THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Information-Theoretic Generalization Bounds: Tightness and Expressiveness

Fredrik Hellström

Communication Systems Group
Department of Electrical Engineering
Chalmers University of Technology
Gothenburg, Sweden, 2022

Information-Theoretic Generalization Bounds:
Tightness and Expressiveness

Fredrik Hellström

ISBN 978-91-7905-782-4
Series No. 5248 in Doktorsavhandlingar vid Chalmers tekniska högskola
ISSN 0346-718X

This thesis has been prepared using LaTeX and Tikz.

Communication Systems Group
Department of Electrical Engineering
Chalmers University of Technology
SE-412 96 Gothenburg, Sweden
Phone: +46 (0)31 772 80 62
`www.chalmers.se`

Front cover illustration:
Illustration of the conditional mutual information framework for meta learning.
The complete figure is provided in Part I, Section 5.2.

Printed by Chalmers Reproservice
Gothenburg, Sweden, December 2022

Abstract

Machine learning has achieved impressive feats in numerous domains, largely driven by the emergence of deep neural networks. Due to the high complexity of these models, classical bounds on the generalization error—that is, the difference between training and test performance—fail to explain this success. This discrepancy between theory and practice motivates the search for new generalization guarantees, which must rely on other properties than function complexity. Information-theoretic bounds, which are intimately related to probably approximately correct (PAC)-Bayesian analysis, naturally incorporate a dependence on the relevant data distributions and learning algorithms. Hence, they are a promising candidate for studying generalization in deep neural networks.

In this thesis, we derive and evaluate several such information-theoretic generalization bounds. First, we derive both average and high-probability bounds in a unified way, obtaining new results and recovering several bounds from the literature. We also develop new bounds by using tools from binary hypothesis testing. We extend these results to the conditional mutual information (CMI) framework, leading to results that depend on quantities such as the conditional information density and maximal leakage.

While the aforementioned bounds achieve a so-called slow rate with respect to the number of training samples, we extend our techniques to obtain bounds with a fast rate. Furthermore, we show that the CMI framework can be viewed as a way of automatically obtaining data-dependent priors, an important technique for obtaining numerically tight PAC-Bayesian bounds. A numerical evaluation of these bounds demonstrate that they are nonvacuous for deep neural networks, but diverge as training progresses.

To obtain numerically tighter results, we strengthen our bounds through the use of the samplewise evaluated CMI, which depends on the information captured by the losses of the neural network rather than its weights. Furthermore, we make use of convex comparator functions, such as the binary relative entropy, to obtain tighter characterizations for low training losses. Numerically, we find that these bounds are nearly tight for several deep neural network settings, and remain stable throughout training. We demonstrate the expressiveness of the evaluated CMI framework by using it to rederive nearly optimal guarantees for multiclass classification, known from classical learning theory.

Finally, we study the expressiveness of the evaluated CMI framework for meta learning, where data from several related tasks is used to improve performance on new tasks from the same task environment. Through the use of a one-step derivation and the evaluated CMI, we obtain new information-theoretic generalization bounds for meta learning that improve upon previous results. Under certain assumptions on the function classes used by the learning algorithm, we obtain convergence rates that match known classical results. By extending our analysis to oracle algorithms and considering a notion of task diversity, we obtain excess risk bounds for empirical risk minimizers.

Keywords: Machine learning, statistical learning, generalization, information theory, PAC-Bayes, neural networks, meta learning.

List of Publications

This thesis is based on the following publications:

[A] **F. Hellström**, G. Durisi, "Generalization Bounds via Information Density and Conditional Information Density," published in *IEEE Journal on Selected Areas of Information Theory*, Nov. 2020.

[B.1] **F. Hellström**, G. Durisi, "Fast-Rate Loss Bounds via Conditional Information Measures with Applications to Neural Networks," presented at *IEEE International Symposium for Information Theory*, July 2021.

[B.2] **F. Hellström**, G. Durisi, "Data-Dependent PAC-Bayesian Bounds in the Random-Subset Setting with Applications to Neural Networks," *International Conference on Machine Learning: Workshop on Information-Theoretic Methods for Rigorous, Responsible, and Reliable Machine Learning*, July 2021.

[C] **F. Hellström**, G. Durisi, "A New Family of Generalization Bounds Using Samplewise Evaluated CMI," presented at *Conference on Neural Information Processing Systems*, Nov. 2022.

[D] **F. Hellström**, G. Durisi, "Evaluated CMI Bounds for Meta Learning: Tightness and Expressiveness," presented at *Conference on Neural Information Processing Systems*, Nov. 2022.

Publications by the author not included in the thesis:

[E] R. Catena, **F. Hellström**, "New Constraints on Inelastic Dark Matter from IceCube," *Journal of Cosmology and Astroparticle Physics*, Oct. 2018.

[F] **F. Hellström**, G. Durisi, "Generalization Error Bounds via mth Central Moments of the Information Density," *IEEE International Symposium on Information Theory*, June 2020.

Acknowledgements

I would like to express my deep gratitude to my supervisor, Prof. Giuseppe Durisi, for being a reliable provider of guidance, support, and collaboration throughout my PhD education. Your valuable insights and attention to detail have contributed greatly to my research journey, the work in this thesis, and my own academic development. If it were not for you, this work would not have been possible. To Fredrik Kahl, Cristopher Zach, and Benjamin Guedj: thank you for all the helpful discussions that have provided me with a wider perspective. Moreover, I want to thank Rui Castro and Tim van Erven for everything during my visit to the Netherlands.

I appreciate my current and former office mates, the Communication Systems group, and WASP for improving this journey. Thanks to Peter Grünwald, Yevgeny Seldin, and Mikael Skoglund for serving as committee, and to Gergely Neu for being opponent.

Sincerely: thank you to my family* for your continual encouragement throughout. Having you by my side has provided a constant source of motivation. Your belief in me, invaluable support, and love has my profound gratitude.

Thank you.

Fredrik Hellström
Göteborg, December 2022

Financial Support

This work was supported by Wallenberg Artificial Intelligence, Autonomous Systems and Software Program (WASP) and Chalmers AI Research Center (CHAIR).

Acronyms

CMI:	conditional mutual information
CNN:	convolutional neural network
DNN:	deep neural network
e-CMI:	evaluated conditional mutual information
f-CMI:	functional conditional mutual information
FCNN:	fully connected neural network
GD:	gradient descent
GPU:	graphics processing unit
i.i.d.:	independent and identically distributed
IM:	information measure
KL:	Kullback-Leibler
ML:	machine learning
NN:	neural network
PAC:	probably approximately correct
ReLU:	rectified linear unit
SGD:	stochastic gradient descent
SGLD:	stochastic gradient Langevin dynamics
VC:	Vapnik-Chervonenkis

Contents

Part I

Overview

CHAPTER 1

Background

A fundamental building block of human learning is our ability to accurately generalize knowledge from past experiences to new situations. For instance, when we observe adverse health effects following the consumption of a poisonous mushroom, we do not necessarily think that this is an isolated incident connected to this individual mushroom: we grow suspicious of the entire species. If we lacked the ability to identify relevant factors in one scenario and recognize them in a similar event, every moment of our lives would appear brand new, wholly separated from our history. For human infants, it suffices to be presented with only a handful examples from a category—sometimes as few as three—before learning the general concept [1]. Without this ability to generalize, it would be hard to imagine any possibility of efficient action in an ever-changing environment.

In recent years, machine learning (ML) methods have found enormous success in a variety of areas, such as medical diagnosis, chess, and protein structure prediction [2–4]. The basic idea underpinning modern ML is to create a computer program that can perform some objective, defined on the basis of a large data set referred to as the *training data*. The program is often referred to as a *hypothesis*, and the process of selecting it is called a *learning algorithm*. How well the hypothesis performs its objective, given some data, is measured by a *loss function*, where a lower value implies better performance. The true goal of ML is to choose a learning algorithm such that the loss function of the hypothesis is small not only for the training data, but for new, unseen data—like humans, the hypothesis should be able to generalize.

The study of generalization within ML is the main goal of *statistical learning theory*. Several classical results in this field have successfully established conditions under which generalization can be guaranteed. These results typically rely on the hypothesis class, from which the hypothesis is chosen, not being too complex [5]. A celebrated complexity measure is the Vapnik-Chervonenkis (VC) dimension, named after two pioneers within the field. The fact that complexity is tied to generalization can be intuitively motivated by Occam's razor: in the same way that the simplicity of an explanation can be predictive of its veracity, the simplicity of an ML hypothesis that performs well on the training data should be indicative of how similar its performance on new data will be. In contrast, a learning algorithm that utilizes a sufficiently complex hypothesis class can memorize a training set, without actually learning any generalizable pattern. This is related to the phenomenon known as *overfitting*—the hypothesis fits the training data *too* well. In such a scenario, achieving good performance on training data does not necessarily imply that something of value has been extracted from the data.

Intriguingly, when it comes to modern ML, this classical theoretical machinery is of little explanatory value. Many of the success stories of recent years make use of *deep neural networks* (DNNs), which are often able to generalize despite boasting enormous complexity. While the performance achieved in practice speaks for itself, theory has yet to catch up. A common criticism against DNNs is that they are used as a black box: we simply feed training data into the learning algorithm and use the results that emerge from the procedure, without any detailed understanding of how and why it works. This can hinder the adoption of ML solutions in safety-critical applications, such as health care or self-driving cars, where more rigorous performance guarantees are desired.

The need for new performance guarantees that are applicable even for DNNs has spurred a flurry of research activity. The lesson that is learned from the failure of the classical theory is that relying on model complexity alone is not enough. For this reason, new bounds are *data*- or *algorithm-dependent*. The basic insight underlying this approach is that, while generalization may fail for a worst-case data distribution or poor learning algorithm, it may work excellently for natural data distributions and practically relevant learning algorithms. This data-dependence is necessary for bounds to apply to DNNs. Consider, as an example, a classification setting, where each datum consists of an example and an associated label. Then, typical DNNs can accurately classify a training set both in the setting where the examples are paired with their true labels and the setting where the labels are determined randomly [6]. In the true-label setting, the DNN performs well on unseen data, while this is obviously impossible in the random-label setting—randomized labels mean that there is nothing to learn from the data! Since the only thing that separates these settings is the data distribution, this is a necessary ingredient of any bound that hopes to explain this phenomenon.

One drawback of deep neural networks is the heavy computational burden that they incur. For some of the largest models currently used, hundreds of thousands of graphics processing unit (GPU) hours are needed [7]. This is further aggravated by the practice of *hyperparameter search*, where different values for parameters of the learning algorithm, such as learning rate and network architecture, are evaluated to find the most suitable ones for the task at hand. One approach to automate this procedure and make it more efficient is *meta learning*. In meta learning, a meta learner has access to data from several different, but related, tasks. This can, for example, be different instances of image classification tasks. The objective of the meta learner is to find good hyperparameters for a base learner, which is applied to each task. The aim, then, is to find hyperparameters that generalize in the sense that they also yield good performance on new related tasks that were not used for training.

In this thesis, we take some steps toward explaining generalization for randomized learning algorithms, and in particular, we present new results for DNNs and meta learning. In Contribution A, we present a framework that can recover several of the information-theoretic bounds available in the literature, while also allowing us to derive new bounds. This framework is based on exponential inequalities, from which generalization bounds follow from simple manipulations. We combine this framework with the conditional mutual information (CMI) setting introduced by Steinke and Zakynthinou [8], where we can derive even tighter bounds. Additionally, we derive bounds using results from hypothesis testing and an approach based on Hölder's inequality, inspired by [9]. In Contribution B, we strengthen the previously obtained bounds for the CMI setting even further, improving their dependence on the size of the training data set. We also show how they naturally lead to data-dependent priors, which has recently been shown to be vital to obtain numerically accurate bounds. We demonstrate how to evaluate the bounds both from Contribution A and Contribution B in the setting of DNNs, and show that for some simple neural network setups, the obtained results predict the true generalization fairly accurately, and are in line with the best previously reported results. However, they diverge as training progresses. In Contribution C, we obtain new bounds for the average generalization error in the CMI framework that are tighter in several ways. First, the bounds are explicit in the disintegrated, samplewise, evaluated CMI, leading to a tighter characterization than the ordinary CMI. Second, we allow for arbitrary convex comparator functions, whereas previous bounds used the (weighted) absolute difference. We demonstrate that this leads to a numerically accurate characterization of the generalization error for some neural network settings, which remains stable throughout training. We also analytically study a multiclass classification setting, where we show that our bounds are expressive enough to recover essentially optimal minimax bounds. Finally, in Contribution D, we extend our analysis to meta learning. By combining the CMI framework with a one-step approach, where previous studies used a two-step approach, we obtain several novel bounds that are tighter than previous work. By specializing our results to a representation learning setting, we show that our bounds recover the convergence rates of classical results for meta learning, demonstrating their expressiveness.

1.1 Thesis Structure

This thesis is comprised of two parts. Part I contains an introduction to the field, and serves the purpose of putting the sequel into context. Part II comprises the research contributions upon which this thesis is based.

Part I is organized as follows. In Chapter 1, we first give an informal overview of the field in order to contextualize the specific problems that we discuss in this thesis, before introducing necessary notation. Then, in Chapter 2, we present the paradigm of statistical learning theory and review some of the classical generalization guarantees. In Chapter 3, we define the main information measures that appear in this thesis and discuss some of their properties, before presenting some concentration inequalities that are used in the remainder of the thesis. Next, in Chapter 4, we survey information-theoretic generalization guarantees from the literature, beginning with average bounds before proceeding with probably approximately correct (PAC)-Bayesian and single-draw bounds, concluding with the CMI framework. In Chapter 5, we give a brief description of neural networks and meta learning, along with some generalization guarantees for these settings. In Chapter 6, we conclude the first part of the thesis by detailing the contributions given in Part II and discussing possible future directions to investigate.

