Generalization Bounds: Perspectives from Information Theory and PAC-Bayes

Other titles in Foundations and Trends® in Machine Learning

An Introduction to Deep Survival Analysis Models for Predicting Time-to-Event Outcomes
George H. Chen
ISBN: 978-1-63828-454-3

Automated Deep Learning: Neural Architecture Search Is Not the End
Xuanyi Dong, David Jacob Kedziora, Katarzyna Musial and Bogdan Gabrys
ISBN: 978-1-63828-318-8

AutonoML: Towards an Integrated Framework for Autonomous Machine Learning
David Jacob Kedziora, Katarzyna Musial and Bogdan Gabrys
ISBN: 978-1-63828-316-4

Causal Fairness Analysis: A Causal Toolkit for Fair Machine Learning
Drago Plečko and Elias Bareinboim
ISBN: 978-1-63828-330-0

User-friendly Introduction to PAC-Bayes Bounds
Pierre Alquier
ISBN: 978-1-63828-326-3

A Friendly Tutorial on Mean-Field Spin Glass Techniques for Non-Physicists
Andrea Montanari and Subhabrata Sen
ISBN: 978-1-63828-212-9

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
University College London
benjamin.guedj@inria.fr

Maxim Raginsky
University of Illinois
maxim@illinois.edu

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Machine Learning

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

F. Hellström *et al.*. *Generalization Bounds: Perspectives from Information Theory and PAC-Bayes*. Foundations and Trends® in Machine Learning, vol. 18, no. 1, pp. 1–223, 2025.

ISBN: 978-1-63828-421-5

Foundations and Trends® in Machine Learning
Volume 18, Issue 1, 2025

Editorial Scope

Foundations and Trends® welcomes monographs that touch on fundamental problems in machine learning from theoretical, methodological, and/or computational perspectives. We are particularly interested in monographs that seek to bridge such problems and perspectives with those from related fields, including (but not limited to) statistics, economics, and optimization.

Information for Librarians

Foundations and Trends® in Machine Learning, 2025, Volume 18, 6 issues. ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.

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, Department of Computer Science and Centre for Artificial Intelligence, UK; f.hellstrom@ucl.ac.uk*
[2] *Chalmers University of Technology, Sweden; durisi@chalmers.se*
[3] *Inria, France and University College London, Department of Computer Science and Centre for Artificial Intelligence, UK; benjamin.guedj@inria.fr*
[4] *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.

Fredrik Hellström, Giuseppe Durisi, Benjamin Guedj and Maxim Raginsky (2025), "Generalization Bounds: Perspectives from Information Theory and PAC-Bayes", Foundations and Trends® in Machine Learning: Vol. 18, No. 1, pp 1–223. DOI: 10.1561/2200000112.

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 in many industries and daily life. As these technologies continue to gain momentum, the need to develop a deeper understanding of their underlying principles, capabilities, and limitations grows. 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 considered 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 Shawe-Taylor and Williamson (1997), McAllester (1998), and McAllester (1999), 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), Barron and Cover (1991) and 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 conventions apply, 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 an ordered list: 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 Sections 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 $\text{gen}(W, \boldsymbol{Z})$ 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 $\mathbb{E}_{P_{W\boldsymbol{Z}}}[\text{gen}(W, \boldsymbol{Z})]$. 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 Section 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, $\mathbb{E}_{P_{W|\boldsymbol{Z}}}[\text{gen}(W, \boldsymbol{Z})]$, 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 Section 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 (Salmon and Dalalyan, 2011; Dalalyan and Salmon, 2012) 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 (Catoni, 2007; Alquier and Biau, 2013; 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.*, Neyshabur *et al.*, 2019; Negrea *et al.*, 2020).

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 Gine and Zinn (1984), Koltchinskii and Panchenko (2000), Koltchinskii (2001), Bartlett and Mendelson (2001), and Bartlett and Mendelson (2002) 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 Shalev-Shwartz and Ben-David (2014) and Mohri *et al.* (2018) 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 Theorem 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 Theorem 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* (Gine and Zinn, 1984; Koltchinskii and Panchenko, 2000; Koltchinskii, 2001; Bartlett and Mendelson, 2001; Bartlett and Mendelson, 2002). 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 $w \in \mathcal{W}$,

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

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 (Rogers and Wagner, 1978; Devroye and Wagner, 1979). As for the section on uniform convergence, we will only provide a very short presentation to provide context for upcoming sections, 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 Sections 2 to 6, we cover the foundations of information-theoretic and PAC-Bayesian generalization bounds for standard supervised learning. Specifically, in Section 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 sections. In Section 3, we overview the core tools that are used in deriving generalization bounds in the upcoming sections, in the form of information measures, change of measure techniques, and concentration inequalities. We use these tools to derive generalization bounds in expectation in Section 4 and generalization bounds in probability in Section 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 Sections 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 Section 7, we examine the *information complexity* of several learning algorithms, that is, the value of information measures that the learning algorithms induce. In Section 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 Section 9, we derive bounds for alternative learning models, namely meta learning, out-of-distribution generalization, federated learning, and reinforcement learning. Finally, in Section 10, we provide concluding remarks and a broader discussion of information-theoretic and PAC-Bayesian generalization bounds as a whole.

Part II

Applications and Additional Topics

7

The Information Complexity of Learning Algorithms

As argued in Section 2.2, one benefit 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 section, 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. Next, 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 sections. 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 Section 8.

7.1 The Gibbs Posterior

Given a generalization bound, it is tempting to design a learning algorithm 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 mainly 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.

Theorem 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 Theorem 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 $P_{\boldsymbol{Z}}$, while the learning algorithm has to be designed without knowledge of $P_{\boldsymbol{Z}}$. 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

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

Thus, applying Theorem 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 (McAllester, 1998; McAllester, 1999; Zhang, 2006; Catoni, 2007). 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 Raginsky (2019).

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

$$\begin{aligned}&\mathbb{E}[L_{P_Z}(W)] - \min_{w \in \mathcal{W}} L_{P_Z}(w) \\ &\leq \frac{M\pi\rho^2 d}{\lambda} + \frac{1}{2\lambda\rho^2}\left(R + \sqrt{\frac{2\pi\rho^2 d}{\lambda}}\right)^2 + \frac{d}{2\lambda}\log\frac{\lambda}{d} - \frac{1}{\lambda}\log V_d + \frac{\lambda\sigma^2}{2n},\end{aligned} \tag{7.11}$$

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 $\|w^*_{\boldsymbol{Z}}\| \leq R$ 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

$$L_{\boldsymbol{Z}}(w) - L_{\boldsymbol{Z}}(w^*_{\boldsymbol{Z}}) \leq \frac{M}{2}\|w - w^*_{\boldsymbol{Z}}\|^2. \tag{7.12}$$

Therefore,

$$\begin{aligned}\Psi_\lambda(\boldsymbol{Z}) &= -\lambda L_{\boldsymbol{Z}}(w^*_{\boldsymbol{Z}}) + \log \mathbb{E}_{Q_W}[\exp\left(-\lambda\left(L_{\boldsymbol{Z}}(W) - L_{\boldsymbol{Z}}(w^*_{\boldsymbol{Z}})\right)\right)] &(7.13)\\ &\geq -\lambda L_{\boldsymbol{Z}}(w^*_{\boldsymbol{Z}}) + \log \mathbb{E}_{Q_W}\left[\exp\left(-\frac{\lambda M}{2}\|W - w^*_{\boldsymbol{Z}}\|^2\right)\right], &(7.14)\end{aligned}$$

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).[13] With this, Aminian *et al.* (2021a) derived the following exact characterization of the average generalization error of the Gibbs posterior.

