).

Jose, S. T. and Simeone, O. (2021a). “Information-Theoretic Generalization Bounds for Meta-Learning and Applications”. *Entropy*. 23(1). DOI: 10.3390/e23010126 (cit. on pp. 166, 183).

Jose, S. T., Simeone, O., and Durisi, G. (2022b). “Transfer Meta-Learning: Information- Theoretic Bounds and Information Meta-Risk Minimization”. *IEEE Trans. Inf. Theor.* 68(1): 474–501. DOI: 10.1109/TIT.2021.3119605 (cit. on pp. 166, 183).

Jose, S. T. and Simeone, O. (2021b). “A Unified PAC-Bayesian Framework for Machine Unlearning via Information Risk Minimization”. In: *Proc. IEEE Int. Workshop Mach. Learn. Sign. Processing (MLSP)*. Gold Coast, Australia. DOI: 10.1109/MLSP52302.2021.9596170 (cit. on p. 184).

Jose, S. T. and Simeone, O. (2021c). “An Information-Theoretic Analysis of the Impact of Task Similarity on Meta-Learning”. In: *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*. Melbourne, Australia. DOI: 10.1109/ISIT45174.2021.9517767 (cit. on pp. 167, 183).

Jose, S. T. and Simeone, O. (2021d). “Information-Theoretic Bounds on Transfer Generalization Gap Based on Jensen-Shannon Divergence”. In: *European Signal Processing Conference*. Dublin, Ireland. DOI: 10.23919/EUSIPCO54536.2021.9616270 (cit. on p. 171).

Jose, S. T. and Simeone, O. (2023). “Transfer Learning for Quantum Classifiers: An Information-Theoretic Generalization Analysis”. In: *Proc. IEEE Inf. Theory Workshop (ITW)*. Saint-Malo, France. DOI: 10.1109/ITW55543.2023.10160236 (cit. on p. 183).

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D’Oliveira, R. G. L., Eichner, H., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konecný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Qi, H., Ramage, D., Raskar, R., Raykova, M., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. (2021). “Advances and Open Problems in Federated Learning”. *Foundations and Trends in Machine Learning*. 14(1–2): 1–210. ISSN: 1935-8237. DOI: 10.1561/9781680837896 (cit. on p. 175).

Kawaguchi, K., Deng, Z., Ji, X., and Huang, J. (2023). “How Does Information Bottleneck Help Deep Learning?” In: *Proc. Int. Conf. Mach. Learning (ICML)*. Honolulu, HI, USA (cit. on p. 29).

Kingma, D. P. and Welling, M. (2019). “An Introduction to Variational Autoencoders”. *Foundations and Trends in Machine Learning*. 12(4): 307–392. DOI: 10.1561/9781680836233 (cit. on p. 93).

Kolmogorov, A. N. (1963). “On tables of random numbers”. *Sankhyā (Statistics). The Indian Journal of Statistics. Series A*. 25: 369–376 (cit. on p. 5).

Koltchinskii, V. (2001). “Rademacher penalties and structural risk minimization”. *IEEE Trans. Inf. Theory*. 47(5): 1902–1914. DOI: 10.1109/18.930926 (cit. on pp. 11, 14).

Koltchinskii, V. and Panchenko, D. (2000). “Rademacher Processes and Bounding the Risk of Function Learning”. In: *High Dimensional Probability II*. Boston, MA: Birkhäuser. 443–457. ISBN: 978-1-4612-1358-1 (cit. on pp. 11, 14).

Kontorovich, A. and Raginsky, M. (2017). "Concentration of Measure Without Independence: A Unified Approach Via the Martingale Method". In: *Convexity and Concentration.* New York, NY: Springer. 183–210. ISBN: 978-1-4939-7005-6 (cit. on p. 46).

Kontorovich, A. and Ramanan, K. (2008). "Concentration Inequalities for Dependent Random Variables via the Martingale Method". *The Annals of Probability.* 36(6): 2126–2158 (cit. on p. 46).

Koolen, W. M., Grünwald, P., and van Erven, T. (2016). "Combining Adversarial Guarantees and Stochastic Fast Rates in Online Learning". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS).* Barcelona, Spain (cit. on p. 89).

Kouw, W. M. and Loog, M. (2019). "An introduction to domain adaptation and transfer learning". *arXiv.* Dec. DOI: 10.48550/arxiv.1812.11806 (cit. on p. 167).

Kutin, S. and Niyogi, P. (2002). "Almost-Everywhere Algorithmic Stability and Generalization Error". In: *Proc. Conf. Uncertainty in Artif. Intell. (UAI).* Edmonton, Canada (cit. on p. 16).

Lacasse, A., Laviolette, F., Marchand, M., Germain, P., and Usunier, N. (2006). "PAC-Bayes Bounds for the Risk of the Majority Vote and the Variance of the Gibbs Classifier". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS).* Vancouver, Canada. DOI: 10.7551/mitpress/7503.003.0101 (cit. on p. 91).

Langford, J. (2002). "Quantitatively Tight Sample Complexity Bounds". *PhD thesis.* Carnegie Mellon University (cit. on p. 78).

Langford, J. and Caruana, R. (2001). "(Not) Bounding the True Error". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS).* Vancouver, Canada (cit. on pp. 154, 156).

Langford, J. and Seeger, M. (2001). "Bounds for Averaging Classifiers". *CMU Technical report.* CMU-CS-01-102 (cit. on pp. 27, 78, 90).

Langford, J. and Shawe-Taylor, J. (2002). "PAC-Bayes & margins". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS).* Vancouver, Canada (cit. on pp. 89, 91, 92).

Lecué, G. and Mendelson, S. (2017). "Regularization and the small-ball method II: complexity dependent error rates". *Journal of Machine Learning Research (JMLR).* 18(146): 1–48 (cit. on p. 52).

Lecué, G. and Mendelson, S. (2018). "Regularization and the small-ball method I: sparse recovery". *The Annals of Statistics*. 46(2): 611–641. DOI: 10.1214/17-AOS1562 (cit. on p. 52).

Lee, J., Bahri, Y., Novak, R., Schoenholz, S., Pennington, J., and Sohl-dickstein, J. (2018). "Deep Neural Networks as Gaussian Processes". In: *Proc. Int. Conf. Learn. Representations (ICLR)*. Vancouver, Canada (cit. on p. 161).