1.2 Notation

Unless otherwise stated, the probability distribution of a random variable X is denoted by P_X. We indicate the fact that the random variable X is distributed according to P_X by writing $X \sim P_X$. For two random variables X and Y, their joint distribution is denoted by P_{XY} and the product of their marginal distributions is denoted by $P_X P_Y$. For a probability measure P_X, we denote the probability operator under it as $P_X[\,\cdot\,]$. For a function $f(X)$, we denote its expectation over $X \sim P_X$ as $\mathbb{E}_{X\sim P_X}[f(X)]$. When there is no risk of confusion, we write $\mathbb{E}_{P_X}[f(X)]$ or simply $\mathbb{E}[f(X)]$. The indicator function of an event E is denoted by $1\{E\}$.

CHAPTER 2

Statistical Learning

In this chapter, we begin by more formally introducing the learning setting that we consider throughout most of this thesis. We then discuss various flavors of generalization guarantees, before presenting classical generalization bounds that are based on the VC dimension and the Rademacher complexity. In Chapter 4, these classical results will be contrasted with more recently obtained information-theoretic guarantees.

2.1 The Learning Setting

We start by discussing the general ingredients that are common to most of the learning settings considered in this thesis, before giving some specific examples. Assume that there is an unknown data distribution P_Z on some instance space $\mathcal{Z}$, and that from this distribution, we have obtained a data set $\boldsymbol{Z} = (Z_1, Z_2, \dots, Z_n)$, consisting of n samples drawn independenetly from P_Z. We will refer to $\boldsymbol{Z}$ as the *training set.* Based on this training set, we want to choose a hypothesis W from a hypothesis space $\mathcal{W}$. This is done by using a learning algorithm, characterized by a conditional distribution $P_{W|\boldsymbol{Z}}$ on $\mathcal{W}$ given $\boldsymbol{Z}$. To measure how good a particular choice W is, we use a loss function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}^+$. The averaged loss of a given $w \in \mathcal{W}$ for a specific training set $\boldsymbol{z} = (z_1, \dots, z_n)$ is given by $L_{\boldsymbol{z}} = \frac{1}{n}\sum_{i=1}^n \ell(w, z_i)$, and is referred to as the *training loss.* The expected loss on a new sample, the *population loss*, is given by $L_{P_Z}(w) = \mathbb{E}_{P_Z}[\ell(w, Z)]$. The *generalization error* is the difference between these, $\operatorname{gen}(w, \boldsymbol{z}) = L_{P_Z}(w) - L_{\boldsymbol{z}}(w)$. This generic learning setting is illustrated in Figure 2.1.

A commonly considered learning algorithm is that of *empirical risk minimization*,

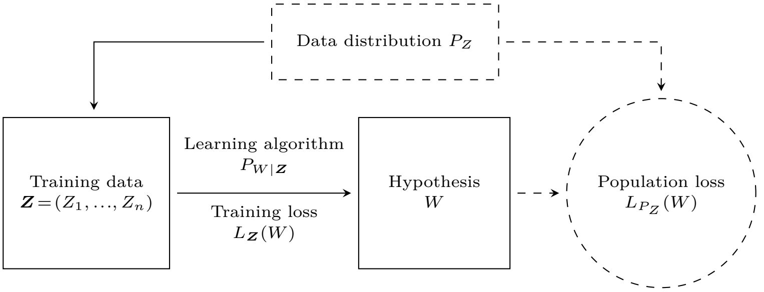

Figure 2.1: A schematic illustration of the learning setup considered in this thesis.

in which the support of $P_{W|\boldsymbol{Z}}$ is limited to $\arg\min_{w \in \mathcal{W}} L_{\boldsymbol{Z}}(W)$. Since there may be imperfections such as noise in the training data, one may not want to perform exact empirical risk minimization, but rather an approximate variant. For example, one may add a regularizer, which limits the model selection, or add noise to the output of the training algorithm.

We now give some specific examples that fit into the general learning setup.

Estimating the mean of a Gaussian distribution: In this setting, the data $Z \in \mathbb{R}$ are samples drawn independently from some Gaussian distribution $\mathcal{N}(\mu, \sigma)$. Here, the hypothesis space is $\mathcal{W} = \mathbb{R}$, and the goal is to find a w that approximates μ. A possible choice for the loss function is $\ell(w, z) = (w - z)^2$. A reasonable learning algorithm in this setting is to use the sample mean: for a training set $\boldsymbol{z}$, set $w = \frac{1}{n}\sum_{i=1}^{n} z_i$. Notice that this is an example of an empirical risk minimizer. The average generalization error of this learning algorithm can be exactly computed as

$$\mathbb{E}_{P_{W\boldsymbol{Z}}}[\operatorname{gen}(W, \boldsymbol{Z})] = \mathbb{E}_{P_{\boldsymbol{Z}}P_Z}\left[\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - Z\right)^2\right] = \frac{2\sigma^2}{n}. \tag{2.1}$$

Regression: In regression, the data are decomposed as $Z = (X, Y)$ where $X \in \mathcal{X}$ is an example from some space $\mathcal{X}$ and $Y \in \mathcal{Y}$ is a label from a continuous space $\mathcal{Y}$. As an example, $\mathcal{X} = \mathbb{R}^3$ can be the coordinate of a point in space, while $\mathcal{Y} = \mathbb{R}^+$ is the temperature in Kelvin. The goal is to learn a function $W : \mathcal{X} \to \mathcal{Y}$ that predicts the temperature at each point in space. For regression, a typical loss function is the squared loss given by $\ell(w, z) = \frac{1}{2}(w(X) - Y)^2$. A possible learning algorithm for this setting is to use a linear predictor given by the least-squares solution.

Classification: In classification, the data are again decomposed as $Z = (X, Y)$, where $X \in \mathcal{X}$ is an example from some space $\mathcal{X}$, but now $Y \in \mathcal{Y}$ is a label from a discrete set $\mathcal{Y}$. In the well-studied setting of *binary classification*, $|\mathcal{Y}| = 2$. As an example, $\mathcal{X} = [0, 1]^{3P}$ can be the normalized RGB values of images with P pixels depicting either cats or dogs, while $\mathcal{Y} = \{0, 1\}$, where 0 corresponds to cats and 1 to dogs. The goal is to learn a function $w : \mathcal{X} \to \mathcal{Y}$ that classifies pictures as either cats or dogs. A typical choice for the loss function is the *classification error*, given by $\ell(w, z) = 1\{w(X) \neq Y\}$. A learning algorithm that has found great success for image recognition tasks, such as classifying cats and dogs, is to train a convolutional neural network (CNN) using some variant of stochastic gradient descent (SGD) [10].

While the learning setting described in this section is quite general, it does not cover all possible settings of interest. An alternative setting that we consider in this thesis is meta learning, where a meta learner has access to data from several tasks, drawn from the same task distribution, and its goal is to select a hyperparameter that improves the performance of a base learner on new tasks from the task distribution. We introduce this setting in more detail in Section 5.2. Beyond this, the assumption that the training data $Z_1, Z_2, \dots, Z_n$ are independent and identically distributed can be lifted [11]. Other settings include online learning, where the data arrives sequentially and the aim is to achieve a small total loss on the observed samples [5, Chapter 21]; active learning, where the learner has access to unlabeled data samples and chooses which labels to request [12]; and unsupervised learning, where unlabeled data is used [13].

2.2 Classical Generalization Guarantees

As previously mentioned, the goal of learning is to find a hypothesis W that achieves a small population loss $L_{P_Z}(W)$. This is complicated by the fact that we only have access to an estimate of the population loss, the training loss $L_{Z^n}(W)$, which is based on n independent samples drawn from P_Z. In this section, we present some classical results which guarantee that, under some conditions, the training loss is a good proxy for the population loss.

2.2.1 Different Flavors of Generalization

Due to the stochastic nature of learning algorithms that we consider, results relating to generalization do not come in a single form. We now present the different flavors of generalization guarantees that we discuss throughout this thesis.

PAC learnability: We begin by presenting the probably approximately correct (PAC) framework for studying learning, since this is the setting of the classical results that we will discuss. A hypothesis class $\mathcal{W}$ is PAC learnable if, for every distribution P_Z, there exists a learning algorithm $P_{W|\boldsymbol{Z}}$ such that, for every $\epsilon, \delta \in (0, 1)$, there exists an $m(\epsilon, \delta)$ such that if $n \geq m(\epsilon, \delta)$,

$$L_{P_Z}(W) \leq \inf_{w \in W} L_{P_Z}(w) + \epsilon \tag{2.2}$$

with probability at least $1 - \delta$ over $P_{\boldsymbol{Z}}$. Here, $m(\epsilon, \delta)$ is referred to as the *sample complexity.* We now see the motivation for the name: the hypothesis W that we choose will *probably* (with probability at least $1-\delta$) be *approximately* (with a margin of ϵ) *correct* (in the sense of obtaining the smallest population loss achievable in the hypothesis class). If we assume that our learning problem is *realizable*, there is a hypothesis in the class that has zero population loss, so that $\inf_{w \in W} L_{P_Z}(w) = 0$. Note that the PAC formulation of generalization is focused on properties of the hypothesis class $\mathcal{W}$ itself.

Average bounds: In the average setting, the quantity of interest is the expected value of the population loss averaged over both the training sample and the randomness of the algorithm, i.e., $\mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{P_Z}(W)]$. In some settings, this quantity is relatively easy to analyze, but a drawback is that average guarantees may not give much relevant information in practice. Typically, one has a single instance of a training set, and wants to know whether or not one can achieve generalization based on this particular instance. Bounds on the average loss do not necessarily imply good guarantees on the tail of the loss distribution with respect to the data. Still, bounds on the average population loss can provide insight into how and when learning algorithms are expected to work.

PAC-Bayesian bounds: The PAC-Bayesian setting was introduced by McAllester [14] in an effort to derive PAC-style bounds for Bayesian-flavored estimators. In this setting, we assume that the algorithm $P_{W|\boldsymbol{Z}}$ is used to select a new W for each time that the hypothesis is used. Therefore, the quantity of interest is $\mathbb{E}_{P_{W|\boldsymbol{Z}}}[L_{P_Z}(W)]$. Since this is a random variable in $\boldsymbol{Z}$, we note that any bound on it will have to hold only with some probability $1-\delta$ over $P_{\boldsymbol{Z}}$. An attractive feature of the PAC-Bayesian setting is that it can incorporate correlation between and uncertainty about hypotheses, since we do not consider a single, fixed W [15].

Single-draw bounds: In the single-draw setting, we instead consider a single training set $\boldsymbol{Z}$ and a single hypothesis W drawn from our algorithm $P_{W|\boldsymbol{Z}}$, which we will use for all future predictions. The quantity of interest is therefore simply $L_{P_Z}(W)$, and bounds on this random variable will hold with some probability $1-\delta$ over $P_{W\boldsymbol{Z}}$. This setting describes many real-world applications of machine learning. For instance, the standard procedure when using neural networks is to optimize the weights using a stochastic algorithm, and then use the fixed weights that one obtains for future applications.

Data-dependent or data-independent: When it comes to the two tail bounds, i.e., the PAC-Bayesian and single-draw settings, results can be either data-dependent, when bounds on the population loss depend on the particular instance of the training set $\boldsymbol{Z}$, or data-independent, when they do not depend on the specific instance. The benefit of data-dependent bounds is that they can be used as regularizers: adjusting the algorithm to make the bound small may lead to improved generalization. Furthermore, data-independent bounds can often be obtained as weakened versions of data-dependent ones. Data-independent bounds, however, can be used to compute the *sample complexity*, i.e., the number of samples needed to guarantee a given precision with a given probability. Of course, the ability to make statements about generalization guarantees without referring to a specific training set can also be useful.

Test loss or population loss: So far, we have discussed guarantees related to the population loss $L_{P_Z}(W)$. However, in some circumstances it is more convenient to obtain bounds on a *test loss* $L_{\bar{\boldsymbol{Z}}}(W)$, i.e., the loss evaluated on a sample $\bar{\boldsymbol{Z}}$ that is independent of W. When empirically evaluating learning algorithms, the true data distribution P_Z

is typically unknown, so in practice one usually has to resort to using a test loss as an estimate. For many settings of interest, any bound on the test loss can be converted into a bound on the population loss through the use of concentration inequalities.

2.2.2 VC Dimension

The Vapnik-Chervonenkis (VC) dimension, named after two pioneers of statistical learning, is a geometric property of the hypothesis class $\mathcal{W}$ that can be used to characterize when generalization can be guaranteed. It applies to the setting of binary classification, where the data Z consist of examples X and labels $Y \in \{0,1\}$ and $\mathcal{W}$ is a set of functions from $\mathcal{X}$ to $\{0,1\}$, as described in the previous section.[1] While our discussion in this section is restricted to the binary classification setting, analogous quantities have been studied in other settings, such as the fat-shattering dimension for regression and the Natarajan dimension for multi-class classification [5, Sec. 6.7]. In a sense, the VC dimension characterizes how many functions there are in $\mathcal{W}$. If the VC dimension is infinite, any function from $\mathcal{X}^n$ to $\{0,1\}^n$ can be expressed by a member of $\mathcal{W}$ for all values of n. However, if it is small, the number of expressible functions are limited in some sense. Below, we give the formal definition of the VC dimension. In so doing, we will also introduce the closely related *growth function* and the concept of *shattering*.