[13] This provides a strong incitement to refer to $I_{\mathrm{SKL}}(\cdot;\cdot)$ as the mutualautum information, but we digress.

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

$$\mathbb{E}_{P_{\boldsymbol{Z}} P^G_{W|\boldsymbol{Z}}}[L_{P_{\boldsymbol{Z}}}(W) - L_{\boldsymbol{Z}}(W)] = \frac{I_{\mathrm{SKL}}(W;\boldsymbol{Z})}{\lambda}. \tag{7.20}$$

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

$$\begin{aligned} I_{\mathrm{SKL}}(W;\boldsymbol{Z}) &= \mathbb{E}_{P_{\boldsymbol{Z}} P^G_{W|\boldsymbol{Z}}}\left[\log \frac{P^G_{W|\boldsymbol{Z}}}{P^G_W}\right] + \mathbb{E}_{P_{\boldsymbol{Z}} P^G_W}\left[\log \frac{P^G_W}{P^G_{W|\boldsymbol{Z}}}\right] \qquad (7.21)\\ &= \mathbb{E}_{P_{\boldsymbol{Z}} P^G_{W|\boldsymbol{Z}}}\left[\log P^G_{W|\boldsymbol{Z}}\right] - \mathbb{E}_{P_{\boldsymbol{Z}} P^G_W}\left[\log P^G_{W|\boldsymbol{Z}}\right]. \qquad (7.22) \end{aligned}$$

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

$$\begin{aligned} &\mathbb{E}_{P_{\boldsymbol{Z}} P^G_{W|\boldsymbol{Z}}}\left[\log P^G_{W|\boldsymbol{Z}}\right] - \mathbb{E}_{P_{\boldsymbol{Z}} P^G_W}\left[\log P^G_{W|\boldsymbol{Z}}\right]\\ =&\,\mathbb{E}_{P_{\boldsymbol{Z}} P^G_{W|\boldsymbol{Z}}}[-\lambda L_{\boldsymbol{Z}}(W)] - \mathbb{E}_{P_{\boldsymbol{Z}} P^G_W}[-\lambda L_{\boldsymbol{Z}}(W)] \qquad (7.25)\\ =&\,\lambda\, \mathbb{E}_{P_{\boldsymbol{Z}} P^G_{W|\boldsymbol{Z}}}[L_{P_{\boldsymbol{Z}}}(W) - L_{\boldsymbol{Z}}(W)]. \qquad (7.26) \end{aligned}$$

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, 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 Sections 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, as mentioned in Section 7.1, note that 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 Theorem 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 Section 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 $\overline{\text{gen}} \le 13\sigma^2/n$, 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 appears to be 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 Section 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, the VC dimension is a fundamental quantity that characterizes distribution- and algorithm-independent learnability for binary classification. 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 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.*, Theorem 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 Theorem 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 (Theorem 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 Section 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 Section 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

$$I(W;\boldsymbol{S}|\tilde{\boldsymbol{Z}}) \leq I(W;\boldsymbol{S}|\tilde{\boldsymbol{Z}} = \tilde{z}_*) \leq \log\left|\hat{\mathcal{W}}\right|. \tag{7.39}$$

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 (Theorem 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. To address 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 $\boldsymbol{\Lambda}$ 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, $I(\boldsymbol{\Lambda}; \boldsymbol{S}|\tilde{\boldsymbol{Z}}) = I(\mathbf{F}; \boldsymbol{S}|\tilde{\boldsymbol{Z}})$, 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 $n > d_{\mathrm{VC}}$,

$$I(\mathbf{F}; \boldsymbol{S}|\tilde{\boldsymbol{Z}}) \leq d_{\mathrm{VC}} \log\left(\frac{2en}{d_{\mathrm{VC}}}\right). \tag{7.41}$$

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

$$I(\mathbf{F}; \boldsymbol{S}|\tilde{\boldsymbol{Z}}) \leq I(F; \boldsymbol{S}|\tilde{\boldsymbol{Z}} = \tilde{\boldsymbol{z}}_*) \leq \log\left|\hat{\mathcal{F}}\right|. \tag{7.42}$$

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 (Theorem 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

$$\hat{R}_{\text{loo}} = \mathbb{E}_{P_U}\left[\mathbb{E}_{P_{\dot{\mathbf{\Lambda}}|U\dot{\boldsymbol{Z}}}}\left[\dot{\Lambda}_U\right]\right]. \tag{7.43}$$

In words, given a supersample $\dot{\boldsymbol{Z}}$, $\hat{R}_{\text{loo}}$ is the test loss when leaving out the Uth sample, averaged over U and the randomness of the learning algorithm. Notice that $\hat{R}_{\text{loo}} \in [0,1]$. It can be shown that the loo-e-CMI $I(\dot{\mathbf{\Lambda}}; U|\dot{\boldsymbol{Z}})$ can be bounded by $H_b(\hat{R}_{\text{loo}}) + \hat{R}_{\text{loo}} \log(n+1)$, where $H_b(\hat{R}_{\text{loo}})$ denotes the binary entropy (*i.e.*, the entropy of a Bernoulli random variable with parameter $\hat{R}_{\text{loo}}$) (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) and Haghifam *et al.* (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) and Haghifam *et al.* (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 $\kappa(\boldsymbol{Z_S})$,

$$I(W; \boldsymbol{S}|\tilde{\boldsymbol{Z}}) \leq I(\kappa(\boldsymbol{Z_S}); \boldsymbol{S}|\tilde{\boldsymbol{Z}}) \leq H(\kappa(\boldsymbol{Z_S})|\tilde{\boldsymbol{Z}}) \leq k \log(2n). \tag{7.46}$$

Here, the last step follows since, given $\tilde{\boldsymbol{Z}}$, there are at most $\binom{2n}{k} \leq (2n)^k$ possible values of $\kappa(\boldsymbol{Z_S})$. 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

$$
\begin{aligned}
\overline{\mathrm{gen}}_\sigma = & \Bigg| \mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}_{\boldsymbol{S}}}P_{Z'}} \Bigg[\mathbb{E}_{P_{N_\sigma}} \Big[\ell_f(f^\sigma_{W|\boldsymbol{Z}_{\boldsymbol{S}},N}(X'), Y') \Big] \\
& \qquad - \frac{1}{n} \sum_{i\in[n]} \mathbb{E}_{P_N} \Big[\ell_f(f^\sigma_{W|\boldsymbol{Z}_{\boldsymbol{S}},N}(X_{i+S_in}), Y_{i+S_in}) \Big] \Bigg] \Bigg| \\
= & \Bigg| \mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}_{\boldsymbol{S}}}P_{Z'}} \Bigg[\ell_f(f_{W|\boldsymbol{Z}_{\boldsymbol{S}}}(X'), Y') + \mathbb{E}_{P_{N_\sigma}}[\Delta'] \qquad (7.50) \\
& \qquad - \frac{1}{n} \sum_{i\in[n]} \Big(\ell_f(f_{W|\boldsymbol{Z}_{\boldsymbol{S}}}(X_{i+S_in}), Y_{i+S_in}) + \mathbb{E}_{P_{N_\sigma}}[\Delta_i] \Big) \Bigg] \Bigg|,
\end{aligned}
$$

where

$$
\Delta' = \ell_f(f^\sigma_{W|\boldsymbol{Z}_{\boldsymbol{S}},N}(X'), Y') - \ell_f(f_{W|\boldsymbol{Z}_{\boldsymbol{S}}}(X'), Y'), \qquad (7.51)
$$

$$
\Delta_i = \ell_f(f^\sigma_{W|\boldsymbol{Z}_{\boldsymbol{S}},N}(X_{i+S_in}), Y_{i+S_in}) - \ell_f(f_{W|\boldsymbol{Z}_{\boldsymbol{S}}}(X_{i+S_in}), Y_{i+S_in}). \quad (7.52)
$$