Lee, W. S., Bartlett, P., and Williamson, R. (1998). "The importance of convexity in learning with squared loss". *IEEE Trans. Inf. Theory*. 44(5): 1974–1980. DOI: 10.1109/18.705577 (cit. on p. 14).

Letarte, G., Germain, P., Guedj, B., and Laviolette, F. (2019). "Dichotomize and Generalize: PAC-Bayesian Binary Activated Deep Neural Networks". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Vancouver, Canada (cit. on p. 161).

Leung, G. and Barron, A. (2006). "Information Theory and Mixing Least-Squares Regressions". *IEEE Trans. Inf. Theory*. 52(8): 3396–3410. DOI: 10.1109/TIT.2006.878172 (cit. on pp. 5, 91).

Lever, G., Laviolette, F., and Shawe-Taylor, J. (2010). "Distribution-Dependent PAC-Bayes Priors". In: *Proc. Conf. Alg. Learn. Theory (ALT)*. Canberra, Australia. DOI: 10.1007/978-3-642-16108-7_13 (cit. on p. 90).

Lever, G., Laviolette, F., and Shawe-Taylor, J. (2013). "Tighter PAC-Bayes bounds through distribution-dependent priors". *Theoretical Computer Science*. 473: 4–28. DOI: 10.1016/j.tcs.2012.10.013 (cit. on p. 90).

Li, C., Jiang, W., and Tanner, M. (2013). "General Oracle Inequalities for Gibbs Posterior with Application to Ranking". In: *Proc. Conf. Learn. Theory (COLT)*. Princeton, NJ, USA (cit. on p. 184).

Li, J., Luo, X., and Qiao, M. (2020). "On Generalization Error Bounds of Noisy Gradient Methods for Non-Convex Learning". In: *Proc. Int. Conf. Learn. Representations (ICLR)*. Addis Ababa, Ethiopia (cit. on p. 160).

Li, L., Guedj, B., and Loustau, S. (2018). "A Quasi-Bayesian Perspective to Online Clustering". *Electronic Journal of Statistics*. 12(2). DOI: 10.1214/18-EJS1479 (cit. on p. 184).

Li, Q. J. (1999). “Estimation of Mixture Models”. *PhD thesis.* Yale University (cit. on p. 14).

Liao, R., Urtasun, R., and Zemel, R. (2021). “A PAC-Bayesian Approach to Generalization Bounds for Graph Neural Networks”. In: *Proc. Int. Conf. Learn. Representations (ICLR).* Vienna, Austria (cit. on p. 162).

Littlestone, N. and Warmuth, M. K. (2003). “Relating Data Compression and Learnability”. *Technical Report* (cit. on p. 138).

Liu, J., Shen, Z., He, Y., Zhang, X., Xu, R., Yu, H., and Cui, P. (2021a). “Towards Out-Of-Distribution Generalization: A Survey”. *arXiv.* Aug. DOI: 10.48550/arxiv.2108.13624 (cit. on p. 167).

Liu, T., Lu, J., Yan, Z., and Zhang, G. (2021b). “Statistical Generalization Performance Guarantee for Meta-Learning with Data Dependent Prior”. *Neurocomputing.* (C): 391–405. DOI: 10.1016/j.neucom.2021.09.018 (cit. on p. 183).

Livni, R. and Moran, S. (2017). “A Limitation of the PAC-Bayes Framework”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS).* Long Beach, CA, USA (cit. on pp. 119, 133, 144).

London, B. (2017). “A PAC-Bayesian Analysis of Randomized Learning with Application to Stochastic Gradient Descent”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS).* Long Beach, CA (cit. on p. 145).

London, B., Huang, B., Taskar, B., and Getoor, L. (2014). “PAC-Bayesian Collective Stability”. In: *Proc. Artif. Intell. Statist. (AISTATS).* Reykjavik, Iceland (cit. on p. 145).

Lopez, A. T. and Jog, V. (2018). “Generalization error bounds using Wasserstein distances”. In: *Proc. IEEE Inf. Theory Workshop (ITW).* Guangzhou, China. DOI: 10.1109/ITW.2018.8613445 (cit. on pp. 66, 67, 71).

Lotfi, S., Finzi, M., Kapoor, S., Potapczynski, A., Goldblum, M., and Wilson, A. G. (2022). “PAC-Bayes Compression Bounds So Tight That They Can Explain Generalization”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS).* New Orleans, LA, USA (cit. on pp. 155, 159, 160, 186).

Lugosi, G. and Neu, G. (2022). “Generalization Bounds via Convex Analysis”. In: *Proc. Conf. Learn. Theory (COLT)*. London, United Kingdom (cit. on p. 184).

Lugosi, G. and Neu, G. (2023). “Online-to-PAC Conversions: Generalization Bounds via Regret Analysis”. *arXiv*. May. DOI: 10.48550/arxiv.2305.19674 (cit. on pp. 184, 186).

Marton, K. (1996). “A Measure Concentration Inequality for Contracting Markov Chains”. *Geometric and functional analysis*. 6(3): 556–571 (cit. on p. 46).

Massart, P. (2007). *Concentration inequalities and model selection*. Ed. by J. Picard. *Lecture Notes in Mathematics: Saint-Flour Summer School on Probability Theory XXXIII 2003* (cit. on p. 46).

Maurer, A. (2004). “A Note on the PAC Bayesian Theorem”. *arXiv*. Nov. DOI: 10.48550/arxiv.cs/0411099 (cit. on pp. 49, 78, 90).

Mbacke, S. D., Clerc, F., and Germain, P. (2023a). “PAC-Bayesian Generalization Bounds for Adversarial Generative Models”. In: *Proc. Int. Conf. Mach. Learning (ICML)*. Honolulu, HI, USA (cit. on p. 93).

Mbacke, S. D., Clerc, F., and Germain, P. (2023b). “Statistical Guarantees for Variational Autoencoders using PAC-Bayesian Theory”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. New Orleans, LO, USA (cit. on p. 93).

McAllester, D. A. (1998). “Some PAC-Bayesian Theorems”. In: *Proc. Conf. Learn. Theory (COLT)*. Madison, WI, USA (cit. on pp. 4, 8, 27, 75, 124).