Definition 2.1 (Shattering, growth function, and VC dimension)**.** *A hypothesis class $\mathcal{W}$ is said to shatter a set $X^n \in \mathcal{X}^n$ if*

$$|\{w(X_1), \dots, w(X_n) : w \in \mathcal{W}\}| = 2^n. \tag{2.3}$$

Let $\tau_{\mathcal{W}}(n)$ denote the growth function defined as

$$\tau_{\mathcal{W}}(n) = \max_{X^n \in \mathcal{X}^n} |\{w(X_1), w(X_2), \dots, w(X_n) : w \in \mathcal{W}\}|. \tag{2.4}$$

The VC dimension d of $\mathcal{W}$ is the largest integer such that $\tau_{\mathcal{W}}(d) = 2^d$. If there is no such integer, we say that $d = \infty$. Thus, if d is finite, $\mathcal{W}$ shatters some set of size d but no set of size $d+1$.

The relation between finite VC dimension and generalization can now be intuited. If we find a function w from a space with VC dimension d that achieves a small loss on a training set Z^n with $n \gg d$, we know that we must have identified some structure in the data: it is not possible that we simply memorized the given samples. In contrast, if the VC dimension is infinite, we can not be certain that the function we found does anything more than encode the training samples. This intuition is formalized in the following theorem [5, Thm. 6.8].

[1] Alternatively, $\mathcal{W}$ can be a parameter space, the members of which characterize parametric functions from $\mathcal{X}$ to $\{0,1\}$. For simplicity of notation, we will consider $\mathcal{W}$ to be the function space.

Theorem 2.1 (Generalization guarantee from VC dimension). *Let $\mathcal{W}$ be of finite VC dimension d. Then, for every distribution P_Z, there exists a learning algorithm $P_{W|\boldsymbol{Z}}$ and constant C such that, for every $\epsilon, \delta \in (0,1)$, we have that with probability at least $1-\delta$ over $P_{\boldsymbol{Z}}$,*

$$L_{P_Z}(W) \leq \inf_{w \in W} L_{P_Z}(w) + \epsilon \tag{2.5}$$

provided that

$$n \geq C \frac{d + \log \frac{1}{\delta}}{\epsilon^2}. \tag{2.6}$$

Furthermore, $\mathcal{W}$ is PAC learnable, with sample complexity bounded above and below as

$$C' \frac{d + \log \frac{1}{\delta}}{\epsilon^2} \leq m(\epsilon, \delta) \leq C \frac{d + \log \frac{1}{\delta}}{\epsilon^2} \tag{2.7}$$

for some constants C, C'.

In the realizable setting, where there is a hypothesis $w^* \in \mathcal{W}$ that achieves zero population loss, i.e. $L_{P_{\boldsymbol{Z}}}(w^*) = 0$, bounds on the sample complexity with a more beneficial dependence on the approximation error ϵ can be obtained. These bounds can be inverted to obtain high-probability bounds on the population loss, which have an n-dependence of $\tilde{O}(1/n)$, where the $\tilde{O}(\,\cdot\,)$ notation indicates that we are ignoring logarithmic factors. In comparison, the corresponding population loss bound that can be obtained from Theorem 2.1 has a $\tilde{O}(1/\sqrt{n})$ dependence. The rate $\tilde{O}(1/n)$ is typically referred to as a *fast rate*, while $\tilde{O}(1/\sqrt{n})$ is a *slow rate*. Below, we present the VC dimension-based sample complexity for the realizable setting, which can be used to obtain fast-rate population loss bounds.

Theorem 2.2 (Fast-rate generalization guarantee from VC dimension). *Let $\mathcal{W}$ be of finite VC dimension d. Assume that there is a hypothesis $w^* \in \mathcal{W}$ such that $L_{P_{\boldsymbol{Z}}}(w^*) = 0$. Then, $\mathcal{W}$ is PAC learnable, with sample complexity bounded above and below as*

$$C' \frac{d + \log \frac{1}{\delta}}{\epsilon} \leq m(\epsilon, \delta) \leq C \frac{d \log(1/\epsilon) + \log \frac{1}{\delta}}{\epsilon} \tag{2.8}$$

for some constants C, C'.

For further discussion about fast-rate bounds and the conditions under which they can be obtained, see [16, 17].

Due to the existence of both upper and lower bounds on the sample complexity of $\mathcal{W}$ in terms of d, the VC dimension completely characterizes learnability in the PAC sense. This is a remarkable feature of the VC-based generalization guarantee, but as previously discussed, it is not enough to explain the successes of modern machine learning algorithms. This indicates that standard PAC learnability may not be the pertinent concept to study when it comes to modern machine learning.

2.2.3 Rademacher Complexity

Another classical metric that can be used for guaranteeing generalization is the *Rademacher complexity.* Notably, the Rademacher complexity of a hypothesis class $\mathcal{W}$ is defined with respect to a given data set. Given the arguments for the necessity of incorporating some kind of data-dependence into our generalization guarantees, this seems like a promising approach to obtain tight generalization bounds. We now give the definition of Rademacher complexity. Unless otherwise specified, all of the material in this section is based on [5, Chap. 26].

Definition 2.2 (Rademacher complexity)**.** *Let $\boldsymbol{Z} \in \mathcal{Z}^n$ be a set of data samples and let $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}^+$ be a loss function. Let σ_i for $i = 1, \dots, n$ 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*

$$\mathrm{Rad}_{\boldsymbol{Z}}(\mathcal{W}) = \frac{1}{n}\mathbb{E}_{P_{\sigma_1 \dots \sigma_n}}\left[\sup_{w \in \mathcal{W}} \sum_{i=1}^{n} \sigma_i \ell(w, Z_i)\right]. \tag{2.9}$$

One way to understand the Rademacher complexity is to think of randomly splitting the data set $\boldsymbol{Z}$ into a training set and a test set. 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, if we are equally likely to assign each data point to either the training set or the test set. 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.

Theorem 2.3 (Generalization guarantee from Rademacher complexity)**.** *Assume that, for all $z \in \mathcal{Z}$ and all $w \in \mathcal{W}$, $|\ell(w, z)| \le c$. With probability at least $1 - \delta$ over $P_{\boldsymbol{Z}}$, for all $w \in \mathcal{W}$,*

$$L_{P_Z}(w) - L_{\boldsymbol{Z}}(w) \le 2\mathrm{Rad}_{\boldsymbol{Z}}(\mathcal{W}) + c\sqrt{\frac{2\log(2/\delta)}{n}}. \tag{2.10}$$

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

As discussed in [5, 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. It has also been used to study generalization in neural networks found by gradient descent [18], albeit without providing nonvacuous guarantees. One issue with the Rademacher complexity is that, while being data-dependent, it is still a worst-case measure over the hypothesis class. This leads to generalization estimates for modern machine learning algorithms that are overly pessimistic.

CHAPTER 3

Information Measures and Concentration Inequalities

In this section, we will introduce the tools that will be used to derive the information-theoretic generalization bounds in Part I of this thesis. Specifically, we will introduce some common information measures in Section 3.1, discuss change of measure techniques in Section 3.2, and present concentration inequalities in 3.3.

3.1 Information Measures

Formally, given a measurable space $\mathcal{X}$ and the associated family $\mathcal{M}_1(\mathcal{X})$ of probability measures on $\mathcal{X}$, an information measure is a mapping $\mathrm{IM} : \mathcal{M}_1(\mathcal{X}) \times \mathcal{M}_1(\mathcal{X}) \to \mathbb{R}$. Typically, for all $P \in \mathcal{M}_1(\mathcal{X})$, we have $\mathrm{IM}(P, P) = 0$. Thus, an information measure is some way to quantify the discrepancy between two probability measures. Often, these information measures are not metrics, as they may not satisfy symmetry, the triangle inequality, or even non-negativity. Throughout information theory and machine learning, such quantities are exceedingly useful and abundant. In the context of information-theoretic generalization bounds, we find that they can provide upper bounds on the loss of learning algorithms. In this section, we will introduce some information measures along with their properties that will be useful in later sections. For a more detailed review, the reader is referred to, for example, [19, 20], upon which much of the material in this section is based.

A basic building block of many information measures is some kind of likelihood ratio. For two probability mass functions P and Q on a common space $\mathcal{X}$, their likelihood ratio at a point $x \in \mathcal{X}$ is defined as $P(x)/Q(x)$. Similarly, if p and q are probability densities, the likelihood ratio is $p(x)/q(x)$. For generic measures P and Q, this concept is captured by Radon-Nikodym derivative, denoted by $\mathrm{d}P/\,\mathrm{d}Q$. For the cases of discrete or continuous random variables, it reduces to the aforementioned likelihood ratios. The precise meaning of this object is captured by the Radon-Nikodym theorem, a change of measure that relates probabilities of events under P with their probabilities under Q. We will present this result in Theorem 3.3. The Radon-Nikodym derivative exists whenever P is absolutely continuous with respect to Q, denoted by $P \ll Q$. This means that for any measurable set $\mathcal{E}$ such that $Q(\mathcal{E}) = 0$, we also have $P(\mathcal{E}) = 0$. In other words, the support of P is contained in the support of Q.

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

Definition 3.1 (Information density)**.** *The information density between two random variables X and Y with joint distribution P_{XY} and marginal distributions P_X and P_Y is given by*

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

The conditional information density between X and Y given Z is

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

One very commonly used information measure is the Kullback-Leibler (KL) divergence, sometimes referred to as the relative entropy. We provide its definition below.

Definition 3.2 (The KL divergence)**.** *Consider two probability distributions P and Q defined on a common measurable space such that P is absolutely continuous with respect to Q, denoted by $P \ll Q$. The KL divergence between P and Q is given by*

$$D(P \,||\, Q) = \mathbb{E}_P\left[\log \frac{\mathrm{d}P}{\mathrm{d}Q}\right]. \tag{3.3}$$

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

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

$$D(P_{Y|X} \,||\, Q_{Y|X} \,|\, P_X) = \mathbb{E}_{P_X}\left[D(P_{Y|X} \,||\, Q_{Y|X})\right]. \tag{3.4}$$

The KL divergence satisfies a useful property called the *chain rule.*

Theorem 3.1 (The chain rule of KL divergence)**.** *Given the distributions $P_{XY} = P_X P_{Y|X}$ and $Q_{XY} = Q_X Q_{Y|X}$, we have*

$$D(P_{XY} \,||\, Q_{XY}) = D(P_{Y|X} \,||\, Q_{Y|X} \,|\, P_X) + D(P_X \,||\, Q_X). \tag{3.5}$$

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

Definition 3.3 (Mutual information)**.** *The mutual information between two random variables X and Y with joint distribution P_{XY} and marginal distributions P_X and P_Y is given by*

$$I(X;Y) = D(P_{XY} \,||\, P_X P_Y) = \mathbb{E}_{P_{XY}}[\imath(X,Y)]\,. \tag{3.6}$$

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

$$I(X;Y|Z) = D(P_{XY|Z} \,||\, P_{X|Z}P_{Y|Z} \,|\, P_Z) = \mathbb{E}_{P_{XYZ}}[\imath(X,Y|Z)]\,. \tag{3.7}$$

We now see the motivation behind the name information density—its average is the mutual information (with an analogous correspondence for the conditional information density). The mutual information is one of the most fundamental quantities in information theory, and famously characterizes the capacity of any communication channel. Recently, it has garnered interest in the statistical learning community as a measure of generalization. This correspondence comes about by viewing the randomized learning algorithm as a channel—mathematically, both are just conditional probability laws.

The mutual information inherits the chain rule from KL divergence, so that $I(X,Y;Z) = I(X;Z) + I(Y;Z|X) = I(Y;Z) + I(X;Z|Y)$.

The KL divergence is a special case of a wider class of information measures called f-divergences, which share many of the desirable properties of the KL divergence.

Definition 3.4 (f-divergence)**.** *Let P and Q be two probability distributions on a common measurable space $\mathcal{X}$ such that $P \ll Q$. Let $f : [0,\infty) \to \mathbb{R}$ be a convex and lower semi-continuous function with $f(1) = 0$. Then, the f-divergence between P and Q is defined as*

$$D_f(P \,||\, Q) = \mathbb{E}_Q\left[f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\right]. \tag{3.8}$$

With $f(x) = x\log x$, we recover the KL divergence. Other notable examples include the total variation $TV(P,Q) = \mathbb{E}_Q\left[\left|\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\right|\right]/2$, obtained by setting $f(x) = |x-1|/2$, and the χ^2-divergence $\chi^2(P \,||\, Q) = \mathbb{E}_Q\left[(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1)^2\right]$, obtained by setting $f(x) = (1-\sqrt{x})^2$. We now review some of the useful properties of f-divergences. For proofs, see [19, Thm. 6.1, 6.2].

Theorem 3.2 (Properties of f-divergences)**.** *For any f-divergence, the following properties hold.*

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

2. ***Data-processing:*** *Let P_X and Q_X be two distributions on $\mathcal{X}$, and let P_Y and Q_Y be the corresponding distributions on $\mathcal{Y}$ induced by a kernel $P_{Y|X}$, that is, $P_Y(y) = \int_{\mathcal{X}} \mathrm{d}P_X(x) P_{Y|X=x}$ and $Q_Y(y) = \int_{\mathcal{X}} \mathrm{d}Q_X(x) P_{Y|X=x}$. Then,*
$$D_f(P_X \,||\, Q_X) \leq D_f(P_Y \,||\, Q_Y). \tag{3.9}$$
3. ***Conditioning increases divergence:*** *Let P_X be a distribution on $\mathcal{X}$, and let P_Y and Q_Y be the distributions induced on $\mathcal{Y}$ by two kernels $P_{Y|X}$ and $Q_{Y|X}$ respectively, that is $P_Y(y) = \int_{\mathcal{X}} \mathrm{d}P_X(x) P_{Y|X=x}$ and $Q_Y(y) = \int_{\mathcal{X}} \mathrm{d}P_X(x) Q_{Y|X=x}$. The* conditional *$f$-divergence is defined as*
$$D_f(P_{Y|X} \,||\, Q_{Y|X} \,|\, P_X) \equiv \mathbb{E}_{P_X}\big[D_f(P_{Y|X} \,||\, Q_{Y|X})\big] \tag{3.10}$$
and it satisfies the inequality
$$D_f(P_Y \,||\, Q_Y) \leq D_f(P_{Y|X} \,||\, Q_{Y|X} \,|\, P_X). \tag{3.11}$$

Notably, unlike the KL divergence, general f-divergences do *not* satisfy the chain rule.