Due to the Lipschitz assumption, we have $|\Delta'| \le \gamma \, \|{N_\sigma}'\|$, where ${N_\sigma}' \sim \mathcal{N}(0, \sigma^2 I_d)$. Similarly, $|\Delta_i| \le \gamma \, \|{N_\sigma}'\|$. Since $\mathbb{E}[\|{N_\sigma}'\|] \le 2\sigma\sqrt{d}$, we find that

$$
\overline{\mathrm{gen}}_\sigma \ge \overline{\mathrm{gen}} - 2\gamma\sigma\sqrt{d}. \qquad (7.53)
$$

We now need to bound $\overline{\mathrm{gen}}_\sigma$. Let $\mathbf{F}^\sigma$ denote the vector of predictions on $\tilde{\boldsymbol{X}}$ induced by $f^\sigma_{W|\boldsymbol{Z}_{\boldsymbol{S}},N}$. By the individual-sample f-CMI version of Theorem 6.12, we have

$$
\begin{aligned}
\overline{\mathrm{gen}}_\sigma &\le \frac{1}{n} \sum_{i\in[n]} \sqrt{2I(F_i^\sigma, F_{i+n}^\sigma; S_i | \tilde{\boldsymbol{Z}})} \qquad (7.54) \\
&\le \frac{1}{n} \sum_{i\in[n]} \sqrt{2I(F_i^\sigma, F_{i+n}^\sigma; S_i | \boldsymbol{S}_{-i}, \tilde{\boldsymbol{Z}})}, \qquad (7.55)
\end{aligned}
$$

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 $(F_i^\sigma, F_{i+n}^\sigma)$ 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 by Dwork *et al.* (2015) as

$$I_{\max}(W;\boldsymbol{Z}) = \operatorname*{ess\,sup}_{P_{W\boldsymbol{Z}}} \imath(W,\boldsymbol{Z}). \tag{7.60}$$

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). A discussion of generalization bounds for Gibbs posteriors regularized with arbitrary complexity measures can be found in the work of Viallard *et al.* (2024).

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 Zhou *et al.* (2021), Zhou *et al.* (2022), and Wu *et al.* (2022a). 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 Livni and Moran (2017), Bassily *et al.* (2018), 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 Theorem 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 *et al.* (2014), London (2017), 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 Steinke and Zakynthinou (2020), Oneto *et al.* (2020), Hellström and Durisi (2020a), Esposito *et al.* (2021a), and Rodríguez-Gálvez *et al.* (2021a).

8

Neural Networks and Iterative Algorithms

In this section, we apply the bounds from Sections 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 Sections 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 (Theorem 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 Theorems 8.1 and 8.2 and the mutual information generalization bound in Theorem 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

$$\mathbb{E}_{P_{W\boldsymbol{Z}}}[\mathrm{gen}(W, \boldsymbol{Z})] \le \sqrt{\frac{\sigma^2}{n}\sum_{t=1}^{T}\frac{\eta_t^2 L^2}{\rho_t^2}}. \tag{8.24}$$

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

$$\begin{aligned} V_0 &= 0\\ V_t &= V_{t-1} - \eta_t \nabla\ell(V_{t-1}, Z_{J_t}) + \xi_t, \qquad t = 1, \dots, T && (8.25)\\ W &= V_T \end{aligned}$$

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:

$$\sup_{z \in \mathcal{Z}} |\ell(w, z) - \ell(w', z)| \leq L\|w - w'\|. \tag{8.26}$$

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

$$\mathbb{E}_{P_{W\boldsymbol{Z}}}[\mathrm{gen}(W, \boldsymbol{Z})] \leq \sqrt{\frac{\beta\sigma^2 L^2}{n} \sum_{t=1}^{nk} \frac{1}{t}} \tag{8.27}$$

$$\leq \sqrt{\frac{\beta\sigma^2 L^2}{n}(\log n + \log k + 1)}. \tag{8.28}$$

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 constructed 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 using gradient-based algorithms, this activation function is often required to be differentiable almost everywhere. 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 as a vector. 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 approach 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 practical 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 suggests that any generalization guarantee that is uniform over all data distributions is doomed to be vacuous, as it 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 using 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 approach 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 Theorem 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 Section 6 can be viewed as an alternative path to data-dependent priors. This was exploited in Hellström and Durisi (2021a) and Hellström and Durisi (2021b), wherein an approach similar to that of Dziugaite and Roy (2017) and Dziugaite *et al.* (2021) 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 the samples 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, where noise is added to the parameters, and not to the underlying, deterministic ones 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, rather than 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. 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 Figure 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

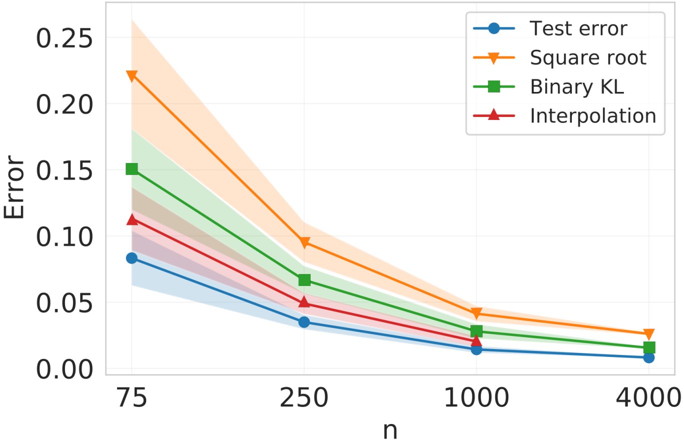

Figure 8.1: Numerical evaluation for a CNN trained on a binary version of MNIST (Hellström and Durisi, 2022a, Fig. 2(a)).

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 assigns greater weight to 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 increase 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, Mou *et al.* (2018), Li *et al.* (2020), Bu *et al.* (2020), Negrea *et al.* (2019), Haghifam *et al.* (2020), Wang *et al.* (2021b), Wang *et al.* (2021a), Wang *et al.* (2023), Issa *et al.* (2023), and Futami and Fujisawa (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 (Neyshabur *et al.*, 2015; Bartlett *et al.*, 2017). 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.*, Tsuzuku *et al.* (2020) and Foret *et al.* (2021). 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. Recently, Mitarchuk *et al.* (2024) derived generalization bounds for a class of recurrent NNs, while Mustafa *et al.* (2024) obtained nonvacuous bounds for the *adversarial* risk.

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 various information metrics for NNs, such as $I(W;\boldsymbol{Z})$. Huang *et al.* (2023), Clerico *et al.* (2023), and Clerico and Guedj (2024) 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 Neyshabur *et al.* (2017), Jiang *et al.* (2020), and Dziugaite *et al.* (2020).

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 section, 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 isolation: the learner has access to n training samples from the task, and this is all it has to rely on. 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 the processing of visual data, which may be similar across many different tasks.