McAllester, D. A. (1999). “PAC-Bayesian Model Averaging”. In: *Proc. Conf. Comp. Learn. Theory (COLT)*. Santa Cruz, CA, USA. DOI: 10.1145/307400.307435 (cit. on pp. 4, 124).

McAllester, D. A. (2003a). “PAC-Bayesian Stochastic Model Selection”. *Mach. Learn.* 51(Apr.): 5–21 (cit. on pp. 27, 42, 90).

McAllester, D. A. (2003b). “Simplified PAC-Bayesian margin bounds”. In: *Proc. Conf. Comp. Learn. Theory (COLT)*. Santa Cruz, CA, USA (cit. on p. 49).

McAllester, D. A. (2013). “A PAC-Bayesian Tutorial with a Dropout Bound”. *arXiv*. July. DOI: 10.48550/arxiv.1307.2118 (cit. on pp. 50, 70, 79, 90).

Mendelson, S. (2014). “Learning without concentration”. In: *Proc. Conf. Learn. Theory (COLT).* Barcelona, Spain. DOI: 10.1145/2699439 (cit. on p. 52).

Mendelson, S. (2018). “Learning without concentration for general loss functions”. *Probability Theory and Related Fields.* 171(1-2): 459–502. DOI: 10.1007/s00440-017-0784-y (cit. on p. 52).

Meunier, D. and Alquier, P. (2021). “Meta-Strategy for Learning Tuning Parameters with Guarantees”. *Entropy.* 23(10). DOI: 10.3390/e23101257 (cit. on p. 183).

Mhammedi, Z., Grünwald, P., and Guedj, B. (2019). “PAC-Bayes Un-Expected Bernstein Inequality”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS).* Vol. 32. Vancouver, Canada (cit. on pp. 53, 89, 90, 155).

Mhammedi, Z., Guedj, B., and Williamson, R. C. (2020). “PAC-Bayesian Bound for the Conditional Value at Risk”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS).* Vol. 33 (cit. on p. 92).

Modak, E., Asnani, H., and Prabhakaran, V. M. (2021). “Rényi Divergence Based Bounds on Generalization Error”. In: *Proc. IEEE Inf. Theory Workshop (ITW).* Kanazawa, Japan. DOI: 10.1109/ITW48936.2021.9611387 (cit. on p. 71).

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of Machine Learning.* 2nd ed. *Adaptive Computation and Machine Learning.* Cambridge, MA: MIT Press (cit. on p. 11).

Mou, W., Wang, L., Zhai, X., and Zheng, K. (2018). “Generalization Bounds of SGLD for Non-convex Learning: Two Theoretical Viewpoints”. In: *Proc. Conf. Learning Theory (COLT).* Stockholm, Sweden (cit. on p. 160).

Murphy, K. P. (2022). *Probabilistic Machine Learning: An introduction.* MIT Press. URL: probml.ai (cit. on p. 151).

Nachum, I., Shafer, J., and Yehudayoff, A. (2018). “A Direct Sum Result for the Information Complexity of Learning”. In: *Proc. Conf. Learning Theory (COLT).* Stockholm, Sweden (cit. on pp. 133, 144).

Nachum, I. and Yehudayoff, A. (2019). “Average-Case Information Complexity of Learning”. In: *Proc. Conf. Alg. Learn. Theory (ALT).* Chicago, IL, USA (cit. on p. 133).

Nagarajan, V. and Kolter, J. Z. (2019). "Deterministic PAC-Bayesian generalization bounds for deep networks via generalizing noise-resilience". In: *Proc. Int. Conf. Learn. Representations (ICLR)*. New Orleans, LA (cit. on p. 162).

Neal, R. M. (1994). "Bayesian Learning for Neural Networks". *PhD thesis*. University of Toronto. DOI: 10.1007/978-1-4612-0745-0 (cit. on p. 161).

Negrea, J., Haghifam, M., Dziugaite, G. K., Khisti, A., and Roy, D. M. (2019). "Information-Theoretic Generalization Bounds for SGLD via Data-Dependent Estimates". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Vancouver, Canada (cit. on pp. 62, 70, 160).

Negrea, J., Dziugaite, G. K., and Roy, D. M. (2020). "In Defense of Uniform Convergence: Generalization via Derandomization with an Application to Interpolating Predictors". In: *Proc. Int. Conf. Mach. Learn. (ICML)*. Virtual Conference (cit. on p. 10).

Neu, G., Dziugaite, G. K., Haghifam, M., and Roy, D. M. (2021). "Information-Theoretic Generalization Bounds for Stochastic Gradient Descent". In: *Proc. Conf. Learn. Theory (COLT)*. Boulder, CO, USA (cit. on pp. 142, 160).

Neyshabur, B., Bhojanapalli, S., Mcallester, D. A., and Srebro, N. (2017). "Exploring Generalization in Deep Learning". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Long Beach, CA (cit. on p. 162).

Neyshabur, B., Bhojanapalli, S., and Srebro, N. (2018). "A PAC-Bayesian Approach to Spectrally-Normalized Margin Bounds for Neural Networks". In: *Proc. Int. Conf. Learn. Representations (ICLR)*. Vancouver, Canada (cit. on p. 160).

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., and Srebro, N. (2019). "The role of over-parametrization in generalization of neural networks". In: *Proc. Int. Conf. Learn. Representations (ICLR)*. New Orleans, LA (cit. on p. 10).

Neyshabur, B., Tomioka, R., and Srebro, N. (2015). "Norm-Based Capacity Control in Neural Networks". In: *Proc. Conf. Learn. Theory (COLT)*. Paris, France (cit. on pp. 16, 160).

Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010). "Estimating Divergence Functionals and the Likelihood Ratio by Convex Risk Minimization". *IEEE Trans. Inf. Theory.* 56(11): 5847–5861. DOI: 10.1109/TIT.2010.2068870 (cit. on p. 44).

Nozawa, K., Germain, P., and Guedj, B. (2020). "PAC-Bayesian Contrastive Unsupervised Representation Learning". In: *Proc. Conf. Uncertainty in Artif. Intell. (UAI)* (cit. on p. 92).