Another special instance of f-divergences are the Rényi divergences, also known as α-divergences [21].

Definition 3.5 (Rényi divergence). *Let $\alpha \in (0,1) \cup (1,\infty)$. Then, the Rényi divergence of order α between P and Q is defined as*
$$D_\alpha(P \,||\, Q) = \frac{1}{\alpha - 1} \log \mathbb{E}_Q\left[\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)^\alpha\right]. \tag{3.12}$$
For $\alpha = 1$, motivated by continuity, the Rényi divergence of order 1 is defined as the KL divergence:
$$D_1(P \,||\, Q) = D(P \,||\, Q). \tag{3.13}$$
The conditional Rényi divergence of order α between $P_{Y|X}$ and $Q_{Y|X}$ given P_X is
$$D_\alpha(P_{Y|X} \,||\, Q_{Y|X} \,|\, P_X) = D_\alpha(P_{Y|X} P_X \,||\, Q_{Y|X} P_X). \tag{3.14}$$

When $P = P_{XY}$ and $Q = P_X P_Y$ are the joint distribution of two random variables and the product of their marginals respectively and $\alpha \to \infty$, the Rényi divergence reduces to the maximal leakage [22].

Definition 3.6 (Maximal leakage). *The maximal leakage from X to Y is defined as*
$$\mathcal{L}(X \to Y) = \log \mathbb{E}_{P_Y}\left[\operatorname*{ess\,sup}_{P_X} \frac{\mathrm{d}P_{XY}}{\mathrm{d}P_X P_Y}\right]. \tag{3.15}$$
Here, the essential supremum of a measurable function $f(\,\cdot\,)$ of a random variable X distributed as P_X is defined as
$$\operatorname*{ess\,sup}_{P_X} f(X) = \inf_{a \in \mathbb{R}}\Big[P_X\big(\{X : f(X) > a\}\big) = 0\Big]. \tag{3.16}$$

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

$$\mathcal{L}(X \to Y|Z) = \log \operatorname*{ess\,sup}_{P_Z} \mathbb{E}_{P_{X|Z}}\left[\operatorname*{ess\,sup}_{P_{Y|Z}} \frac{\mathrm{d}P_{XYZ}}{\mathrm{d}P_{X|Z}P_{Y|Z}P_Z}\right]. \tag{3.17}$$

While the maximal leakage is obtained as the infinite limit of the Rényi divergence, the same does not hold for the conditional maximal leakage. Instead, the conditional maximal leakage is the infinite limit of the conditional α-mutual information [23].

Definition 3.7 (α-mutual information)**.** *For $\alpha \in (0,1) \cup (1,\infty)$, the α-mutual information between X and Y is given by*

$$I_\alpha(X;Y) = \frac{1}{\alpha-1} \log \mathbb{E}^{\alpha}_{P_X}\left[\mathbb{E}^{1/\alpha}_{P_Y}\left[\exp\left(\frac{\mathrm{d}P_{XY}}{\mathrm{d}P_X P_Y}\right)^{\alpha}\right]\right]. \tag{3.18}$$

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

$$I_\alpha(X;Y|Z) = \frac{1}{\alpha-1} \log \mathbb{E}_{P_Z}\left[\mathbb{E}^{\alpha}_{P_{X|Z}}\left[\mathbb{E}^{1/\alpha}_{P_{Y|Z}}\left[\left(\frac{\mathrm{d}P_{XYZ}}{\mathrm{d}P_{X|Z}P_{Y|Z}P_Z}\right)^{\alpha}\right]\right]\right]. \tag{3.19}$$

It should be noted that the version of the conditional α-mutual information that we give here is not the only possible definition, and many others have been considered [24,25]. Our main reasons for focusing on this particular definition is its role in generalization bounds and its connection to the conditional maximal leakage.

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

$$I_\alpha(X;Y) \le D_\alpha(P_{Y|X} \,||\, P_Y \,|\, P_X) \tag{3.20}$$

$$I_\alpha(X;Y|Z) \le D_\alpha(P_{XY} \,||\, P_X P_Y). \tag{3.21}$$

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

3.2 Change of Measure

While the quantity of interest in statistical learning is the high-error event under the joint distribution of the hypothesis and the data, this can be difficult to control directly. Instead, there may be other, auxiliary distributions that allow for direct control of the high-error event—for instance, when one considers the hypothesis and the data to be drawn independently from each other, there are many situations where the concentration inequalities that we introduce in Section 3.3 readily apply. The technique of relating an event under one distribution to its corresponding value under another distribution is referred to as *change of measure.* To properly account for the fact that we are no longer working with the original distribution of interest, we need to have some handle on the discrepancy between the two distributions. As it turns out, the information measures that we described in the previous section often take this role.

In this section, we introduce two key instances of the change of measure technique. After defining absolute continuity, we introduce the Radon-Nikodym theorem, which is the backbone of many change of measure techniques. Then, we present the celebrated Donsker-Varadhan variational formula, which can be used to express averages under different distributions as a function of the KL divergence between the distributions.

3.2.1 Absolute Continuity

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

Definition 3.8 (Absolute continuity)**.** *A measure P is absolutely continuous with respect to a measure Q if, for any measurable set $\mathcal{E}$ such that $Q(\mathcal{E}) = 0$, we also have $P(\mathcal{E}) = 0$. This is denoted $P \ll Q$.*

Throughout this section, this property will be crucial for virtually every result. The importance of the absolute continuity property is that it guarantees the existence of the Radon-Nikodym derivative.

3.2.2 The Radon-Nikodym Theorem

Perhaps the most direct way to compare an event under two different distributions is to use the Radon-Nikodym theorem, sometimes simply referred to as “change of measure.” Provided that an absolute continuity requirement holds, it exactly relates the measure of an event under two distributions [26, Thm. 6.9(b)].

Theorem 3.3 (Radon-Nikodym theorem)**.** *Let P and Q be probability distributions on a common space such that $P \ll Q$. Then, there exists a function* $\mathrm{d}P/\,\mathrm{d}Q$ *such that, for any measurable event E,*

$$P(E) = \int_E \frac{\mathrm{d}P}{\mathrm{d}Q}\,\mathrm{d}Q. \tag{3.22}$$

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

For discrete random variables, a valid choice for $\mathrm{d}P/\,\mathrm{d}Q$ is simply the ratio between the probability mass functions of the two distributions. Similarly, for continuous random variables, we can choose the ratio between the probability densities. If the absolute continuity criterion $P \ll Q$ does not hold, we say that $\mathrm{d}P/\,\mathrm{d}Q = \infty$.

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

$$\imath(X,Y) = \log \frac{\mathrm{d}P_{XY}}{\mathrm{d}P_X P_Y}. \tag{3.23}$$

This can be used for the following change of measure: assume that we have $f(X,Y) = 0$ whenever $\imath(X,Y) = -\infty$. Note that, if we assume that $P_X P_Y \ll P_{XY}$, we always have $\imath(X,Y) > -\infty$ so that the condition is satisfied for any function f. Then, [19, Prop. 17.1]

$$\mathbb{E}_{P_{XY}}[f(X,Y)] = \mathbb{E}_{P_X P_Y}\left[e^{-\imath(X,Y)} f(X,Y)\right]. \tag{3.24}$$

Of course, the same type of result holds if we replace the product of marginals $P_X P_Y$ with an auxiliary distribution Q_{XY}, provided that the suitable absolute continuity requirements hold.

3.2.3 The Donsker-Varadhan Variational Formula

The celebrated Donsker-Varadhan variational formula for the KL divergence has its origins in the work of [27]. It has a rich history both in the fields of information theory and machine learning, and is a core tool of many influential concepts. It has been repeatedly rediscovered in many contexts, including the PAC-Bayesian literature, being referred to as the shift of measure lemma [28] and the compression lemma [29]. We state this important result below.

Theorem 3.4 (Donsker-Varadhan variational formula)**.** *Let P and Q be two probability distributions on a common measurable space $\mathcal{X}$ such that $P \ll Q$. Let F denote the set of functions $f : \mathcal{X} \to \mathbb{R}$ such that $\mathbb{E}_Q\left[e^{f(X)}\right] < \infty$,*

$$D(P\,||\,Q) = \sup_{f \in F} \mathbb{E}_P[f(X)] - \log \mathbb{E}_Q\left[e^{f(X)}\right]. \tag{3.25}$$

Theorem 3.4 relates the expectation of $f(X)$ under P to the moment-generating function of $f(X)$ under Q, in terms of the KL divergence between the two distributions. This is abundantly useful throughout information theory and machine learning.

3.3 Concentration Inequalities

While the change of measure techniques discussed in Section 3.2 are useful for going from a hard probability distribution to an easier auxiliary one, this is of little use if we cannot control the expressions with the auxiliary distribution. In this section, we present methods for controlling expected values and tail probabilities for various categories of random variables. For a more detailed review of this vast topic, we refer the reader to, for example, [30–32], where many of the proofs of the results presented here can be found.

3.3.1 Sub-Gaussian Random Variables

A commonly studied category are sub-Gaussian random variables. A random variable is said to be sub-Gaussian with parameter σ, or σ-sub-Gaussian, if its tail is dominated by that of a Gaussian random variable with variance σ^2. Below, we give several equivalent characterizations of sub-Gaussian random variables [32, Thm. 2.6].

Definition 3.9 (Sub-Gaussian random variable). *A random variable X is called σ-sub-Gaussian if*

$$P(X - \mathbb{E}[X] > \varepsilon) \leq e^{-\frac{\eta^2}{2\sigma^2}}. \tag{3.26}$$

An equivalent condition is that, for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq e^{\frac{\lambda^2\sigma^2}{2}}. \tag{3.27}$$

A third equivalent characterization is that, for all $\lambda \in [0,1)$,

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])^2/2\sigma^2}\right] \leq \frac{1}{\sqrt{1-\lambda}}. \tag{3.28}$$

A useful property of sub-Gaussian random variables is that the sub-Gaussianity parameter σ behaves like a variance under averaging: if we let S denote the average of n samples of X, then S is $\sigma/\sqrt{n}$-sub-Gaussian.

Proposition 3.1 (Averaging sub-Gaussian random variables). *Let X be a σ-sub-Gaussian random variable and let $S = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the average of n independent instances of X. Then, S is $\sigma/\sqrt{n}$-sub-Gaussian.*

3.3.2 Bounded Random Variables

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

We begin by observing that any bounded random variable is sub-Gaussian [32, Sec. 2.1.2].

Proposition 3.2 (Bounded random variables are sub-Gaussian)**.** *Let X be a random variable whose range is restricted to $[0,1]$. Then, X is $1/2$-sub-Gaussian.*

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

Definition 3.10 (Binary KL divergence)**.** *Let $p, q \in [0,1]$. Then $d(q \,||\, p)$ denotes the KL divergence between two Bernoulli random variables with parameters q and p respectively, that is,*

$$d(q \,||\, p) = D(\mathrm{Bern}(q) \,||\, \mathrm{Bern}(p)) \tag{3.29}$$

$$= q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p}. \tag{3.30}$$

A “relaxed” version of the binary KL divergence can be expressed as

$$d_\gamma(q \,||\, p) = \gamma q - \log(1 - p + pe^\gamma), \tag{3.31}$$

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

The binary KL divergence between a sample mean and its expectation can be shown to display a useful concentration behavior. The following result is due to [33].

Theorem 3.5 (KL concentration)**.** *Let X be a random variable with range $[0,1]$ and mean μ. Let S denote the mean of n independent draws of X. For $n \geq 8$,*

$$\mathbb{E}\left[e^{nd(S \,||\, \mu)}\right] \leq 2\sqrt{n}. \tag{3.32}$$

An even tighter concentration result can be derived by considering the aforementioned relaxed binary KL divergence. This turns out to be particularly useful in the derivation of average generalization bounds. The following result is due to [34].

Theorem 3.6 (Parametric KL concentration)**.** *Let X be a random variable with range $[0,1]$ and mean μ. Let S denote the mean of n independent draws of X. For any fixed γ,*

$$\mathbb{E}\left[e^{nd_\gamma(S \,||\, \mu)}\right] \leq 1. \tag{3.33}$$

3.3.3 Binary Random Variables

While we previously considered any bounded random variables within $[0,1]$, we now restrict our attention to binary random variables, which can only take two values within this range. For such random variables, a fast concentration result on the weighted difference between the random variable and its complement can be derived. While this may seem quite esoteric at first glance, it can be used to derive fast-rate generalization bounds with sharp constants for interpolating learning algorithms. The following is due to [8].

Theorem 3.7 (Concentration of complementary random variables)**.** *Let X be a random variable satisfying $P(X = a) = P(X = b) = 1/2$ where $a, b \in [0, 1]$. Let $\bar{X} = a + b - X$ denote its complement in the set $\{a, b\}$. Finally, let $\lambda, \gamma > 0$ be constants such that $\lambda(1 - \gamma) + (e^\lambda - 1 - \lambda)(1 + \gamma^2) \leq 0$. Then,*

$$\mathbb{E}\left[e^{\lambda\left(X - \gamma\bar{X}\right)}\right] \leq 1. \tag{3.34}$$

In the following chapter, as well as the appended research contributions in Part II, we will put the tools discussed in this chapter to use to derive generalization guarantees.