This idea is captured by the framework of meta learning (Caruana, 1997; Thrun and Pratt, 1998; Baxter, 2000). 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_τ, 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_τ, 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_τ, 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

$$L = \mathbb{E}_{P_{\hat{\boldsymbol{Z}}} P_{U|\hat{\boldsymbol{Z}}} P_{\boldsymbol{Z}^\tau} P_{W|\boldsymbol{Z}^\tau U} P_{Z^\tau}}\left[\ell(W, Z^\tau)\right] = \mathbb{E}_{P_W P_{Z^\tau}}\left[\ell(W, Z^\tau)\right]. \tag{9.1}$$

While the meta learner does not have access to L, it can compute the meta-training loss, defined as

$$\hat{L} = \mathbb{E}_{P_{\hat{\boldsymbol{Z}}} P_{U|\hat{\boldsymbol{Z}}}} \left[\frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{P_{W_i|\hat{\boldsymbol{Z}}_{i,:},U}} \left[\frac{1}{n} \sum_{j=1}^{n} \ell(W_i, \hat{Z}_{i,j}) \right] \right]. \tag{9.2}$$

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 Section 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,

$$\left| L - \hat{L} \right| \leq \sqrt{\frac{2\sigma^2 I(U, \hat{W}; \hat{\boldsymbol{Z}})}{nm}}. \tag{9.3}$$

Proof. The proof follows the same approach as the proof of Theorem 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., we have

$$\frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{P_{W_i|U} P_{\hat{\boldsymbol{Z}}} P_U} \left[\frac{1}{n} \sum_{j=1}^{n} \ell(W_i, \hat{Z}_{i,j}) \right] = \mathbb{E}_{P_W P_{Z^\tau}} [\ell(W, Z^\tau)] = L. \tag{9.6}$$

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 using 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 into 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.*, Pentina and Lampert (2014), Amit and Meir (2018), Rothfuss *et al.* (2021), and Rezazadeh (2022). 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 described 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 is drawn. While this is a natural assumption to make from a theoretical standpoint, there are many situations where a distribution shift is expected when deploying a model. There are also scenarios where there is an abundance of data from a surrogate distribution, but 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, when 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 sometimes overloaded and used to refer to OOD generalization more broadly, or even to certain variations of meta learning.

For simplicity, we 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 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, for any $\lambda \in \mathbb{R}$, we have

$$D(P_Z^T \,||\, P_Z) \geq \mathbb{E}_{P_Z^T}\left[\lambda\, \mathbb{E}_{P_{W\boldsymbol{Z}}}\left[\ell(W,Z^T)\right]\right] - \log \mathbb{E}_{P_Z}\left[e^{\lambda\, \mathbb{E}_{P_{W\boldsymbol{Z}}}[\ell(W,Z)]}\right]. \tag{9.12}$$

Due to the sub-Gaussianity assumption, we have

$$\begin{aligned}
&\log \mathbb{E}_{P_Z}\left[e^{\lambda\, \mathbb{E}_{P_{W\boldsymbol{Z}}}[\ell(W,Z)]}\right] \\
&= \log \mathbb{E}_{P_Z}\left[e^{\lambda(\mathbb{E}_{P_{W\boldsymbol{Z}}}[\ell(W,Z)] - \mathbb{E}_{P_Z}[\mathbb{E}_{P_{W\boldsymbol{Z}}}[\ell(W,Z)]] + \mathbb{E}_{P_Z}[\mathbb{E}_{P_{W\boldsymbol{Z}}}[\ell(W,Z)]])}\right] && (9.13) \\
&\geq \lambda\, \mathbb{E}_{P_Z}[\mathbb{E}_{P_{W\boldsymbol{Z}}}[\ell(W,Z)]] + \frac{\lambda^2\sigma^2}{2}. && (9.14)
\end{aligned}$$

By combining these steps and optimizing over λ for the two cases $\lambda > 0$ and $\lambda < 0$, we obtain the final result. □

Theorem 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 distributions 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 roles 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 since we cannot minimize these discrepancy measures and do not have 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 $f_W : \mathcal{X} \to \mathcal{Y}$, with its loss depending 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

$$\hat{L} = \mathbb{E}_{P_{W\boldsymbol{Z}\boldsymbol{X}^T}}[L_{\boldsymbol{Z}}(W)], \qquad L^T = \mathbb{E}_{P_{W\boldsymbol{Z}\boldsymbol{X}^T}}\left[\mathbb{E}_{P_Z^T}\left[\ell(W,Z^T)\right]\right]. \tag{9.16}$$

Following Wang and Mao (2023a), we can derive bounds on L^T directly from $\hat{L}$, *i.e.*, without relying on 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 Theorem 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 relationship 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—*i.e.*, 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

$$L^T(W) = \mathbb{E}_{P_Z^T}[\ell(W,Z)] = \mathbb{E}_{P_{h_W}^T P_{Y|h_W}^T}\left[\ell(g_W(h_W(X)))\right], \tag{9.22}$$

$$L(W) = \mathbb{E}_{P_Z}[\ell(W,Z)] = \mathbb{E}_{P_{h_W} P_{Y|h_W}}\left[\ell(g_W(h_W(X)))\right]. \tag{9.23}$$

Therefore, by repeating the argument of Theorem 9.2 with this reformulation at the start, we obtain

$$\left|L^T - L\right| \le \mathbb{E}_{P_W}\left[\sqrt{2\sigma^2 D(P^T_{h_W} P^T_{Y|h_W} \,||\, P_{h_W} P_{Y|h_W})}\right]. \tag{9.24}$$

The result in Theorem 9.4 can be adapted similarly. Next, note that the relative entropy can be decomposed as

$$D(P^T_{h_W} P^T_{Y|h_W} || P_{h_W} P_{Y|h_W}) = D(P^T_{h_W} || P_{h_W}) + D(P^T_{Y|h_W} || P_{Y|h_W}). \tag{9.25}$$

Consequently, we have two components of the discrepancy measure: the representation discrepancy $D(P^T_{h_W} || P_{h_W})$ and the conditional discrepancy $D(P^T_{Y|h_W} || P_{Y|h_W})$. 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, then for any fixed W we have

$$L^T(W) - L(W) \le \sqrt{2\sigma^2 D(P^T_X \,||\, P_X)} + \min_{w^* \in \mathcal{W}} \{L^T(w^*) + L(w^*)\}. \tag{9.26}$$

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.*,

$$|\ell(w, Z)|_\infty = \inf\{s : P_Z^T(\ell(w, Z) > s) = 0\} \le \sigma. \tag{9.29}$$

Then, we have

$$\begin{aligned}\left|L^T - \hat{L}\right| \le \frac{2\alpha\sigma}{m}\sum_{i\in[m]}\mathrm{TV}(P_{WZ_i}, P_W P_{Z_i^T}) \\ + \frac{2(1-\alpha)\sigma}{n}\sum_{i\in[n]}\Big(\mathrm{TV}(P_{WZ_i}, P_W P_{Z_i}) + \mathrm{TV}(P_Z, P_Z^T)\Big).\end{aligned} \tag{9.30}$$