Ohnishi, Y. and Honorio, J. (2021). "Novel Change of Measure Inequalities with Applications to PAC-Bayesian Bounds and Monte Carlo Estimation". In: *Proc. Artif. Intell. Statist. (AISTATS).* San Diego, CA, USA (cit. on pp. 82, 83, 90).

Oneto, L., Donini, M., Pontil, M., and Shawe-Taylor, J. (2020). "Randomized learning and generalization of fair and private classifiers: From PAC-Bayes to stability and differential privacy". *Neurocomputing.* 416: 231–243. DOI: 10.1016/j.neucom.2019.12.137 (cit. on p. 145).

Orabona, F. (2023). "A modern introduction to online learning". *arXiv.* May. DOI: 10.48550/arxiv.1912.13213 (cit. on pp. 184, 186).

Palomar, D. P. and Verdu, S. (2008). "Lautum Information". *IEEE Trans. Inf. Theory.* 54(3): 964–975. DOI: 10.1109/TIT.2007.915715 (cit. on p. 126).

Parrado-Hernández, E., Ambroladze, A., Shawe-Taylor, J., and Sun, S. (2012). "PAC-Bayes Bounds with Data Dependent Priors". *Journal of Machine Learning Research (JMLR).* 13(112): 3507–3531. DOI: 10.1007/978-3-7908-2604-3_21 (cit. on p. 90).

Pensia, A., Jog, V., and Loh, P.-L. (2018). "Generalization Error Bounds for Noisy, Iterative Algorithms". In: *Proc. IEEE Int. Symp. Inf. Theory (ISIT).* Vail, CO, USA. DOI: 10.1109/ISIT.2018.8437571 (cit. on pp. 124, 150, 151, 160).

Pentina, A. and Lampert, C. (2014). "A PAC-Bayesian bound for Lifelong Learning". In: *Proc. Int. Conf. Mach. Learning (ICML).* Beijing, China (cit. on pp. 166, 183).

Pérez, G. V., Camargo, C. Q., and Louis, A. A. (2019). "Deep Learning Generalizes Because the Parameter-Function Map is Biased Towards Simple Functions". In: *Proc. Int. Conf. Learn. Representations (ICLR).* New Orleans, LA (cit. on p. 161).

Pérez-Ortiz, M., Rivasplata, O., Shawe-Taylor, J., and Szepesvári, C. (2021). "Tighter Risk Certificates for Neural Networks". *Journal of Machine Learning Research (JMLR)*. 22(227): 1–40 (cit. on p. 155).

Perlaza, S. M., Esnaola, I., Bisson, G., and Poor, H. V. (2023). "On the Validation of Gibbs Algorithms: Training Datasets, Test Datasets and their Aggregation". In: *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*. Taipei, Taiwan. DOI: 10.1109/ISIT54713.2023.10206506 (cit. on p. 128).

Pflug, G. C. (2000). "Some Remarks on the Value-at-Risk and the Conditional Value-at-Risk". In: *Probabilistic Constrained Optimization: Methodology and Applications*. Springer US. 272–281. ISBN: 978-1-4757-3150-7. DOI: 10.1007/978-1-4757-3150-7_15 (cit. on p. 93).

Pitas, K. (2020). "Dissecting Non-Vacuous Generalization Bounds Based on the Mean-Field Approximation". In: *Proc. Int. Conf. Mach. Learn. (ICML)*. Virtual Conference (cit. on p. 161).

Polyanskiy, Y. and Wu, Y. (2022). *Lecture Notes On Information Theory*. Cambridge, U.K.: Cambridge Univ. Press (cit. on pp. 31, 35, 36, 41, 43, 44, 85, 149).

Pradeep, A., Nachum, I., and Gastpar, M. (2022). "Finite Littlestone Dimension Implies Finite Information Complexity". In: *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*. Espoo, Finland. DOI: 10.1109/ISIT50566.2022.9834457 (cit. on p. 144).

Raginsky, M., Rakhlin, A., Tsao, M., Wu, Y., and Xu, A. (2016). "Information-theoretic analysis of stability and bias of learning algorithms". In: *Proc. IEEE Inf. Theory Workshop (ITW)*. DOI: 10.1109/ITW.2016.7606789 (cit. on pp. 28, 70).

Raginsky, M., Rakhlin, A., and Xu, A. (2021). "Information-Theoretic Stability and Generalization". In: *Information-Theoretic Methods in Data Science*. Ed. by M. R. D. Rodrigues and Y. C. Eldar. Cambridge University Press. 302–329. DOI: 10.1017/9781108616799.011 (cit. on pp. 71, 143).

Raginsky, M. and Sason, I. (2013). "Concentration of Measure Inequalities in Information Theory, Communications, and Coding". *Foundations and Trends in Communications and Information Theory*. 10(1-2): 1–246. DOI: 10.1561/0100000064 (cit. on p. 46).

Rakhlin, A., Mukherjee, S., and Poggio, T. (2005). "Stability results in learning theory". *Analysis and Applications.* 14(Oct.): 397–417. DOI: 10.1142/S0219530505000650 (cit. on p. 16).

Ralaivola, L., Szafranski, M., and Stempfel, G. (2010). "Chromatic PAC-Bayes Bounds for Non-IID Data: Applications to Ranking and Stationary beta-Mixing Processes". *Journal of Machine Learning Research (JMLR).* 11(Aug.): 1927–1956 (cit. on p. 184).

Rammal, M. R., Achille, A., Diggavi, S., Soatto, S., and Golatkar, A. (2022). "On Leave-One-Out Conditional Mutual Information For Generalization". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS).* New Orleans, LA, USA (cit. on pp. 114, 115, 119).

Redko, I., Morvant, E., Habrard, A., Sebban, M., and Bennani, Y. (2022). "A survey on domain adaptation theory: learning bounds and theoretical guarantees". *arXiv.* July. DOI: 10.48550/arxiv.2004.11829 (cit. on p. 167).

Rezazadeh, A., Jose, S. T., Durisi, G., and Simeone, O. (2021). "Conditional Mutual Information-Based Generalization Bound for Meta Learning". In: *Proc. IEEE Int. Symp. Inf. Theory (ISIT).* Melbourne, Australia. DOI: 10.1109/ISIT45174.2021.9518020 (cit. on p. 167).