CHAPTER 4

Information-Theoretic Generalization Guarantees

In this chapter, we overview some of the information-theoretic generalization guarantees that are available in the literature. We start by motivating the need for new generalization guarantees, beyond the classical results discussed in Chapter 2, and discuss why the information-theoretic approach is a promising way forward. We then describe some of the main results that are available in the literature, beginning with average generalization bounds, before proceeding to tail bounds of the PAC-Bayesian and single-draw varieties. We conclude by presenting the CMI framework, where the training data is randomly selected from a larger set of data samples. This framework plays an important role in the appended papers.

4.1 Motivation

The celebrated fundamental theorem of statistical learning [5] shows that the VC dimension completely characterizes PAC learnability. However, this result has a uniform flavor: the guarantees hold for all hypotheses in the class, and for all possible data distributions.

In [6], two experiments are performed with deep neural networks for image classification tasks. In the first, the networks are trained on training sets with *true* labels. In this setting, the networks achieve zero training loss and a low test loss, meaning that they generalize. In the second experiment, the labels of the training set are *randomized*. Now, there is nothing to be learned from the training set, as the information carried by the correctly labelled pairs has been erased. Still, the networks are able to achieve zero training loss, but in this setting, their test loss is no better than random guessing—

they do not generalize. This experiment illustrates that, to explain generalization in modern machine learning algorithms, uniform results are not sufficient. Deep neural networks, which achieve the state-of-the-art results in a myriad of applications, operate and generalize in a regime that cannot be explained by their VC dimension. Indeed, networks whose VC dimension is estimated to be in the millions can generalize based on a few thousand training examples.

This motivates the need for new generalization guarantees. Unlike the classical results, we do not want to restrict ourselves to properties of the hypothesis class, and we want to be less uniform in some sense. In particular, we want to incorporate the data distribution and the learning algorithm into our bounds. The information-theoretic bounds that we present in this section do exactly this: if the algorithm or data distribution are altered, the generalization performance that is guaranteed by the bound will also change. Unlike the classical generalization guarantees, these information-theoretic results can thus distinguish between the settings with true and random labels that are studied in [6], providing hope that we can explain the discrepancy in generalization. The intuition behind this approach is as follows. If a learner achieves good performance on training data, but captures all of the information contained therein, it may simply have memorized the specific training samples, without identifying any generally useful structure. Thus, it may be unable to generalize. In contrast, if a learner performs well on training data, while extracting a low amount of information from it, it must have captured some fundamental relation in the data rather than just memorizing it. Hence, it will generalize to new data.

4.2 Average Generalization Bounds

We begin by looking at information-theoretic bounds on the average generalization error. Initial work on explicitly tying generalization guarantees to the mutual information, a core quantity within information theory, was performed by Russo and Zou [35]. Although the main focus of their investigation is on adaptive data analysis, the statements can be adapted to the learning setting, but only for finite data domains. Xu and Raginsky [36] extended this to uncountable domains, and highlighted the connection to learning. We present the main result from Xu and Raginsky [36, Thm. 1] below.

Theorem 4.1 (Average bound in terms of mutual information)**.** *Assume that $\ell(w, Z)$ is σ-sub-Gaussian under P_Z for all $w \in \mathcal{W}$ and that $P_{W\boldsymbol{Z}} \ll P_W P_{\boldsymbol{Z}}$. Then,*

$$\mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{P_Z}(W)] \leq \mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{\boldsymbol{Z}}(W)] + \sqrt{\frac{2\sigma^2 I(W; \boldsymbol{Z})}{n}}. \tag{4.1}$$

Proof. We begin by applying the Donsker-Varadhan variational representation (3.25) with $P_{W\boldsymbol{Z}}$, $Q = P_W P_{\boldsymbol{Z}}$ and $f(X) = \lambda(L_{P_Z}(W) - L_{\boldsymbol{Z}}(W))$, to see that for all λ,

$$\mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{P_Z}(W) - L_{\boldsymbol{Z}}(W)] \leq \frac{D(P_{W\boldsymbol{Z}} \,||\, P_W P_{\boldsymbol{Z}}) + \log \mathbb{E}_{P_W P_{\boldsymbol{Z}}}\left[e^{\lambda(L_{P_Z}(W) - L_{\boldsymbol{Z}}(W))}\right]}{\lambda}. \tag{4.2}$$

This allows us to change measure from $P_{W\boldsymbol{Z}}$ to $P_W P_{\boldsymbol{Z}}$. Next, we need to use a concentration inequality to obtain a bound that decays with n. Note that for a fixed w, $L_{\boldsymbol{Z}}(w)$ is the average of n independent σ-sub-Gaussian random variables. Therefore, by Proposition 3.1, $L_{\boldsymbol{Z}}(w)$ is $\sigma/\sqrt{n}$-sub-Gaussian. Therefore, by (3.27),

$$\log \mathbb{E}_{P_W P_{\boldsymbol{Z}}}\left[e^{\lambda(L_{P_Z}(W)-L_{\boldsymbol{Z}}(W))}\right] \leq \frac{\lambda^2\sigma^2}{2n}. \tag{4.3}$$

The desired result follows after combining (4.2)-(4.3) and optimizing over λ. □

Many of the later results in this chapter follow by similar proofs, but with the change of measure step and concentration inequality substituted for suitable alternatives. For the tail bounds, a step involving Markov's inequality is also required, and in some cases there are additional technical subtleties. In the remainder of this chapter, we will only give brief descriptions of the tools used for each proof. For many of the bounds we present, detailed derivations are provided in Part II.

The big advantage of generalization guarantees based on information measures like the mutual information when compared to, for instance, the one based on the VC dimension, is that it takes into account the learning algorithm. As an extreme case, consider a learning algorithm that picks the hypothesis W independently of the training data $\boldsymbol{Z}$. Then, the mutual information $I(W;\boldsymbol{Z})$ will be 0, and we are guaranteed to generalize in expectation even if the hypothesis is selected from a class with infinite VC dimension. Of course, this specific learner is not very interesting. A discussion of more relevant scenarios where the mutual information can be bounded, such as noisy empirical risk minimization, use of stable algorithms, and compression schemes can be found in [36].

As discussed in [36], an analogy can be drawn between learning and communication by identifying the learning algorithm $P_{W|\boldsymbol{Z}}$ with a channel. In communication applications of information theory, the mutual information between the input and output, when supremized over input distributions, is the channel capacity. The channel capacity is the maximum rate at which information can be reliably transmitted over a communication channel [19, Chap. 19]. Thus, Theorem 4.1 shows that the capacity of a learning algorithm provides an upper bound on generalization error, in a worst-case sense with respect to the data distribution. However, this view eliminates the data-dependence inherent in Theorem 4.1, and thus leads to a less tight characterization.

A drawback of bounds expressed in terms of the mutual information is that they can often be unbounded. For instance, if W is a deterministic function of $\boldsymbol{Z}$ and both are separately continuous random variables, the mutual information will be infinite, even if generalization can be guaranteed through, for instance, the VC dimension bound. This issue was alleviated by Bu and Veeravalli [37], who used the methods of Xu and Raginsky to derive a generalization guarantee in terms of the samplewise mutual informatia, that is, $I(W;Z_i)$ for $i=1,\dots,n$. Since W is typically undetermined given any individual Z_i, even when it is a deterministic function of the whole training set $\boldsymbol{Z}$, this leads to a finite bound in situations where the original mutual information-based bound fails. We present this result below.

Theorem 4.2 (Average bound in terms of samplewise mutual information)**.** *Assume that $\ell(w, Z)$ is σ-sub-Gaussian under P_Z for all w and that $P_{W|\boldsymbol{Z}} \ll P_W$. Then,*

$$\mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{P_Z}(W)] \leq \mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{\boldsymbol{Z}}(W)] + \frac{1}{n}\sum_{i=1}^{n} \sqrt{2\sigma^2 I(W; Z_i)}. \tag{4.4}$$

This result relies on the decomposition

$$\mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{P_Z}(W)] - \mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{\boldsymbol{Z}}(W)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{P_W P_Z}[\ell(W, Z)] - \mathbb{E}_{P_{WZ_i}}[\ell(W, Z_i)]\,. \tag{4.5}$$

Applying the same arguments as were used to prove Theorem 4.1 to each term in this composition, we obtain the desired result. By using Jensen's inequality, the chain rule of mutual information, and the independence of the Z_i, we see that the samplewise mutual information guarantee is always tighter than the original mutual information result [37, Prop. 1].

One point to note is that the decomposition above can be seen as weighting all training samples with a uniform distribution. However, this may not always be the best approach. For instance, for sequential algorithms like stochastic gradient descent, the training samples that are processed first may have an outsized influence on the selected hypothesis, even if the training loss is similar for all samples. Using a non-uniform weighting of the training samples yields the following result.

Proposition 4.1 (Average bound in terms of weighted samplewise mutual information)**.** *Assume that $\ell(w, Z)$ is σ-sub-Gaussian under P_Z for all w and that $P_{W|\boldsymbol{Z}} \ll P_W$. Let I be distributed according P_I, where P_I is an arbitrary probability mass function on $\{1, \dots, n\}$. Then,*

$$\mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{P_Z}(W)] \leq \mathbb{E}_{P_I}\left[\mathbb{E}_{P_{WZ_i}}[\ell(W, Z_i)]\right] + \mathbb{E}_{P_I}\left[\sqrt{2\sigma^2 I(W; Z_I)}\right]. \tag{4.6}$$

The bound above can be applied to, for instance, a generalized notion of compression schemes. A compression scheme of size k is a learning algorithm where the output based on the full training set, consisting of $n > k$ samples, is always identical to the output based on some size-k subset of the training set [5, Chapter 30]. Thus, for these algorithms, only k samples affect the output. The main insight behind generalization bounds for compression schemes is that the remaining $n - k$ samples do not affect the selected hypothesis, and thus serve as an independent test set. However, the requirement that these $n - k$ samples are completely independent from the selected hypothesis may be too strict. Using Proposition 4.1, we can instead consider learning algorithms that depend freely on k input samples, but where the remaining $n - k$ samples have bounded samplewise mutual information with the output.

An alternative approach to get information-theoretic bounds that are always finite is to use the CMI framework, which we present at the end of this chapter.

4.3 PAC-Bayesian Generalization Bounds

The genesis of information-theoretic approaches to generalization guarantees can be found within the PAC-Bayesian literature. While the initial ideas can be glimpsed in [38], the PAC-Bayesian approach is usually said to have started with McAllester [14], who worked on developing PAC-style bounds for classifiers of a Bayesian flavor. These bounds typically rely on a divergence, such as the KL divergence, between a *posterior* $P_{W|\boldsymbol{Z}}$, i.e., the output distribution from the learning algorithm, and some *prior* Q_W, which has to be independent of $\boldsymbol{Z}$. Philosophically, this prior reflects some belief about which hypotheses are seen as reasonable before any data is seen. While the usage of the terms prior and posterior do not exactly match their original meanings in a Bayesian sense, we will use them for historical reasons. Since the advent of the PAC-Bayesian approach, research output in the field has been torrential. Despite the name, the approach applies not only to Bayesian classifiers, but to a large class of learning algorithms, both deterministic and randomized. Furthermore, the results are often amenable to numerical evaluation, and can also provide new insights into algorithm design by way of regularization methods. The PAC-Bayesian framework also allows for several extensions, where results can be adapted to new settings or strengthened for certain learning problems [15, 34, 39, 40].

Below, we give a somewhat more modern version of the basic PAC-Bayesian bound [14] for sub-Gaussian losses.

Theorem 4.3 (PAC-Bayesian bound for sub-Gaussian losses)**.** *Assume that $\ell(w, Z)$ is σ-sub-Gaussian under P_Z for all $w \in \mathcal{W}$ and let Q_W be some distribution on $\mathcal{W}$ that satisfies $P_{W|\boldsymbol{Z}} \ll Q_W$. Then, with probability at least $1-\delta$ under $P_{\boldsymbol{Z}}$,*

$$\mathbb{E}_{P_{W|\boldsymbol{Z}}}[L_{P_Z}(W)] \leq \mathbb{E}_{P_{W|\boldsymbol{Z}}}[L_{\boldsymbol{Z}}(W)] + \sqrt{\frac{2\sigma^2}{n-1}\left(D(P_{W|\boldsymbol{Z}} \,||\, Q_W) + \log\frac{\sqrt{2n}}{\delta}\right)}. \tag{4.7}$$

The proof of this result is similar to that of Theorem 4.1, but includes an extra Markov step and relies on (3.28) rather than (3.27). The reason for this difference is that the optimization over λ would not be possible due to the probabilistic nature of the bound. This is presented in more detail in Contribution A in Part II.

As observed by Bassily et al. [41], the PAC-Bayesian bound in (4.7) can be converted into a bound in terms of mutual information, by selecting the prior Q_W to be the marginal P_W and using Markov's inequality. The price to pay for this conversion is a highly undesirable linear dependence on $1/\delta$. Of course, the same conversion can be performed in the rest of the PAC-Bayesian bounds that we present.

Theorem 4.4 (PAC-Bayesian bound in terms of mutual information)**.** *Assume that $\ell(w, Z)$ is σ-sub-Gaussian under P_Z for all $w \in \mathcal{W}$ and that $P_{W|\boldsymbol{Z}} \ll P_W$. Then,*

$$\mathbb{E}_{P_{W|\boldsymbol{Z}}}[L_{P_Z}(W)] \leq \mathbb{E}_{P_{W|\boldsymbol{Z}}}[L_{\boldsymbol{Z}}(W)] + \sqrt{\frac{2\sigma^2}{n}\left(\frac{2I(W;\boldsymbol{Z})}{\delta} + \log\frac{2}{\delta}\right)}. \tag{4.8}$$

Note that the improved dependence on n in Theorem 4.4 compared to Theorem 4.3 can be achieved since the right-hand side is no longer data-dependent. This allows us to use (3.27) in the derivation instead of (3.28).