Proof. First, we decompose the generalization gap as

$$\begin{aligned}\left|L^T - \hat{L}\right| &= \left|L^T - \frac{\alpha}{m}\sum_{i=1}^{m} \mathbb{E}_{P_{WZ_i^T}}\left[\ell(W, Z_i^T)\right] - \frac{1-\alpha}{n}\sum_{i=1}^{n} \mathbb{E}_{P_{WZ_i}}[\ell(W, Z_i)]\right| \\ &\leq \frac{\alpha}{m}\sum_{i=1}^{m}\left|\mathbb{E}_{P_W P_{Z_i^T}}\left[\ell(W, Z^T)\right] - \mathbb{E}_{P_{WZ_i^T}}\left[\ell(W, Z_i^T)\right]\right| \qquad (9.31) \\ &\quad + \frac{1-\alpha}{n}\sum_{i=1}^{n}\left|\mathbb{E}_{P_W P_{Z_i^T}}\left[\ell(W, Z^T)\right] - \mathbb{E}_{P_{WZ_i}}[\ell(W, Z_i)]\right|.\end{aligned}$$

The terms in the first sum are individual-sample generalization gaps. By applying Theorem 4.4 to each term, we can bound them as

$$\mathbb{E}_{P_W P_{Z_i^T}}\left[\ell(W, Z^T)\right] - \mathbb{E}_{P_{WZ_i^T}}\left[\ell(W, Z_i^T)\right] \leq \mathrm{TV}(P_{WZ_i^T}, P_W P_{Z_i^T}). \tag{9.32}$$

Proceeding similarly with the second sum, we can bound each term as

$$\mathbb{E}_{P_W P_{Z_i^T}}\left[\ell(W, Z^T)\right] - \mathbb{E}_{P_{WZ_i}}[\ell(W, Z_i)] \leq \mathrm{TV}(P_{WZ_i}, P_W P_{Z_i^T}). \tag{9.33}$$

To isolate the effect of the distribution shift, we can decompose this last upper bound as

$$\begin{aligned}\mathrm{TV}(P_{WZ_i}, P_W P_{Z_i^T}) &= \frac{1}{2}\int_{\mathcal{W}\times\mathcal{Z}}\left|\mathrm{d}P_{WZ_i} - \mathrm{d}P_W P_{Z_i^T}\right| && (9.34) \\ &\leq \frac{1}{2}\int_{\mathcal{W}\times\mathcal{Z}}\left|\mathrm{d}P_{WZ_i} - \mathrm{d}P_W P_{Z_i}\right| && (9.35) \\ &\quad + \frac{1}{2}\int_{\mathcal{W}\times\mathcal{Z}}\left|\mathrm{d}P_W P_{Z_i} - \mathrm{d}P_W P_{Z_i^T}\right| \\ &= \mathrm{TV}(P_{WZ_i}, P_W P_{Z_i}) + \mathrm{TV}(P_Z, P_Z^T). && (9.36)\end{aligned}$$

The desired result is obtained 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 Wu *et al.* (2022a), Aminian *et al.* (2022a), and Wang and Mao (2023a).

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, each with access to its 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, 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)

$$\begin{aligned}\mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{P_Z}(W)] &= \frac{1}{nK}\sum_{k,i} \mathbb{E}_{P_{W\boldsymbol{Z}}P_{\boldsymbol{Z}'}}\left[\ell(W, Z'_{k,i})\right] \qquad (9.41)\\ &= \frac{1}{nK}\sum_{k,i} \mathbb{E}_{P_{W\boldsymbol{Z}}P_{\boldsymbol{Z}'}}[f(W) - f(Z'_{k,i}) - \langle \nabla f(Z'_{k,i}), W - Z'_{k,i}],\end{aligned}$$

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,

$$\begin{aligned}\mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{\boldsymbol{Z}}(W)] &= \frac{1}{nK}\sum_{k,i} \mathbb{E}_{P_{W^{k,i}\boldsymbol{Z}\boldsymbol{Z}'}}\left[\ell(W^{k,i}, Z'_{k,i})\right] \qquad (9.42)\\ &= \frac{1}{nK}\sum_{k,i} \mathbb{E}_{P_{W^{k,i}\boldsymbol{Z}\boldsymbol{Z}'}}\Big[f(W^{k,i}) - f(Z'_{k,i}) \qquad (9.43)\\ &\qquad - \langle \nabla f(Z'_{k,i}), W^{k,i} - Z'_{k,i}\Big],\end{aligned}$$

since $Z'_{k,i}$ is in the training set of $W^{k,i}$. It follows that

$$\begin{aligned}&\mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{P_Z}(W) - L_{\boldsymbol{Z}}(W)]\\ &\qquad = \frac{1}{nK}\sum_{k,i} \mathbb{E}_{P_{WW^{k,i}\boldsymbol{Z}\boldsymbol{Z}'}}\left[\langle \nabla f(Z'_{k,i}), W^{k,i} - W\rangle\right]. \qquad (9.44)\end{aligned}$$

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,

$$\begin{aligned}\mathbb{E}_{P_{W\boldsymbol{Z}}}&[L_{P_Z}(W) - L_{\boldsymbol{Z}}(W)]\\ &= \frac{1}{nK^2}\sum_{k,i}\mathbb{E}_{P_{WW^{k,i}\boldsymbol{Z}\boldsymbol{Z}'}}\left[\langle \nabla f(Z'_{k,i}), W_k^i - W_k\rangle\right].\end{aligned} \tag{9.45}$$

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 Theorem 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 Theorem 4.2 in the proof. For further discussion and extensions of these bounds, see for instance the work of Yagli *et al.* (2020) and Barnes *et al.* (2022).

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, allowing us to capture interactions 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.[14] 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 $\bar{M}_n(W) \le n$). Hence, the result in (9.47) holds trivially in this case. Thus, we only consider $D(P_{W|\boldsymbol{Z}} \,||\, Q_W) \le 2n$. Specifically, assume that $D(P_{W|\boldsymbol{Z}} \,||\, Q_W) \in [k-1, k]$ for $k \in [2n]$. Then, by (9.51), we have

$$\left| \mathbb{E}_{P_{W|\boldsymbol{Z}}} \left[\frac{\bar{M}_n(W)}{n} \right] \right| \le \frac{k + \log \frac{2}{\delta}}{\lambda} + \frac{\lambda}{8n}. \tag{9.52}$$

For a fixed k, this is minimized by $\lambda = 2\sqrt{2n(k + \log \frac{2}{\delta})}$, which gives

$$\left| \mathbb{E}_{P_{W|\boldsymbol{Z}}} \left[\frac{\bar{M}_n(W)}{n} \right] \right| \le \sqrt{\frac{k + \log \frac{2}{\delta}}{2n}}. \tag{9.53}$$

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 $k \le D(P_{W|\boldsymbol{Z}} \,||\, Q_W) + 1$,

$$\left| \mathbb{E}_{P_{W|\boldsymbol{Z}}} \left[\frac{\bar{M}_n(W)}{n} \right] \right| \le \sqrt{\frac{D(P_{W|\boldsymbol{Z}} \,||\, Q_W) + 1 + \log \frac{4n}{\delta}}{2n}} \tag{9.54}$$

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 the applicability of PAC-Bayesian bounds beyond i.i.d. data. For instance, Seldin *et al.* (2011) and Seldin *et al.* (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.

[14] 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.