Rezazadeh, A. (2022). "A Unified View on PAC-Bayes Bounds for Meta-Learning". In: *Proc. Int. Conf. Mach. Learning (ICML).* Baltimore, MD, USA (cit. on pp. 166, 183).

Rigollet, P. and Tsybakov, A. B. (2012). "Sparse Estimation by Exponential Weighting". *Statistical Science.* 27(4): 558–575. DOI: 10.1214/12-STS393 (cit. on pp. 91, 92).

Riou, C., Alquier, P., and Chérief-Abdellatif, B.-E. (2023). "Bayes meets Bernstein at the Meta Level: an Analysis of Fast Rates in Meta-Learning with PAC-Bayes". *arXiv.* Feb. DOI: 10.48550/arxiv.2302.11709 (cit. on p. 183).

Rissanen, J. (1978). "Modeling by shortest data description". *Automatica.* 14(5): 465–471. DOI: https://doi.org/10.1016/0005-1098(78)90005-5 (cit. on p. 5).

Rissanen, J. (1983). "A Universal Prior for Integers and Estimation by Minimum Description Length". *The Annals of Statistics.* 11(2): 416–431 (cit. on p. 5).

Rivasplata, O., Kuzborskij, I., Szepesvari, C., and Shawe-Taylor, J. (2020). “PAC-Bayes Analysis Beyond the Usual Bounds”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS).* Vancouver, Canada (cit. on pp. 69, 76, 77, 90, 91).

Rivasplata, O., Parrado-Hernández, E., Shawe-Taylor, J., Sun, S., and Szepesvári, C. (2018). “PAC-Bayes Bounds for Stable Algorithms with Instance-Dependent Priors”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS).* Montréal, Canada (cit. on pp. 84, 90, 145).

Rivasplata, O., Tankasali, V. M., and Szepesvari, C. (2019). “PAC-Bayes with Backprop”. *arXiv.* Oct. DOI: 10.48550/arxiv.1908.07380 (cit. on p. 161).

Rockafellar, R. T. (1970). *Convex analysis. Princeton Mathematical Series.* Princeton, N. J., USA: Princeton University Press. DOI: 10.1515/9781400873173 (cit. on p. 43).

Rodríguez-Gálvez, B., Bassi, G., and Skoglund, M. (2021a). “Upper Bounds on the Generalization Error of Private Algorithms for Discrete Data”. *IEEE Trans. Inf. Theory.* 67(11): 7362–7379. DOI: 10.1109/TIT.2021.3111480 (cit. on p. 145).

Rodríguez-Gálvez, B., Bassi, G., Thobaben, R., and Skoglund, M. (2020). “On Random Subset Generalization Error Bounds and the Stochastic Gradient Langevin Dynamics Algorithm”. In: *Proc. IEEE Inf. Theory Workshop (ITW).* Riva del Garda, Italy. DOI: 10.1109/ITW46852.2021.9457578 (cit. on pp. 70, 95, 101, 119).

Rodríguez-Gálvez, B., Bassi, G., Thobaben, R., and Skoglund, M. (2021b). “Tighter expected generalization error bounds via Wasserstein distance”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS).* Virtual Conference (cit. on pp. 67–69, 71).

Rodríguez-Gálvez, B., Thobaben, R., and Skoglund, M. (2023). “More PAC-Bayes bounds: From bounded losses, to losses with general tail behaviors, to anytime-validity”. In: *Proc. Int. Conf. Mach. Learning (ICML). Workshop on PAC-Bayes Meets Interactive Learning (PBMIL).* Honolulu, HI, USA (cit. on pp. 77, 179).

Rogers, W. H. and Wagner, T. J. (1978). “A Finite Sample Distribution-Free Performance Bound for Local Discrimination Rules”. *The Annals of Statistics.* 6(3): 506–514. DOI: 10.1214/aos/1176344196 (cit. on p. 15).

Rothfuss, J., Fortuin, V., Josifoski, M., and Krause, A. (2021). "PACOH: Bayes-Optimal Meta-Learning with PAC-Guarantees". In: *Proc. Int. Conf. Mach. Learning (ICML).* Virtual conference (cit. on pp. 166, 183).

Rubner, Y., Tomasi, C., and Guibas, L. (1998). "A metric for distributions with applications to image databases". In: *Int. Conf. Computer Vision.* Mumbai, India. DOI: 10.1109/ICCV.1998.710701 (cit. on p. 39).

Ruderman, A., Reid, M. D., Garcia-Garcia, D., and Petterson, J. (2012). "Tighter Variational Representations of F-Divergences via Restriction to Probability Measures". In: *Proc. Int. Conf. Mach. Learning (ICML).* Edinburgh, Scotland (cit. on p. 44).

Rudin, W. (1987). *Real and Complex Analysis, 3rd Ed.* USA: McGraw-Hill, Inc. (cit. on p. 41).

Russo, D. and Zou, J. (2016). "Controlling bias in adaptive data analysis using information theory". In: *Proc. Artif. Intell. Statist. (AISTATS).* Cadiz, Spain (cit. on pp. 5, 8, 27, 28, 57, 70).

Russo, D. and Van Roy, B. (2016). "An Information-Theoretic Analysis of Thompson Sampling". *Journal of Machine Learning Research (JMLR).* 17(1): 2442–2471 (cit. on p. 181).

Sachs, S., van Erven, T., Hodgkinson, L., Khanna, R., and Şimşekli, U. (2023). "Generalization Guarantees via Algorithm-dependent Rademacher Complexity". In: *Proc. Conf. Learn. Theory (COLT).* Bangalore, India (cit. on p. 119).

Salmon, J. and Dalalyan, A. (2011). "Optimal aggregation of affine estimators". In: *Proc. Conf. Learn. Theory (COLT).* Budapest, Hungary (cit. on pp. 9, 91, 92).

Samson, P.-M. (2000). "Concentration of Measure Inequalities for Markov Chains and Φ-Mixing Processes". *The Annals of Probability.* 28(1): 416–461 (cit. on p. 46).