We note that the dependence on n in (4.7) is[1] $\sqrt{D(P_{W|\boldsymbol{Z}} \,||\, Q_W)/n}$. We will refer to this as a *slow rate*. For classification settings, it is typical to use the accuracy as the loss function. For the bound in (4.7) to be interesting, the square-root term must be smaller than one. It is therefore in our interest to rid ourselves of the square root, since this would yield a tighter bound. This is done in the following result [34], but at the cost of worse multiplicative constants. We will refer to it as a *fast-rate* bound. However, we note that in order for the bound to achieve a fast-rate in the most commonly used sense [16,17], the KL divergence $D(P_{W|\boldsymbol{Z}} \,||\, Q_W)$ must grow at most polylogarithmically in n. For these fast-rate results, we require the loss function to be bounded.

Theorem 4.5 (Fast-rate PAC-Bayesian bound)**.** *Assume that the loss is bounded to* $[0,1]$*. For all* $\lambda \in (0,1)$*, the following holds with probability at least* $1-\delta$ *under* $P_{\boldsymbol{Z}}$*:*

$$\mathbb{E}_{P_{W|\boldsymbol{Z}}}[L_{P_Z}(W)] \leq \frac{1}{\lambda}\left[\mathbb{E}_{P_{W|\boldsymbol{Z}}}[L_{\boldsymbol{Z}}(W)] + \frac{D(P_{W|\boldsymbol{Z}} \,||\, Q_W) + \log\frac{1}{\delta}}{2(1-\lambda)n}\right]. \tag{4.9}$$

The derivation of this bound relies on the concentration inequality in Theorem 3.6, which gives a result in terms of the relaxed binary KL divergence from Definition 3.10. When suitably weakened, this gives the fast-rate bound above.

By considering the binary KL divergence between the training and population loss, one can obtain a bound whose rate interpolates between the fast and slow rate, depending on the value of the training loss. This bound was initially developed by Langford and Seeger [42], and later strengthened by Maurer [33]. It is sometimes referred to as Seeger's bound or the Maurer-Langford-Seeger bound [43,44]. It is derived on the basis of Theorem 3.5.

Theorem 4.6 (Binary KL PAC-Bayesian bound)**.** *Assume that the loss is bounded to* $[0,1]$*. With probability at least* $1-\delta$ *under* $P_{\boldsymbol{Z}}$*,*

$$d(L_{\boldsymbol{Z}}(W) \,||\, L_{P_Z}(W)) \leq \frac{D(P_{W|\boldsymbol{Z}} \,||\, Q_W) + \log\frac{2\sqrt{n}}{\delta}}{2n}. \tag{4.10}$$

If the training loss is assumed to be zero, a fast-rate bound can be obtained based on the above by a straight-forward calculation. In general, Pinsker's inequality can be used to obtain an explicit slow-rate bound. However, for any loss value, the bound can be efficiently numerically inverted to find the maximum value for the population loss that satisfies the bound, given values of n, δ, the training loss, and the KL divergence. Hence, the bound in Theorem 4.6 interpolates between the fast and slow rates.

[1] The dependence of $P_{W|\boldsymbol{Z}}$ on n is implicit, since the learning algorithm has a fixed definition only for a given sample size, and in principle, it is allowed to have different behaviors for different sample sizes.

4.4 Single-Draw Generalization Bounds

In [36, Thm. 3], a single-draw generalization bound in terms of mutual information is also derived, through the use of the *monitor technique*. Bassily et al. [41] also derive such a single-draw bound, but obtain better constants.

Theorem 4.7 (Single-draw bound in terms of mutual information)**.** *Assume that $\ell(w, Z)$ is σ-sub-Gaussian under P_Z for all $w \in \mathcal{W}$ and that $P_{W|\boldsymbol{Z}} \ll P_W$. Then, with probability at least $1-\delta$ under $P_{W\boldsymbol{Z}}$,*

$$L_{P_Z}(W) \leq L_{\boldsymbol{Z}}(W) + \sqrt{\frac{2\sigma^2}{n}\left(\frac{I(W;\boldsymbol{Z}) + H_b(\delta)}{\delta}\right)}. \tag{4.11}$$

Proof. For a pair of probability distributions P_X and Q_X on a common space $\mathcal{X}$ and a measurable event $E \subset \mathcal{X}$, let $p = P[E]$ and $q = Q[E]$ denote the probability of the event under the respective distributions. Let $H_b(p)$ denote the entropy of a Bernoulli random variable with parameter p. Then, the data processing inequality for the KL divergence implies that

$$D(P\,||\,Q) \geq d(p\,||\,q) \geq -H_b(p) + p\log\frac{1}{q}. \tag{4.12}$$

Here, $d(p\,||\,q)$ denotes the KL divergence between two Bernoulli distributions with parameters p and q respectively, while $H_b(p) = -p\log(p) - (1-p)\log(1-p)$ is the entropy of a Bernoulli random variable with parameter p. We now set $P = P_{W\boldsymbol{Z}}$, $Q = P_W P_{\boldsymbol{Z}}$ and take $\mathcal{E}$ to be the high-error event

$$\mathcal{E} = \{(w, \boldsymbol{z}) : L_{P_Z}(w) - L_{\boldsymbol{z}}(w) > \epsilon\}. \tag{4.13}$$

The σ-sub-Gaussianity of the loss function implies that [32, Eq. (2.9)]

$$P_{Z^n}[\mathcal{E} > \epsilon] \leq \exp\bigl(-n\epsilon^2/(2\sigma^2)\bigr)\,. \tag{4.14}$$

From this, it follows that

$$\log\frac{1}{q} \geq n\frac{\epsilon^2}{2\sigma^2} \tag{4.15}$$

which, substituted into (4.12), gives us

$$\epsilon \leq \sqrt{\frac{2\sigma^2}{n}\left(\frac{I(W;Z^n) + H_b(p)}{p}\right)}. \tag{4.16}$$

Since the right-hand side of (4.16) is monotonically decreasing in p, we conclude that the condition

$$\epsilon \geq \sqrt{\frac{2\sigma^2}{n}\left(\frac{I(W;Z^n) + H_b(\delta)}{\delta}\right)} \tag{4.17}$$

implies that $p \leq \delta$. □

As previously mentioned, the tail bounds in terms of mutual information display an undesirable linear dependence on the inverse confidence parameter $1/\delta$. Esposito et al. [9] sought to rectify this by introducing new single-draw bounds in terms of a large family of alternative information-theoretic quantities. Below, we present their bound given in terms of the α-mutual information $I_\alpha(W;\boldsymbol{Z})$.

Theorem 4.8 (Single-draw bound in terms of α-mutual information)**.** *Assume that $\ell(w,Z)$ is σ-sub-Gaussian under P_Z for all $w \in \mathcal{W}$ and that $P_{W|\boldsymbol{Z}} \ll P_W$. Then, for all $\alpha > 1$, with probability at least $1-\delta$ under $P_{W\boldsymbol{Z}}$,*

$$L_{P_Z}(W) \le L_{\boldsymbol{Z}}(W) + \sqrt{\frac{2\sigma^2}{n}\left[I_\alpha(W;\boldsymbol{Z}) + \frac{\alpha}{\alpha-1}\log\frac{1}{\delta}\right]}. \tag{4.18}$$

Here, $I_\alpha(W;Z^n)$ is the α-mutual information

$$I_\alpha(W;Z^n) = \frac{\alpha}{\alpha-1}\log \mathbb{E}_{P_{Z^n}}\left[\mathbb{E}_{P_W}^{1/\alpha}\left[\left(\frac{\mathrm{d}P_{WZ^n}}{\mathrm{d}P_W P_{Z^n}}\right)^{\alpha}\right]\right]. \tag{4.19}$$

The proof of this result relies on repeated uses of Hölder's inequality, combined with a use of Hoeffding's inequality. A similar proof technique can be found in Theorem 7 and Corollary 9 in Paper A.

For a fixed α, we see that the bound achieves a much more beneficial $\log 1/\delta$ dependence on the inverse confidence parameter. Bounds with this logarithmic dependence on $1/\delta$ are sometimes called *high-probability* bounds.[2] In particular, if $\alpha \to \infty$, the constant disappears and the α-mutual information becomes the maximal leakage from Definition 3.6. However, in the limit of $\alpha \to 1$, where the α-mutual information becomes the normal mutual information, we see that the δ-dependent term blows up, rendering the bound completely vacuous. We thus see that there is some kind of trade-off between the value of α and the contribution of the δ-dependent term. In Contribution A, we explore this trade-off further, laying bare a connection between the moment of the information measure under consideration and the effect that δ has on the tightness of the bound.

4.5 The CMI framework

As previously mentioned, an alternative method to obtain information-theoretic bounds that are guaranteed to be finite is to use the CMI framework. The intuitive motivation behind this approach is that we want to normalize the information carried by each sample to 1 bit. It was introduced by Steinke and Zakynthinou [8]. In the CMI framework, we have $2n$ training samples $\tilde{\boldsymbol{Z}} = (\tilde{Z}_1, \dots, \tilde{Z}_{2n})$, referred to as a *supersample.* From

[2] This terminological distinction between tail bounds with a logarithmic dependence on $1/\delta$ and bounds with a polynomial dependence on $1/\delta$ is not universal. In some works, all tail bounds are referred to as high-probability bounds.

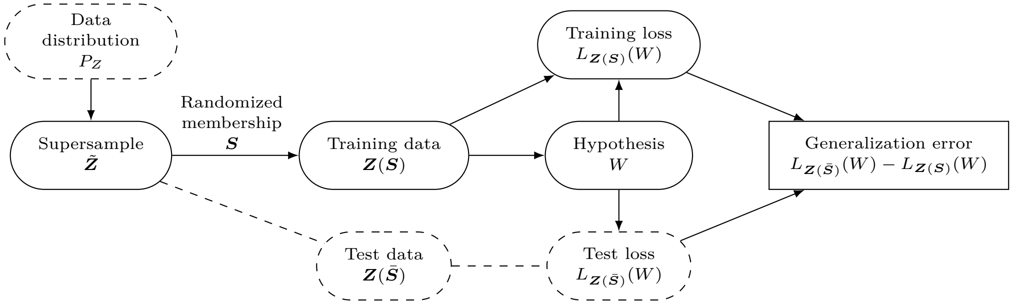
Data distribution P_Z
Supersample $\tilde{\boldsymbol{Z}}$
Randomized membership $\boldsymbol{S}$
Training data $\boldsymbol{Z}(\boldsymbol{S})$
Training loss $L_{\boldsymbol{Z}(\boldsymbol{S})}(W)$
Hypothesis W
Generalization error $L_{\boldsymbol{Z}(\bar{\boldsymbol{S}})}(W) - L_{\boldsymbol{Z}(\boldsymbol{S})}(W)$
Test data $\boldsymbol{Z}(\bar{\boldsymbol{S}})$
Test loss $L_{\boldsymbol{Z}(\bar{\boldsymbol{S}})}(W)$

Figure 4.1: Schematic illustration of the CMI framework.

this, the training set is randomly formed as follows: let $\boldsymbol{S} = (S_1, \dots, S_n)$ be a random vector, where each entry is distributed according to a Bernoulli distribution with parameter 1/2. Then, the ith element of the training set $\boldsymbol{Z}(\boldsymbol{S}) = (Z_1(S_1), \dots, Z_n(S_n))$ is given by $Z_i(S_i) = \tilde{Z}_{i+S_i n}$. In other words, the ith element of the training set can be one of the two elements $\tilde{Z}_i$ or $\tilde{Z}_{i+n}$ from $\tilde{\boldsymbol{Z}}$, and the selection between these two is determined by S_i. The hypothesis W is then chosen based on $\boldsymbol{Z}(\boldsymbol{S})$, and is conditionally independent of $\tilde{\boldsymbol{Z}}$ and $\boldsymbol{S}$ given $\boldsymbol{Z}(\boldsymbol{S})$. We let $\bar{\boldsymbol{S}} = (1 - S_1, \dots, 1 - S_n)$ denote the modulo-2 complement of $\boldsymbol{S}$. Note that $\boldsymbol{Z}(\bar{\boldsymbol{S}})$ is independent of W, and hence is a test set. The loss on this set, $L_{\boldsymbol{Z}(\bar{\boldsymbol{S}})}(W)$, is an unbiased estimate of the population loss under $P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}$, the joint distribution of W, $\tilde{\boldsymbol{Z}}$, and $\boldsymbol{S}$. This setup is illustrated in Figure 4.1.

For this setup, under the additional assumption of a bounded loss function, Steinke and Zakynthinou derived an average bound on the generalization error that is similar to that of Xu and Raginsky [36, Thm. 1], but given in terms of the *conditional* mutual information $I(W; \boldsymbol{S}|\tilde{\boldsymbol{Z}})$. We present this result below.

Theorem 4.9 (Slow-rate bound in terms of CMI)**.** *Assume that $\ell(w, z) \in [0, 1]$ for all $w \in \mathcal{W}$ and $z \in \mathcal{Z}$. Then,*

$$\mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}[L_{P_Z}(W)] \le \mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[L_{\boldsymbol{Z}(\boldsymbol{S})}(W)\right] + \sqrt{\frac{2I(W; \boldsymbol{S}|\tilde{\boldsymbol{Z}})}{n}}. \tag{4.20}$$

The proof of this result again relies on the Donsker-Varadhan variational representation of KL divergence. An alternative proof can be found in Corollary 5 in Paper A.