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 $H_{\leq t} = (H_1, \ldots, H_{t-1})$, where $H_t = (S_t, A_t, R_t)$. Specifically, the policy should be chosen to obtain a high cumulative expected reward $r_c(\varphi)$, defined as

$$r_c(\varphi) = \mathbb{E}\left[\sum_{t\in[T]} r(Y_t, \varphi_t(S_t, H_{\leq t}))\right]. \tag{9.55}$$

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

$$\mathrm{MBR} = R_B^o - R_c. \tag{9.57}$$

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 (Thompson, 1933; Russo and Van Roy, 2016). 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 Section 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.*, Pentina and Lampert (2014), Amit and Meir (2018), Rothfuss *et al.* (2021), Liu *et al.*

(2021b), Farid and Majumdar (2021), Meunier and Alquier (2021), Flynn *et al.* (2022), Rezazadeh (2022), Jose *et al.* (2022a), and Riou *et al.* (2023). The bounds for OOD generalization in Theorems 9.2 to 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 Germain *et al.* (2016b), Achille *et al.* (2021), Jose and Simeone (2021c), Aminian *et al.* (2022b), and Bu *et al.* (2022). 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.*, Seldin and Tishby (2010), Higgs and Shawe-Taylor (2010), and Li *et al.* (2018). 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) and Lugosi and Neu (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. Caro *et al.* (2024) consider the quantum learning setting, and derive generalization bounds in terms of information-theoretic quantities. 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, while Zecchin *et al.* (2024) use information-theoretic metrics to characterize the expected size of conformal prediction sets.

Full text available at: http://dx.doi.org/10.1561/2200000112

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) and Haghifam *et al.* (2023), there are certain settings where 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 in 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 Theorem 5.2? Can the $\log \sqrt{n}$ dependence in Theorem 5.4 be removed? Another question is whether the most suitable information measure 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, several interesting extensions to other settings and connections to other approaches can be explored. While we covered some topics in Section 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 provide 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. The authors are grateful to the team at *Foundations and Trends® in Machine Learning*, in particular Mike Casey and Lucy Wiseman, for their diligent handling of this manuscript.

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.

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.

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.

Alabdulmohsin, I. (2020). “Towards a Unified Theory of Learning and Information”. *Entropy*. 22(4). DOI: 10.3390/e22040438.

Alquier, P. (2006). “Transductive and inductive adaptative inference for regression and density estimation”. *PhD thesis*. University of Paris.

Alquier, P. (2008). “PAC-Bayesian bounds for randomized empirical risk minimizers”. *Mathematical Methods of Statistics*. 17(4): 279–304. DOI: 10.3103/S1066530708040017.

Alquier, P. (2024). “User-friendly Introduction to PAC-Bayes Bounds”. *Foundations and Trends® in Machine Learning*. 17(2): 174–303. DOI: 10.1561/2200000100.

Alquier, P. and Biau, G. (2013). “Sparse single-index model”. *Journal of Machine Learning Research (JMLR)*. 14(1): 243–280.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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).

Audibert, J.-Y. (2004). "A better variance control for PAC-Bayesian classification". URL: certis.enpc.fr/~audibert/Mes%20articles/PhDthesis.pdf.

Audibert, J.-Y. and Bousquet, O. (2007). "Combining PAC-Bayesian and Generic Chaining Bounds". *Journal of Machine Learning Research (JMLR)*. 8(32): 863–889.

Banerjee, A. (2006). "On Bayesian Bounds". In: *Proc. Int. Conf. Mach. Learning (ICML)*. Pittsburgh, PE, USA. DOI: 10.1145/1143844.1143855.

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.

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.

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.

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.

Barron, A. and Cover, T. (1991). “Minimum complexity density estimation”. *IEEE Trans. Info. Theory.* 37(4): 1034–1054. DOI: 10.1109/18.86996.

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.

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.

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. DOI: 10.1007/3-540-44581-1_15.

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.

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.

Baxter, J. (2000). “A Model of Inductive Bias Learning”. *J. Artif. Int. Res.* 12(1): 149–198. DOI: 10.1613/jair.731.

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.

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.

Bernstein, J. and Yue, Y. (2021). “Computing the Information Content of Trained Neural Networks”. In: *Workshop on the Theory of Overparameterized Machine Learning.*

Biggs, F. and Guedj, B. (2021). "Differentiable PAC–Bayes Objectives with Partially Aggregated Neural Networks". *Entropy*. 23(10). DOI: 10.3390/e23101280.

Biggs, F. and Guedj, B. (2022a). "Non-Vacuous Generalisation Bounds for Shallow Neural Networks". In: *Proc. Int. Conf. Mach. Learn. (ICML)*. Baltimore, MD.

Biggs, F. and Guedj, B. (2022b). "On Margins and Derandomisation in PAC-Bayes". In: *Proc. Artif. Intell. Statist. (AISTATS)*. Virtual Conference.

Biggs, F. and Guedj, B. (2023). "Tighter PAC-Bayes Generalisation Bounds by Leveraging Example Difficulty". In: *Proc. Artif. Intell. Statist. (AISTATS)*. Valencia, Spain.

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.

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.

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.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities. A nonasymptotic theory of independence*. Oxford, United Kingdom: Oxford University Press.

Bousquet, O. and Elisseeff, A. (2002). "Stability and Generalization". *Journal of Machine Learning Research (JMLR)*. 2(Mar.): 499–526.

Bretagnolle, J. and Huber, C. (1978). "Estimation des densités : risque minimax". fre. *Séminaire de probabilités de Strasbourg*. 12: 342–363.

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.

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.

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.

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.

Canonne, C. L. (2022). "A short note on an inequality between KL and TV". *arXiv*. DOI: 10.48550/arxiv.2202.07198.

Caro, M., Gur, T., Rouzé, C., França, D. S., and Subramanian, S. (2024). "Information-theoretic generalization bounds for learning from quantum data". In: *Proc. Conf. Learn. Theory (COLT)*. Edmonton, Canada.

Caruana, R. (1997). "Multitask Learning". *Mach. Learn.* 28(1): 41–75. DOI: 10.1007/978-1-4615-5529-2_5.

Catoni, O. (2007). *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. Vol. 56. IMS Lecture Notes Monogr. Ser. 1–163.

Catoni, O. (2004a). "A PAC-Bayesian approach to adaptive classification". URL: yaroslavvb.com/papers/notes/catoni-pac.pdf.

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.

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.*

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.

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.

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.

Clerico, E., Deligiannidis, G., and Doucet, A. (2022a). "Conditionally Gaussian PAC-Bayes". In: *Proc. Artif. Intell. Statist. (AISTATS)*. Virtual conference.

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.

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.

Clerico, E., Shidani, A., Deligiannidis, G., and Doucet, A. (2022b). "Chained generalisation bounds". In: *Proc. Conf. Learn. Theory (COLT)*. Boulder, CO, USA.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. USA: Wiley-Interscience.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

Edgeworth, F. Y. (1908). "On the Probable Errors of Frequency-Constants". *Journal of the Royal Statistical Society*. 71(2): 381–397.

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.

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.

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.

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.

Fard, M. M. and Pineau, J. (2010). "PAC-Bayesian Model Selection for Reinforcement Learning". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Vancouver, Canada.

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.

Feldman, V. and Steinke, T. (2018). "Calibrating Noise to Variance in Adaptive Data Analysis". In: *Proc. Conf. Learning Theory (COLT)*. Stockholm, Sweden.

Fenchel, W. (1949). "On Conjugate Convex Functions". *Canadian Journal of Mathematics*. 1(1): 73–77. DOI: 10.1007/978-3-0348-0439-4_7.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

Grünwald, P. (2007). *The minimum description length principle*. MIT press.

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.

Grünwald, P., Pérez-Ortiz, M. F., and Mhammedi, Z. (2023). "Exponential Stochastic Inequality". *arXiv*. May. DOI: 10.48550/arxiv.2304.14217.