Saunshi, N., Plevrakis, O., Arora, S., Khodak, M., and Khandeparkar, H. (2019). "A Theoretical Analysis of Contrastive Unsupervised Representation Learning". In: *Proc. Int. Conf. Mach. Learning (ICML)* (cit. on p. 92).

Saxe, A. M., Bansal, Y., Dapello, J., Advani, M., Kolchinsky, A., Tracey, B. D., and Cox, D. D. (2019). “On the Information Bottleneck Theory of Deep Learning”. In: *Proc. Int. Conf. Learn. Representations (ICLR)*. New Orleans, LA. DOI: 10.1088/1742-5468/ab3985 (cit. on p. 29).

Schwarz, G. (1978). “Estimating the Dimension of a Model”. *The Annals of Statistics*. 6(2): 461–464 (cit. on p. 5).

Seeger, M. (2002). “PAC-Bayesian Generalisation Error Bounds for Gaussian Process Classification”. *Journal of Machine Learning Research (JMLR)*. 3(Oct.): 233–269 (cit. on pp. 89–91, 183).

Sefidgaran, M., Chor, R., and Zaidi, A. (2022a). “Rate-Distortion Theoretic Bounds on Generalization Error for Distributed Learning”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. New Orleans, LA, USA (cit. on p. 183).

Sefidgaran, M., Gohari, A., Richard, G., and Simsekli, U. (2022b). “Rate-Distortion Theoretic Generalization Bounds for Stochastic Learning Algorithms”. In: *Proc. Conf. Learn. Theory (COLT)*. Boulder, CO, USA (cit. on p. 71).

Sefidgaran, M., Zaidi, A., and Krasnowski, P. (2023). “Minimum Description Length and Generalization Guarantees for Representation Learning”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. New Orleans, LO, USA (cit. on p. 119).

Seldin, Y., Auer, P., Shawe-taylor, J., Ortner, R., and Laviolette, F. (2011). “PAC-Bayesian Analysis of Contextual Bandits”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Vancouver, Canada (cit. on pp. 180, 183).

Seldin, Y., Cesa-Bianchi, N., Auer, P., Laviolette, F., and Shawe-Taylor, J. (2012a). “PAC-Bayes-Bernstein Inequality for Martingales and its Application to Multiarmed Bandits”. In: *Proc. Workshop On-line Trading of Exploration and Exploitation*. 98–111 (cit. on p. 180).

Seldin, Y., Laviolette, F., Cesa-Bianchi, N., Shawe-Taylor, J., and Auer, P. (2012b). “PAC-Bayesian Inequalities for Martingales”. *IEEE Trans. Inf. Theory*. 58(12): 7086–7093. DOI: 10.1109/TIT.2012.2211334 (cit. on pp. 52, 77, 178–180, 183).

Seldin, Y. and Tishby, N. (2010). “PAC-Bayesian Analysis of Co-clustering and Beyond”. *Journal of Machine Learning Research (JMLR).* 11(117): 3595–3646 (cit. on p. 184).

Settles, B. (2012). *Active Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning.* Morgan and Claypool Publishers. DOI: 10.1007/978-3-031-01560-1 (cit. on p. 186).

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms.* Cambridge, U.K.: Cambridge Univ. Press. DOI: 10.1017/CBO9781107298019 (cit. on pp. 10–16, 104).

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. (2010). “Learnability, Stability and Uniform Convergence”. *Journal of Machine Learning Research (JMLR).* 11(Dec.): 2635–2670 (cit. on pp. 16, 176).

Shannon, C. E. (1948). “A Mathematical Theory of Communication”. *The Bell System Technical Journal.* 27: 379–423. DOI: 10.1063/1.3067010 (cit. on pp. 5, 19).

Sharma, A., Veer, S., Hancock, A., Yang, H., Pavone, M., and Majumdar, A. (2023). “PAC-Bayes Generalization Certificates for Learned Inductive Conformal Prediction”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS).* New Orleans, LO, USA (cit. on p. 184).

Shawe-Taylor, J. and Cristianini, N. (1999). “Margin Distribution Bounds on Generalization”. In: *Proc. European Conf. Comp. Learn. Theory (EuroCOLT).* Nordkirchen, Germany. DOI: 10.1007/3-540-49097-3_21 (cit. on p. 16).

Shawe-Taylor, J. and Hardoon, D. (2009). “PAC-Bayes Analysis Of Maximum Entropy Classification”. In: *Proc. Artif. Intell. Statist. (AISTATS).* Clearwater Beach, FL, USA (cit. on p. 183).

Shawe-Taylor, J. and Williamson, R. C. (1997). “A PAC Analysis of a Bayesian Estimator”. In: *Proc. Conf. Learn. Theory (COLT).* Nashville, TN, USA. DOI: 10.1145/267460.267466 (cit. on pp. 4, 8, 27, 75).

Shwartz-Ziv, R. and Alemi, A. A. (2020). “Information in Infinite Ensembles of Infinitely-Wide Neural Networks”. In: *Proc. Symposium on Advances in Approximate Bayesian Inference.* Vancouver, Canada (cit. on p. 161).

Shwartz-Ziv, R. and Tishby, N. (2017). "Opening the Black Box of Deep Neural Networks via Information". *arXiv.* DOI: 10.48550/arxiv.1703.00810 (cit. on p. 29).

Solomonoff, R. (1964). "A formal theory of inductive inference. Part I". *Information and Control.* 7(1): 1–22. ISSN: 0019-9958. DOI: https://doi.org/10.1016/S0019-9958(64)90223-2 (cit. on p. 5).

Steinke, T. and Zakynthinou, L. (2020). "Reasoning About Generalization via Conditional Mutual Information". In: *Proc. Conf. Learn. Theory (COLT).* Graz, Austria (cit. on pp. 50, 94, 96, 98, 99, 110, 118, 119, 130, 134, 135, 139, 144, 145).

Sun, S., Yu, M., Shawe-Taylor, J., and Mao, L. (2022). "Stability-based PAC-Bayes analysis for multi-view learning algorithms". *Information Fusion.* 86-87(Oct.): 76–92. DOI: 10.1016/j.inffus.2022.06.006 (cit. on p. 145).

Thiemann, N., Igel, C., Wintenberger, O., and Seldin, Y. (2017). "A Strongly Quasiconvex PAC-Bayesian Bound". In: *Proc. Conf. Alg. Learn. Theory (ALT).* Kyoto, Japan (cit. on p. 92).