Intuitively, the result in Theorem 4.9 improves upon Theorem 4.1, for the special case of a bounded loss function, because the information of each sample is normalized to 1 bit—indeed, the CMI can be upper-bounded as $I(W; \boldsymbol{S}|\tilde{\boldsymbol{Z}}) \le H(\boldsymbol{S}) = n \log 2$. By the chain rule of mutual information, combined with the Markov property $(\tilde{\boldsymbol{Z}}, \boldsymbol{S}) - Z(\boldsymbol{S}) - W$ and that $\boldsymbol{Z}(\boldsymbol{S})$ is a deterministic function of $(\tilde{\boldsymbol{Z}}, \boldsymbol{S})$, we also have $I(W; \boldsymbol{Z}(\boldsymbol{S})) = I(W; \tilde{\boldsymbol{Z}}) + I(W; \boldsymbol{S}|\tilde{\boldsymbol{Z}})$. Thus, a direct comparison between Theorem 4.9 and Theorem 4.1 reveals that the former is tighter provided that $I(W; \tilde{\boldsymbol{Z}}) > 3I(W; \boldsymbol{S}|\tilde{\boldsymbol{Z}})$.

As previously mentioned, information-theoretic generalization guarantees can have slow rates, where the dependence on n is $\sqrt{\mathrm{IM}/n}$, where IM is shorthand for some information measure, or fast rates, where the dependence is IM/n. In [8, Cor. 5(3)], Steinke and Zakynthinou also derive a bound with such a fast rate, at the expense of less beneficial multiplicative constants. In particular, the training loss is multiplied by a factor greater than one.

Theorem 4.10 (Fast-rate bound in terms of CMI)**.** *Assume that $\ell(w,z) \in [0,1]$ for all $w \in \mathcal{W}$ and $z \in \mathcal{Z}$. Then,*

$$\mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}[L_{P_Z}(W)] \leq 2\,\mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[L_{\boldsymbol{Z}(\boldsymbol{S})}(W)\right] + \frac{3I(W;\boldsymbol{S}|\tilde{\boldsymbol{Z}})}{n}. \tag{4.21}$$

To achieve a fast rate, Theorem 3.7 is used, where the boundedness of the loss function is used more directly than in the derivation of the slow-rate bound in Theorem 4.9. For learning algorithms that achieve zero training loss, a fast-rate bound with sharp constants is also derived in [8].

Similar to the samplewise extension of the average mutual information bound performed by Bu and Veeravalli [37], Haghifam et al. [45, Thm. 3.4] extended the CMI result to a samplewise CMI bound, using the same decomposition as in Theorem 4.2. They also use the disintegration ideas introduced in [46] to pull the expectation over $P_{\tilde{\boldsymbol{Z}}}$ outside of the square root, which tightens the resulting bound.

Theorem 4.11 (Slow-rate bound in terms of samplewise CMI)**.** *Assume that $\ell(w,z) \in [0,1]$ for all $w \in \mathcal{W}$ and $z \in \mathcal{Z}$. Then,*

$$\mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}[L_{P_Z}(W)] \leq \mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[L_{\boldsymbol{Z}(\boldsymbol{S})}(W)\right] + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{P_{\tilde{\boldsymbol{Z}}}}\left[\sqrt{2D(P_{WS_i|\tilde{\boldsymbol{Z}}}\,||\,P_{W|\tilde{\boldsymbol{Z}}}P_{S_i})}\right]. \tag{4.22}$$

Again, Jensen's inequality, the chain rule of mutual information, and the independence between the S_i implies that this bound is stronger than the CMI bound in Theorem 4.9.

Finally, a tighter characterization can be obtained by considering the *evaluated* CMI. In the evaluated CMI, the hypothesis is replaced with the losses that the hypothesis induces on the supersample [8]. We denote these losses by λ, so that for $i = 1, \dots, n$ and $j = 0, 1$, we have $\lambda_{i,j} = \ell(W, \tilde{\boldsymbol{Z}}_{i+jn})$. The derivation of bounds in terms of the evaluated CMI rely on the observation that the hypothesis enters the proof only through the losses it induces. While this idea can be used in all of the aforementioned CMI bounds, we present it below only for the slow-rate bound.

Theorem 4.12 (Slow-rate bound in terms of evaluated CMI)**.** *Assume that $\ell(w,z) \in [0,1]$ for all $w \in \mathcal{W}$ and $z \in \mathcal{Z}$. Then,*

$$\mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}[L_{P_Z}(W)] \leq \mathbb{E}_{P_{W\tilde{\boldsymbol{Z}}\boldsymbol{S}}}\left[L_{\boldsymbol{Z}(\boldsymbol{S})}(W)\right] + \sqrt{\frac{2I(\lambda;\boldsymbol{S}|\tilde{\boldsymbol{Z}})}{n}}. \tag{4.23}$$

Similar results can also be derived on the basis of the *predictions* that the hypothesis induces on the supersample [47]. The expressiveness of the CMI and evaluated CMI has been studied in terms of the VC dimension and related quantities in [48], which proves several positive results for the ability of the framework to capture classical generalization guarantees.

CHAPTER 5

Applications and Extensions

In the previous chapter, we covered several information-theoretic generalization bounds. These results were stated for the generic learning setting introduced in Section 2.1, and we provided no explicit characterization of the information measures that appeared in the bounds. However, the usefulness of the bounds depend on the behavior of the information measure. For instance, consider the bound in Theorem 4.9, which depends on the quotient $I(W;\boldsymbol{S}|\tilde{\boldsymbol{Z}})/n$. The conditional mutual information in the numerator has an implicit dependence on n, and in the worst case, it may grow linearly with n. If this is the case, the bound is non-decreasing as the sample size grows, and will typically be vacuous. Hence, evaluating the information terms for specific learning algorithms, sometimes referred to as the information complexity of an algorithm, is of high importance. Furthermore, the generic setup from Section 2.1 does not cover all settings of interest, as previously discussed.

In this thesis, we numerically evaluate our information-theoretic generalization bounds for neural networks. We also consider meta learning, which is an extension that goes beyond the learning setup from Section 2.1. In this chapter, we provide a brief overview of generalization results for neural networks and meta learning, both in terms of classical generalization bounds and their information-theoretic counterpart.

5.1 Neural Networks

Neural networks are parametric models that can represent highly complex functions through the composition of several simple operations. A very simple neural network of depth d consists of an activation function f and a set of matrices $W_1, \dots, W_d$. The output of the network for an input x is given by $N(x) = W_d f(W_{d-1} f(W_{d-2} \dots f(W_1 x) \dots))$, where the activation function is applied elementwise and the size of all matrices are such that the output is well-defined. The values of the matrices are updated by performing gradient descent on the training loss. This basic structure has been extended in several ways, such as with convolutional neural networks [49] and transformers [50]. In general, the number of parameters of a neural network are much greater than the number of training samples, and the resulting hypothesis class is highly complex.

As mentioned in the previous chapter, one motivation for studying new types of generalization guarantees, beyond the classical ones, is that the performance of modern machine learning algorithms, such as deep neural networks, cannot be explained by bounds that rely on the complexity of the model class, such as those based on the VC dimension. New bounds need to exploit properties of the data distribution and learning algorithm, which makes information-theoretic approaches a good candidate. In this section, we survey some success stories where information-theoretic generalization guarantees have been applied to neural networks.

In [39], Dziugaite and Roy considered a stochastic neural network, the weights of which are drawn from a Gaussian distribution for each new prediction that the network makes. The mean and variance of this distribution were found by optimizing a PAC-Bayesian bound similar to the one in Theorem 4.3 using stochastic gradient descent. We thus note that, in this setup, the generalization bound is directly optimized as part of the neural network training procedure. The mean of the prior is chosen to be the random initialization of the neural network, and is independent of any data. This lead to nonvacuous bounds for overparameterized neural networks trained on a binary version of the MNIST data set, where the digits 0 to 4 were combined into one class and 5 to 9 into another.

By exploiting the compressibility of neural networks, Zhou et al. [40] derived a PAC-Bayesian bound that applies to deterministic, pruned networks. To obtain such a network, one first trains a large neural network, before removing parameters that do not affect performance too much. Through this process, one ends up with a similarly well-performing network, the size of which is significantly smaller than the original network size. An impressive aspect of [40] is that a nonvacuous generalization guarantee is obtained even for ImageNet, a relatively challenging setup. However, the bounds obtained are far from tight, even for the simpler MNIST data set, and do not apply to networks trained through a standard procedure.

Negrea et al. [46] applied their disintegrated, samplewise mutual information bound to noisy iterative optimizers, and in particular, provided numerically nonvacuous results for neural networks trained through stochastic gradient Langevin dynamics. This results in bounds on the average generalization error.

In [51], Dziugaite et al. improved upon their previous results by employing a strategy that allows them to construct the prior in a data-dependent fashion. Specifically, they evaluate the PAC-Bayesian bound in Theorem 4.5 using only part of the training data, while still using the full set of training data for choosing the posterior. Leaving part of the training data out when evaluating the bound allows for the prior to be chosen on the basis of the held-out data. This procedure yields relatively accurate bounds when applied to stochastic networks using normal stochastic gradient descent, and an even tighter characterization when the networks are trained by optimizing the bound directly.

Recently, Harutyunyan et al. [47] used the CMI framework to obtain generalization bounds for neural networks. Specifically, they used a variant of the samplewise, disintegrated bound in Theorem 4.11, where the CMI is replaced with the so-called f-CMI, i.e., the CMI based on the predictions that the hypothesis induces on the supersample. In order to estimate the f-CMI, they sample a subset of the available training data to form a supersample, and then sample half of this supersample to form the training set. This procedure is repeated several times. This yields an average generalization bound, unlike the PAC-Bayesian bounds previously mentioned in this section. Notably, the resulting bound is for the deterministic network found by stochastic gradient descent, and no noise needs to be added. A benefit of the f-CMI bound is that the bound remains stable throughout training. This is in contrast to weight-based bounds, which tend to diverge as training progresses.

Thus, the information-theoretic approach has proven to be a promising direction for the study of generalization in modern machine learning algorithms. However, there is still much work to be done. First, there are several symmetries and sources of noise inherent to neural networks and their training. Systematically exploiting this may be a path toward improving the bounds. Furthermore, the results obtained so far do not provide many guidelines regarding network design. A long term goal of the study of generalization would be to be able to predict a priori what design choices lead to a better performing network. As things currently stand, a lot of resources are spent on performing grid searches over hyperparameters to find well-generalizing networks, and many design choices are purely heuristic. A well-developed theory that satisfactorily explains generalization in neural networks should be able to provide more rigorously motivated choices for these parameters, and enable us to find well-performing networks without spending huge computational resources.

5.2 Meta Learning

In the standard supervised learning setup, each task is viewed in isolation, and the selected hypothesis depends only on data from this task. In practice, however, tasks are often related. For instance, a classification task where the goal is to identify cats or dogs is similar to a task where the goal is to identify tigers or wolves. In meta learning, the objective is to use data from related tasks to improve performance on a new, related task [52]. The term is sometimes used interchangeably with transfer learning, although meta learning often refers to learning hyperparameters whereas transfer learning is often

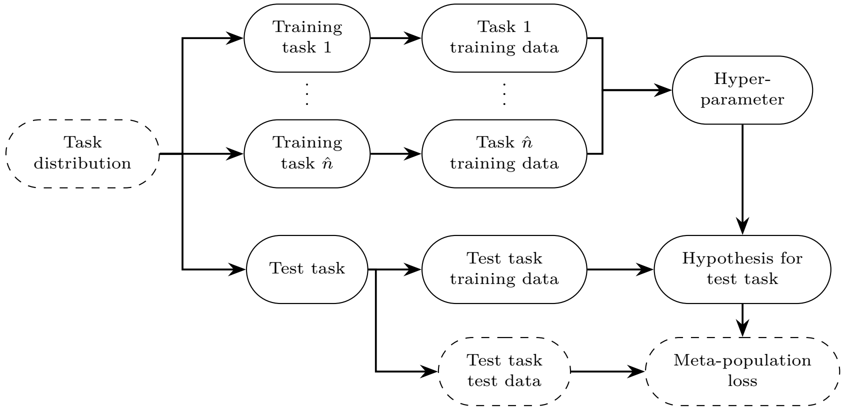

Figure 5.1: Schematic illustration of the general meta-learning setting.

about fine-tuning the parameters themselves. When applying neural networks to image classification tasks, such as those described above, there are several factors that can be learned from related tasks that can improve performance on a specific classification task. This includes the architecture, initialization, optimizer, learning rate, and the parameters of some matrices.

Formally, one considers a task space with an associated task distribution. For each task, there is an associated in-task data distribution. The meta-training set is assumed to be formed by first sampling $\hat{n}$ tasks according to the task distribution, and then drawing n samples within each task from the in-task data distributions. Within each task, a base learner selects a hypothesis on the basis of the in-task training data and a hyperparameter. The meta-learning algorithm has access to the entire meta-training set, and its objective is to select a suitable hyperparameter. The meta-training loss is the loss on the meta-training set. The goal is to have a small meta-population loss, i.e., a small population loss on new tasks. To compute this, a new task is drawn according to the task distribution, and a training set is drawn according to the corresponding in-task data distribution. The meta-population loss, then, is the expected loss on a new sample drawn from the in-task data distribution. This is illustrated in Figure 5.1.

A specific instance of meta learning that is well-studied is representation learning. In this setting, the goal of the meta learner is to select a representation h from a function class $\mathcal{H}$, while the base learner for task i selects a task-specific function f_i from the function class $\mathcal{F}$. The representation is shared among the tasks, and the hypothesis for task i is the composition $f_i \circ h$. This is similar to how neural networks are sometimes first fully trained on a set of tasks, and then the final layer is fine-tuned on a target task.