Guedj, B. (2019). "A primer on PAC-Bayesian learning". *Proc. 2nd Congress Société Mathématique de France*: 391–414. DOI: 10.48550/arxiv.1901.05353.

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.

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.

Haddouche, M. and Guedj, B. (2022). "Online PAC-Bayes Learning". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. New Orleans, LA, USA.

Haddouche, M. and Guedj, B. (2023a). "PAC-Bayes Generalisation Bounds for Heavy-Tailed Losses through Supermartingales". *Transactions on Machine Learning Research (TMLR)*. Apr.

Haddouche, M. and Guedj, B. (2023b). "Wasserstein PAC-Bayes Learning: Exploiting Optimisation Guarantees to Explain Generalisation". *arXiv*. DOI: 10.48550/arXiv.2304.07048.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

Hellström, F. and Guedj, B. (2024). "Comparing Comparators in Generalization Bounds". In: *Proc. Artif. Intell. Statist. (AISTATS)*. Valencia, Spain.

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.

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.

Holland, M. (2019). "PAC-Bayes under potentially heavy tails". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Vancouver, Canada.

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.

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.

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.

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.

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.

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.

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.

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.

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.

Jose, S. T. and Simeone, O. (2021a). “Information-Theoretic Generalization Bounds for Meta-Learning and Applications”. *Entropy*. 23(1). DOI: 10.3390/e23010126.

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.

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.

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.

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.

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.

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. DOI: 10.1561/9781680837896.

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.

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.

Kolmogorov, A. N. (1963). "On tables of random numbers". *Sankhyā (Statistics). The Indian Journal of Statistics. Series A*. 25: 369–376.

Koltchinskii, V. (2001). "Rademacher penalties and structural risk minimization". *IEEE Trans. Inf. Theory*. 47(5): 1902–1914. DOI: 10.1109/18.930926.

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.

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.

Kontorovich, A. and Ramanan, K. (2008). "Concentration Inequalities for Dependent Random Variables via the Martingale Method". *The Annals of Probability*. 36(6): 2126–2158.

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.

Kouw, W. M. and Loog, M. (2019). "An introduction to domain adaptation and transfer learning". *arXiv*. Dec. DOI: 10.48550/arxiv.1812.11806.

Kutin, S. and Niyogi, P. (2002). "Almost-Everywhere Algorithmic Stability and Generalization Error". In: *Proc. Conf. Uncertainty in Artif. Intell. (UAI)*. Edmonton, Canada.

Kuzborskij, I., Jun, K.-S., Wu, Y., Jang, K., and Orabona, F. (2024). "Better-than-KL PAC-Bayes Bounds". In: *Proc. Conf. Learn. Theory (COLT)*. Edmonton, Canada.

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.

Langford, J. (2002). "Quantitatively Tight Sample Complexity Bounds". *PhD thesis*. Carnegie Mellon University.

Langford, J. and Caruana, R. (2001). "(Not) Bounding the True Error". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Vancouver, Canada.

Langford, J. and Seeger, M. (2001). "Bounds for Averaging Classifiers". *CMU Technical report*. CMU-CS-01-102.

Langford, J. and Shawe-Taylor, J. (2002). "PAC-Bayes & margins". In: *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*. Vancouver, Canada.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

Li, Q. J. (1999). "Estimation of Mixture Models". *PhD thesis*. Yale University.

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.

Littlestone, N. and Warmuth, M. K. (2003). "Relating Data Compression and Learnability". *Technical Report*.

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.

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.

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.

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.

London, B., Huang, B., Taskar, B., and Getoor, L. (2014). “PAC-Bayesian Collective Stability”. In: *Proc. Artif. Intell. Statist. (AISTATS)*. Reykjavik, Iceland.

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.

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.

Lugosi, G. and Neu, G. (2022). “Generalization Bounds via Convex Analysis”. In: *Proc. Conf. Learn. Theory (COLT)*. London, United Kingdom.

Lugosi, G. and Neu, G. (2023). “Online-to-PAC Conversions: Generalization Bounds via Regret Analysis”. *arXiv*. May. DOI: 10.48550/arxiv.2305.19674.

Marton, K. (1996). “A Measure Concentration Inequality for Contracting Markov Chains”. *Geometric and functional analysis*. 6(3): 556–571.

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*.

Maurer, A. (2004). “A Note on the PAC Bayesian Theorem”. *arXiv*. Nov. DOI: 10.48550/arxiv.cs/0411099.

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.

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.

McAllester, D. A. (1998). “Some PAC-Bayesian Theorems”. In: *Proc. Conf. Learn. Theory (COLT)*. Madison, WI, USA.

McAllester, D. A. (1999). “PAC-Bayesian Model Averaging”. In: *Proc. Conf. Comp. Learn. Theory (COLT)*. Santa Cruz, CA, USA. DOI: 10.1145/307400.307435.

McAllester, D. A. (2003a). "PAC-Bayesian Stochastic Model Selection". *Mach. Learn.* 51(Apr.): 5–21.

McAllester, D. A. (2003b). "Simplified PAC-Bayesian margin bounds". In: *Proc. Conf. Comp. Learn. Theory (COLT)*. Santa Cruz, CA, USA.

McAllester, D. A. (2013). "A PAC-Bayesian Tutorial with a Dropout Bound". *arXiv*. July. DOI: 10.48550/arxiv.1307.2118.

Mendelson, S. (2014). "Learning without concentration". In: *Proc. Conf. Learn. Theory (COLT)*. Barcelona, Spain. DOI: 10.1145/2699439.

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.

Meunier, D. and Alquier, P. (2021). "Meta-Strategy for Learning Tuning Parameters with Guarantees". *Entropy*. 23(10). DOI: 10.3390/e23101257.

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.

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.

Mitarchuk, V., Lacroce, C., Eyraud, R., Emonet, R., Habrard, A., and Rabusseau, G. (2024). "Length independent PAC-Bayes bounds for Simple RNNs". In: *Proc. Artif. Intell. Statist. (AISTATS)*. Valencia, Spain.

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.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of Machine Learning*. 2nd ed. *Adaptive Computation and Machine Learning*. Cambridge, MA: MIT Press.

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.

Murphy, K. P. (2022). *Probabilistic Machine Learning: An introduction.* MIT Press. URL: probml.ai.

Mustafa, W., Liznerski, P., Ledent, A., Wagner, D., Wang, P., and Kloft, M. (2024). "Non-vacuous Generalization Bounds for Adversarial Risk in Stochastic Neural Networks". In: *Proc. Artif. Intell. Statist. (AISTATS).* Valencia, Spain.

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.

Nachum, I. and Yehudayoff, A. (2019). "Average-Case Information Complexity of Learning". In: *Proc. Conf. Alg. Learn. Theory (ALT).* Chicago, IL, USA.

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.

Neal, R. M. (1994). "Bayesian Learning for Neural Networks". *PhD thesis.* University of Toronto. DOI: 10.1007/978-1-4612-0745-0.

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.

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.

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.

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.

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.

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.

Neyshabur, B., Tomioka, R., and Srebro, N. (2015). “Norm-Based Capacity Control in Neural Networks”. In: *Proc. Conf. Learn. Theory (COLT)*. Paris, France.

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.

Nozawa, K., Germain, P., and Guedj, B. (2020). “PAC-Bayesian Contrastive Unsupervised Representation Learning”. In: *Proc. Conf. Uncertainty in Artif. Intell. (UAI)*.