Thompson, W. R. (1933). "On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples". *Biometrika.* 25(3/4): 285–294. DOI: 10.2307/2332286 (cit. on p. 181).

Thrun, S. and Pratt, L. (1998). *Learning to Learn: Introduction and Overview.* Boston, MA, USA: Springer. DOI: 10.1007/978-1-4615-5529-2_1 (cit. on p. 164).

Tinsi, L. and Dalalyan, A. (2022). "Risk bounds for aggregated shallow neural networks using Gaussian priors". In: *Proc. Conf. Learn. Theory (COLT).* London, United Kingdom (cit. on p. 162).

Tishby, N., Pereira, F. C., and Bialek, W. (1999). "The information bottleneck method". In: *Allerton Conf. Communication, Control, Computing (Allerton).* Monticello, IL, USA (cit. on p. 28).

Tolstikhin, I. O. and Seldin, Y. (2013). "PAC-Bayes-Empirical-Bernstein Inequality". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS).* Lake Tahoe, NV, United States (cit. on pp. 49, 92).

Tomamichel, M. and Hayashi, M. (2018). "Operational interpretation of Rényi information measures via composite hypothesis testing against product and Markov distributions". *IEEE Trans. Inf. Theory.* 64(2): 1064–1082 (cit. on p. 38).

Tsuzuku, Y., Sato, I., and Sugiyama, M. (2020). "Normalized Flat Minima: Exploring Scale Invariant Definition of Flat Minima for Neural Networks Using PAC-Bayesian Analysis". In: *Proc. Int. Conf. Mach. Learn. (ICML)*. Virtual Conference (cit. on p. 161).

Valiant, L. G. (1984). "A Theory of the Learnable". *Commun. ACM*. 27(11): 1134–1142. DOI: 10.1145/1968.1972 (cit. on p. 4).

Van Erven, T., Grünwald, P., Mehta, N., Reid, M., and Williamson, R. (2015). "Fast rates in statistical and online learning". *Journal of Machine Learning Research (JMLR)*. 16(Sept.): 1793–1861 (cit. on p. 14).

Van Erven, T. and Harremoës, P. (2014). "Rényi divergence and Kullback-Leibler divergence". *IEEE Trans. Inf. Theory*. 60(7): 3797–3820. DOI: 10.1109/TIT.2014.2320500 (cit. on p. 36).

Van Handel, R. (2016). *Probability in High Dimension*. URL: web.math.princeton.edu/%7EErvan/APC550.pdf (cit. on p. 69).

Vapnik, V. and Chervonenkis, A. (1971). "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities". *Theory of Probability & Its Applications*. 16(2): 264–280. DOI: 10.1007/978-3-319-21852-6_3 (cit. on p. 4).

Vapnik, V. and Chervonenkis, A. (1974). *Theory of Pattern Recognition [in Russian]*. Moscow: Nauka (cit. on p. 14).

Verdú, S. (2015). "α-Mutual Information". In: *Proc. Inf. Theory Appl. Workshop (ITA)*. San Diego, CA, USA (cit. on pp. 37, 143).

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science. Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press. DOI: 10.1017/9781108231596 (cit. on p. 70).

Viallard, P., Emonet, R., Germain, P., Habrard, A., and Morvant, E. (2019). "Interpreting Neural Networks as Majority Votes through the PAC-Bayesian Theory". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS). Workshop on Machine Learning with guarantees*. Vancouver, Canada (cit. on p. 161).

Viallard, P., Haddouche, M., Şimşekli, U., and Guedj, B. (2023). "Learning via Wasserstein-Based High Probability Generalisation Bounds". *arXiv*. June. DOI: 10.48550/arXiv.2306.04375 (cit. on p. 93).

Viallard, P., Vidot, E. G., Habrard, A., and Morvant, E. (2021). “A PAC-Bayes Analysis of Adversarial Robustness”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Virtual Conference (cit. on p. 162).

Villani, C. (2008). *Optimal transport – Old and new*. Vol. 338. *Grundlehren der mathematischen Wissenschaften*. Springer Science & Business Media (cit. on pp. 38, 45, 68).

Wainwright, M. J. (2019). *High-Dimensional Statistics: a Non-Asymptotic Viewpoint*. Cambridge, U.K.: Cambridge Univ. Press. DOI: 10.1017/9781108627771 (cit. on pp. 46–48, 51).

Wang, B., Zhang, H., Zhang, J., Meng, Q., Chen, W., and Liu, T.-Y. (2021a). “Optimizing Information-theoretical Generalization Bound via Anisotropic Noise of SGLD”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Virtual Conference (cit. on p. 160).

Wang, H., Diaz, M., Santos Filho, J. C. S., and Calmon, F. (2019a). “An Information-Theoretic View of Generalization via Wasserstein Distance”. In: *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*. Paris, France. DOI: 10.1109/ISIT.2019.8849359 (cit. on pp. 66, 71).

Wang, H., Huang, Y., Gao, R., and Calmon, F. (2021b). “Analyzing the Generalization Capability of SGLD Using Properties of Gaussian Channels”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Virtual Conference (cit. on p. 160).

Wang, H., Huang, Y., Gao, R., and Calmon, F. (2023). “Analyzing the Generalization Capability of SGLD Using Properties of Gaussian Channels”. *Journal of Machine Learning Research (JMLR)*. 24(26): 1–43 (cit. on p. 160).

Wang, H., Zheng, S., Xiong, C., and Socher, R. (2019b). “On the Generalization Gap in Reparameterizable Reinforcement Learning”. In: *Proc. Int. Conf. Mach. Learning (ICML)*. Long Beach, CA, USA (cit. on p. 183).

Wang, Z., Huang, S.-L., Kuruoglu, E. E., Sun, J., Chen, X., and Zheng, Y. (2022). “PAC-Bayes Information Bottleneck”. In: *Proc. Int. Conf. Learn. Representations (ICLR)*. Virtual Conference (cit. on p. 161).

Wang, Z. and Mao, Y. (2022). “On the Generalization of Models Trained with SGD: Information-Theoretic Bounds and Implications”. In: *Proc. Int. Conf. Learn. Representations (ICLR).* Virtual Conference (cit. on p. 160).

Wang, Z. and Mao, Y. (2023a). “Information-Theoretic Analysis of Unsupervised Domain Adaptation”. In: *Proc. Int. Conf. Learn. Representations (ICLR).* Kigali, Rwanda (cit. on pp. 168, 170–172, 175, 183).

Wang, Z. and Mao, Y. (2023b). “Sample-Conditioned Hypothesis Stability Sharpens Information-Theoretic Generalization Bounds”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS).* New Orleans, LO, USA (cit. on pp. 119, 160).

Wang, Z. and Mao, Y. (2023c). “Tighter Information-Theoretic Generalization Bounds from Supersamples”. In: *Proc. Int. Conf. Mach. Learning (ICML).* Honolulu, HI, USA (cit. on pp. 95, 110, 112, 114, 119, 157, 158, 185).

Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). “A survey of transfer learning”. *Journal of Big Data.* 3(1): 1–40. DOI: 10.1186/s40537-016-0043-6 (cit. on p. 167).

Wintenberger, O. (2015). “Weak transport inequalities and applications to exponential and oracle inequalities”. *Electronic Journal of Probability.* 20: 1–27. DOI: 10.1214/EJP.v20-3558 (cit. on p. 70).

Wongso, S., Ghosh, R., and Motani, M. (2022). “Understanding Deep Neural Networks Using Sliced Mutual Information”. In: *Proc. IEEE Int. Symp. Inf. Theory (ISIT).* Espoo, Finland. DOI: 10.1109/ISIT50566.2022.9834357 (cit. on p. 71).

Wongso, S., Ghosh, R., and Motani, M. (2023). “Using Sliced Mutual Information to Study Memorization and Generalization in Deep Neural Networks”. In: *Proc. Artif. Intell. Statist. (AISTATS).* Valencia, Spain (cit. on p. 71).

Wu, X., Manton, J. H., Aickelin, U., and Zhu, J. (2022a). “An Information-Theoretic Analysis for Transfer Learning: Error Bounds and Applications”. In: *Proc. IEEE Int. Symp. Inf. Theory (ISIT).* Los Angeles, CA, USA (cit. on pp. 144, 173, 175, 183).

Wu, X., Manton, J. H., Aickelin, U., and Zhu, J. (2022b). "Fast Rate Generalization Error Bounds: Variations on a Theme". In: *Proc. IEEE Inf. Theory Workshop (ITW)*. Mumbai, India. DOI: 10.1109/ITW54588.2022.9965761 (cit. on p. 130).

Xiao, J., Sun, R., and Luo, Z.-Q. (2023). "PAC-Bayesian Spectrally-Normalized Bounds for Adversarially Robust Generalization". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. New Orleans, LO, USA (cit. on p. 162).

Xu, A. and Raginsky, M. (2017). "Information-theoretic analysis of generalization capability of learning algorithms". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Long Beach, CA, USA (cit. on pp. 5, 8, 27, 28, 55, 69–71, 91, 132, 144).

Xu, A. and Raginsky, M. (2022). "Minimum Excess Risk in Bayesian Learning". *IEEE Trans. Inf. Theory*. 68(12): 7935–7955. DOI: 10.1109/TIT.2022.3176056 (cit. on pp. 72, 180).

Yagli, S., Dytso, A., and Poor, H. V. (2020). "Information-Theoretic Bounds on the Generalization Error and Privacy Leakage in Federated Learning". In: *Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*. Atlanta, GA, USA. DOI: 10.1109/SPAWC48557.2020.9154277 (cit. on pp. 177, 183).

Yang, J., Sun, S., and Roy, D. M. (2019). "Fast-rate PAC-Bayes Generalization Bounds via Shifted Rademacher Processes". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Vancouver, Canada (cit. on p. 92).

Yang, Y. and Barron, A. (1999). "Information-theoretic determination of minimax rates of convergence". *The Annals of Statistics*. 27(5): 1564–1599. DOI: 10.1214/aos/1017939142 (cit. on p. 5).

Zantedeschi, V., Viallard, P., Morvant, E., Emonet, R., Habrard, A., Germain, P., and Guedj, B. (2021). "Learning Stochastic Majority Votes by Minimizing a PAC-Bayes Generalization Bound". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Virtual Conference (cit. on pp. 91, 162).

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2021). "Understanding Deep Learning (Still) Requires Rethinking Generalization". *Commun. ACM*. 64(3): 107–115. DOI: 10.1145/3446776 (cit. on p. 153).

Zhang, T. (2006). “Information-theoretic upper and lower bounds for statistical estimation”. *IEEE Trans. Inf. Theory*. 52(4): 1307–1321. DOI: 10.1109/TIT.2005.864439 (cit. on pp. 5, 27, 75, 89, 124).

Zhou, R., Tian, C., and Liu, T. (2021). “Individually Conditional Individual Mutual Information Bound on Generalization Error”. In: *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*. Melbourne, Australia. DOI: 10.1109/TIT.2022.3144615 (cit. on pp. 70, 101, 119, 130, 144).

Zhou, R., Tian, C., and Liu, T. (2022). “Stochastic Chaining and Strengthened Information-Theoretic Generalization Bounds”. In: *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*. DOI: 10.1016/j.jfranklin.2023.02.009 (cit. on pp. 65, 70, 130, 144).

Zhou, R., Tian, C., and Liu, T. (2023a). “Exactly Tight Information-Theoretic Generalization Error Bound for the Quadratic Gaussian Problem”. In: *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*. Taipei, Taiwan. DOI: 10.1109/ISIT54713.2023.10206951 (cit. on pp. 131, 144, 185).

Zhou, S., Lei, Y., and Kaban, A. (2023b). “Toward Better PAC-Bayes Bounds for Uniformly Stable Algorithms”. In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. New Orleans, LO, USA (cit. on p. 145).

Zhou, W., Veitch, V., Austern, M., Adams, R., and Orbanz, P. (2019). “Non-vacuous Generalization Bounds at the ImageNet Scale: a PAC-Bayesian Compression Approach”. In: *Proc. Int. Conf. Learn. Representations (ICLR)*. New Orleans, LA (cit. on p. 159).