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.

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.

Orabona, F. (2023). “A modern introduction to online learning”. *arXiv*. May. DOI: 10.48550/arxiv.1912.13213.

Palomar, D. P. and Verdu, S. (2008). “Lautum Information”. *IEEE Trans. Inf. Theory*. 54(3): 964–975. DOI: 10.1109/TIT.2007.915715.

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.

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.

Pentina, A. and Lampert, C. (2014). "A PAC-Bayesian bound for Lifelong Learning". In: *Proc. Int. Conf. Mach. Learning (ICML)*. Beijing, China.

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.

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.

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.

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. DOI: 10.1007/978-1-4757-3150-7_15.

Pitas, K. (2020). "Dissecting Non-Vacuous Generalization Bounds Based on the Mean-Field Approximation". In: *Proc. Int. Conf. Mach. Learn. (ICML)*. Virtual Conference.

Polyanskiy, Y. and Wu, Y. (2022). *Lecture Notes On Information Theory*. Cambridge, U.K.: Cambridge Univ. Press.

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.

Raginsky, M. (2019). "Information, Concentration, and Learning". In: *North American Summer School on Information Theory (NASIT)*. Boston, MA, USA.

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.

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.

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.

Rakhlin, A., Mukherjee, S., and Poggio, T. (2005). "Stability results in learning theory". *Analysis and Applications*. 14(Oct.): 397–417. DOI: 10.1142/S0219530505000650.

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.

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.

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.

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.

Rezazadeh, A. (2022). "A Unified View on PAC-Bayes Bounds for Meta-Learning". In: *Proc. Int. Conf. Mach. Learning (ICML)*. Baltimore, MD, USA.

Rigollet, P. and Tsybakov, A. B. (2012). "Sparse Estimation by Exponential Weighting". *Statistical Science*. 27(4): 558–575. DOI: 10.1214/12-STS393.

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.

Rissanen, J. (1978). "Modeling by shortest data description". *Automatica*. 14(5): 465–471. DOI: https://doi.org/10.1016/0005-1098(78)90005-5.

Rissanen, J. (1983). "A Universal Prior for Integers and Estimation by Minimum Description Length". *The Annals of Statistics*. 11(2): 416–431.

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.

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.

Rivasplata, O., Tankasali, V. M., and Szepesvari, C. (2019). "PAC-Bayes with Backprop". *arXiv*. Oct. DOI: 10.48550/arxiv.1908.07380.

Rockafellar, R. T. (1970). *Convex analysis. Princeton Mathematical Series*. Princeton, N. J., USA: Princeton University Press. DOI: 10.1515/9781400873173.

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.

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.

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.

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.

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.

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.

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.

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.

Rudin, W. (1987). *Real and Complex Analysis, 3rd Ed.* USA: McGraw-Hill, Inc.

Russo, D. and Zou, J. (2016). "Controlling bias in adaptive data analysis using information theory". In: *Proc. Artif. Intell. Statist. (AISTATS)*. Cadiz, Spain.

Russo, D. and Van Roy, B. (2016). "An Information-Theoretic Analysis of Thompson Sampling". *Journal of Machine Learning Research (JMLR)*. 17(1): 2442–2471.

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.

Salmon, J. and Dalalyan, A. (2011). "Optimal aggregation of affine estimators". In: *Proc. Conf. Learn. Theory (COLT)*. Budapest, Hungary.

Samson, P.-M. (2000). "Concentration of Measure Inequalities for Markov Chains and Φ-Mixing Processes". *The Annals of Probability*. 28(1): 416–461.

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).*

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.

Schwarz, G. (1978). "Estimating the Dimension of a Model". *The Annals of Statistics.* 6(2): 461–464.

Seeger, M. (2002). "PAC-Bayesian Generalisation Error Bounds for Gaussian Process Classification". *Journal of Machine Learning Research (JMLR).* 3(Oct.): 233–269.

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.

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.

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.

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.

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.

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.

Seldin, Y. and Tishby, N. (2010). "PAC-Bayesian Analysis of Co-clustering and Beyond". *Journal of Machine Learning Research (JMLR)*. 11(117): 3595–3646.

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.

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.

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.

Shannon, C. E. (1948). "A Mathematical Theory of Communication". *The Bell System Technical Journal*. 27: 379–423. DOI: 10.1063/1.3067010.

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.

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.

Shawe-Taylor, J. and Hardoon, D. (2009). "PAC-Bayes Analysis Of Maximum Entropy Classification". In: *Proc. Artif. Intell. Statist. (AISTATS)*. Clearwater Beach, FL, USA.

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.

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.

Shwartz-Ziv, R. and Tishby, N. (2017). "Opening the Black Box of Deep Neural Networks via Information". *arXiv*. DOI: 10.48550/arxiv.1703.00810.

Solomonoff, R. (1964). "A formal theory of inductive inference. Part I". *Information and Control.* 7(1): 1–22. DOI: https://doi.org/10.1016/S0019-9958(64)90223-2.

Steinke, T. and Zakynthinou, L. (2020). "Reasoning About Generalization via Conditional Mutual Information". In: *Proc. Conf. Learn. Theory (COLT).* Graz, Austria.

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.

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.

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.

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.

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.

Tishby, N., Pereira, F. C., and Bialek, W. (1999). "The information bottleneck method". In: *Allerton Conf. Communication, Control, Computing (Allerton).* Monticello, IL, USA.

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.

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.

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.

Valiant, L. G. (1984). "A Theory of the Learnable". *Commun. ACM*. 27(11): 1134–1142. DOI: 10.1145/1968.1972.

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.

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.

Van Handel, R. (2016). *Probability in High Dimension*. URL: web.math.princeton.edu/%7EErvan/APC550.pdf.

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.

Vapnik, V. and Chervonenkis, A. (1974). *Theory of Pattern Recognition [in Russian]*. Moscow: Nauka.

Verdú, S. (2015). "α-Mutual Information". In: *Proc. Inf. Theory Appl. Workshop (ITA)*. San Diego, CA, USA.

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.

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.

Viallard, P., Emonet, R., Habrard, A., Morvant, E., and Zantedeschi, V. (2024). "Leveraging PAC-Bayes Theory and Gibbs Distributions for Generalization Bounds with Complexity Measures". In: *Proc. Artif. Intell. Statist. (AISTATS)*. Valencia, Spain.

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.

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.

Villani, C. (2008). *Optimal transport – Old and new*. Vol. 338. *Grundlehren der mathematischen Wissenschaften*. Springer Science & Business Media.

Wainwright, M. J. (2019). *High-Dimensional Statistics: a Non-Asymptotic Viewpoint*. Cambridge, U.K.: Cambridge Univ. Press. DOI: 10.1017/9781108627771.

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.

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.

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.

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.

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.

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.

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.

Wang, Z. and Mao, Y. (2023a). “Information-Theoretic Analysis of Unsupervised Domain Adaptation”. In: *Proc. Int. Conf. Learn. Representations (ICLR)*. Kigali, Rwanda.

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.

Wang, Z. and Mao, Y. (2023c). “Tighter Information-Theoretic Generalization Bounds from Supersamples”. In: *Proc. Int. Conf. Mach. Learning (ICML)*. Honolulu, HI, USA.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

Zecchin, M., Park, S., Simeone, O., and Hellström, F. (2024). “Generalization and Informativeness of Conformal Prediction”. In: *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*. Athens, Greece.

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.

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.

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.

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.

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.

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.

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.