As indicated by the discussion above, one intuitively expects to obtain some benefit from using data from related tasks. This benefit was first theoretically demonstrated by [52], where the notion of task environment was also formally introduced. For the

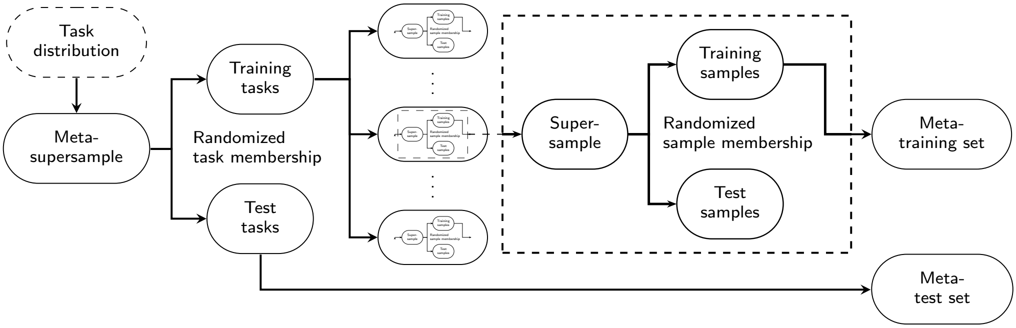

Figure 5.2: Schematic illustration of the CMI framework for meta learning.

representation learning setting above, [53, Thm. 5] derived a generalization bound that scales as $\sqrt{\mathcal{C}(\mathcal{H})/\hat{n}}+\sqrt{\mathcal{C}(\mathcal{F})/n}$, where $\mathcal{C}(\,\cdot\,)$ denotes a complexity measure of the function class and we ignore logarithmic terms. This illustrates a potential benefit over conventional, single-task learning, where the techniques described in the previous chapter would yield a bound that scales as $\sqrt{\mathcal{C}(\mathcal{F}\times\mathcal{H})/n}$, since both h and f_i need to be learned based only on the n samples from task i. However, as $n\to\infty$, the bound in [53, Thm. 5] still does not approach zero. This was rectified in [54], who derived a bound with a scaling of $\sqrt{\mathcal{C}(\mathcal{H})/(n\hat{n})+\mathcal{C}(\mathcal{F})/n}$. This captures the intuitive notion that each sample provides information about the representation h.

A main focus of [54] is to obtain excess risk bounds for a specified target task, that is, a bound on the difference between the population loss achieved by an empirical risk minimizer and the smallest possible population loss for a specific task. To achieve this, they use an assumption of *task diversity*, which essentially states that, for any representation h, the excess risk of the empirical risk minimizer averaged over the training tasks is not too far from the greatest possible excess risk of the empirical risk minimizer for any task in the environment. While [54] discuss this assumption in terms of the training tasks being sufficiently diverse, it can also be seen as the task environment being sufficiently similar, since no task is allowed to deviate significantly from the other tasks in the environment.

Information-theoretic bounds have also been applied to meta learning. Typically, the derivations of these bounds are based on a two-step approach, where one defines an auxiliary loss. This auxiliary loss can either be the population loss on the training tasks or the training loss on unobserved tasks. Using this auxiliary loss, a generalization bound is applied twice: first to bound the difference between the meta-training loss and the auxiliary loss, then to bound the difference between the auxiliary loss and the meta-population loss. Then, one of these steps is purely at the environment level, and one is purely at the task level, where the order depends on the choice of auxiliary loss. Composing these bounds gives a bound on the meta-population loss in terms of the meta-training loss, as desired. In [55], this is done on the basis of the bounds in Theorem 4.1 and Theorem 4.2, and the resulting bounds are applied for several meta-learning settings. PAC-Bayesian bounds have also extensively been used, for instance in [56,57]. In [58], the bound in Theorem 4.1 is combined with a one-step approach, where the auxiliary loss is not needed. This yields a tighter bound, where the task level and environment level are considered jointly. Finally, in [59], the bound in Theorem 4.9 is combined with a two-step approach to derive generalization bounds for meta learning. In order to achieve this, the CMI framework is extended to meta learning by considering a meta-supersample, which contains supersample for $2\hat{n}$ tasks. As for the standard CMI framework, half of these are selected for inclusion in the meta-training set, on the basis of a Bernoulli vector. Then, within each task, the standard CMI framework is used. This is illustrated in Figure 5.2. Through the chain rule for mutual information and the data-processing inequality, these bounds can be shown to be tighter than bounds based on the mutual information. However, due to the two-step derivation, the bounds in [59] have a suboptimal dependence on the number of samples.

CHAPTER 6

Summary

In this chapter, the contributions provided in Part II are summarized. We conclude Part I of this thesis by discussing limitations of our results and explore possible directions for future investigations emanating from the work contained in this thesis.

6.1 Contributions

In this thesis, we derive and analyze information-theoretic generalization bounds. In particular, we:

- extend existing bounds to new settings,
- derive novel bounds that are tighter than existing results,
- demonstrate that some of our bounds are numerically accurate for neural networks, and
- show that our bounds are expressive enough to recover some classical results.

Below, we provide a more detailed summary of each of the appended contributions.

Contribution A: "Generalization Bounds via Information Density and Conditional Information Density"

In this paper, we develop a framework for deriving generalization bounds of various types through the use of exponential inequalities. Not only can this approach be used to derive novel generalization bounds, but it also provides a unified way to recover several of the known results in the literature, both average bounds and tail bounds (PAC-Bayesian and single-draw). Notably, we obtain a new data-dependent single-draw bound in terms of the information density $\imath(W, \boldsymbol{Z})$ between the training data $\boldsymbol{Z}$ and the hypothesis W, which can be weakened to obtain many data-independent bounds. Our results illustrate a trade-off between the magnitude of the high moments of the information measures appearing in the bounds and the confidence levels that can be achieved. We then extend our exponential-inequality approach to the CMI framework introduced by Steinke and Zakynthinou [8], and as a result, we extend their bounds on the average generalization error to the PAC-Bayesian and single-draw settings. This exemplifies how our framework can be used to implement new ideas in bounds of all flavors at once. For this setting, we derive a new data-dependent single-draw bound in terms of the conditional information density $\imath(W, \boldsymbol{S}|\tilde{\boldsymbol{Z}})$ between the hypothesis W and the random vector $\boldsymbol{S}$ determining the training set selection, given the supersample $\tilde{\boldsymbol{Z}}$. When suitably weakened, this leads to a new result in terms of the conditional maximal leakage $\mathcal{L}(\boldsymbol{S} \to W|\boldsymbol{Z})$, which can be tighter than the corresponding bound based on the maximal leakage in [9, Cor. 9].

In addition to this, we present an approach to derive generalization bounds based on a change of measure argument that is used in the binary hypothesis testing literature. This yields a data-independent single-draw bound in terms of the tail of the information density $\imath(W, \boldsymbol{Z})$. This bound can be shown to imply essentially equivalent versions of the data-independent single-draw bounds that we derived through the exponential-inequality approach. We also extend this approach to the CMI framework, deriving a data-independent single-draw bound in terms of the conditional information density $\imath(W, \boldsymbol{S}|\tilde{\boldsymbol{Z}})$. Finally, we extend the Hölder-based approach used by Esposito et al. [9] to the CMI setting, and derive a bound in terms of the conditional α-mutual information, from which results in terms of the conditional Rényi divergence and the conditional maximal leakage follow. We note that the dependence on the training set size n in all bounds presented in this paper is of the form $\sqrt{\mathrm{IM}/n}$, where IM denotes some information measure. Due to the presence of the square root, these results are slow-rate bounds.

Contribution B: "Nonvacuous Loss Bounds with Fast Rates for Neural Networks via Conditional Information Measures"

Building on the work of Steinke and Zakynthinou [8], we obtain fast-rate bounds on the test loss of a randomized learning algorithm in the CMI framework, i.e., bounds with an IM/n-dependence on n where IM is a conditional information measure. Again, we obtain these results through the use of an exponential inequality. The cost of this rate improvement as compared to the bounds in Contribution A is that the multiplicative constants that appear in the bounds are larger, and in particular, the training loss is multiplied by a constant greater than one. This deterioration in multiplicative factors means that, in order for the new fast-rate bounds to be better than the previously obtained slow-rate ones, the training loss and information measure have to be sufficiently small. The same manipulations that were performed in Contribution A to obtain bounds in terms of information-theoretic quantities, such as conditional mutual information and conditional maximal leakage, can also be performed for these fast-rate bounds.

A particular focus of this contribution is how to apply the bounds from the CMI setting in the context of neural networks. We show that the CMI setting naturally enables data-dependent priors, which is an important technique for obtaining numerically tight PAC-Bayesian bounds. Following the approach taken in [39, 51], we model the learning algorithm $P_{W|\tilde{\boldsymbol{Z}}\boldsymbol{S}}$ as a Gaussian distribution centered around the output weights of stochastic gradient descent, and use a data-dependent prior that aims to approximate the true marginal $P_{W|\tilde{\boldsymbol{Z}}}$. With this, both the PAC-Bayesian and single-draw bounds, with either slow or fast rates, can be computed. We see that the resulting bounds essentially coincide with the tightest bounds that were previously obtained for the setups that we consider [51], but unlike previous results, our bounds also apply to the single-draw setting.

Contribution C: "A New Family of Generalization Bounds Using Samplewise Evaluated CMI"

In deriving the bounds in the previous contributions, the training loss and test loss were compared only through their (weighted) absolute difference. In Contribution C, we extend this to allow for arbitrary convex comparator functions, which can lead to significantly tighter bounds. While similar results have previously been derived in the PAC-Bayesian literature, we extend this to the CMI framework, obtaining average generalization bounds in terms of the disintegrated, samplewise, evaluated CMI, that is, the CMI evaluated in terms of the *loss* of the chosen hypothesis, rather than its parameters. In particular, through a novel concentration result for non-identically distributed random variables, we derive a bound where the convex comparator is the binary KL divergence. Additionally, we use our framework to recover and generalize several results from the literature. Again, through the lens of exponential inequalities, we extend these results to obtain PAC-Bayesian and single-draw bounds in terms of pointwise versions of the evaluated CMI.

In order to study the expressiveness of our framework, we consider multiclass classification with a hypothesis class of finite Natarajan dimension. For this setting, we show that the (pointwise) evaluated CMI that appears in our results can be bounded as a function of the Natarajan dimension. Combining this with our generalization bounds, we recover essentially optimal bounds from the literature. Furthermore, we numerically evaluate our bounds for several neural network settings. We find that our bounds are numerically accurate and improve on previous results, including for randomized labels, and remain stable throughout training. We perform experiments where we vary several hyperparameters, and find that our bounds are robust to these changes and induce the same ordering on the hyperparameter values as the true test error.

Contribution D: “Evaluated CMI Bounds for Meta Learning: Tightness and Expressiveness”

In Contribution D, we extend our techniques to meta learning. In meta learning, the meta-training set consists of several training sets from a number of training tasks, drawn from a common task distribution. Within each task, a base learning algorithm is used to select a hypothesis based on the in-task training set, as well as a hyperparameter. The objective of meta learning is to use a meta-learning algorithm to select the hyperparameter on the basis of the meta-training set. The goal is to have a small meta-test loss, i.e., a small average loss on test data for a new task from the task distribution. The meta-learner has access to the meta-training loss, which is the average loss on the training data for all training tasks. Most previous analyses of meta-learning use a two-step derivation, where the meta-training loss is first compared to an auxiliary loss, which is then compared to the meta-test loss. In contrast, we use a one-step derivation, where the meta-training loss and meta-test loss are compared directly. We show that the resulting bounds are tighter than comparable results from the literature. Furthermore, we extend the aforementioned techniques from the standard setting to meta learning, allowing us to obtain bounds in terms of the disintegrated, samplewise, evaluated CMI.

In order to examine the expressiveness of our bounds, we specialize our bounds to a representation learning setting that is well-studied in the literature. We find that our bounds allow us to recover the rates of classical bounds for this setting. As it turns out, the one-step derivation is a crucial ingredient to obtain these rates. By extending our analysis to oracle algorithms and empirical risk minimizers, we also essentially recover the rates of the excess risk bounds found in [54].

6.2 Future Work

As mentioned in the previous chapter, one remaining goal in the study of information-theoretic generalization guarantees is the ability to guide the design of modern machine learning algorithms. In their current form, the bounds discussed in this thesis do not fully exploit the structure of, for instance, neural networks, instead just treating the parameters as a generic vector that could potentially describe anything. It should, however, be noted that such structure can potentially be utilized by suitably selecting the prior distribution. While there is a strength in the aforementioned generality, further specializing the bounds to more concrete setups is needed to gain new insights. One potential improvement is to incorporate the symmetries that are present in most neural network architectures. One such symmetry is the homogeneity of the ReLU activation function, whereby for $a > 0$, we have $\mathrm{ReLU}(a \cdot x) = a \cdot \mathrm{ReLU}(x)$. Another example is permutation symmetry, where different units within layers can be swapped without affecting the functional form of the neural network. Properly utilizing these symmetries may improve the quantitative results that can be obtained, and potentially provide new insights. However, as discussed in [39], the non-isotropic random initialization that is typically used when training neural networks breaks many of the symmetries that are present, and it is unclear to what extent gains can be made by exploiting the remainder.

There are also several promising ways to improve the bounds themselves by using different tools than the ones presented in this thesis. For instance, instead of using the Donsker-Varadhan variational representation, [60] use tools from convex analysis to perform an alternative change of measure. In the resulting bounds, the mutual information is replaced by an arbitrary strongly convex function of the joint distribution. Combining this approach with the ideas discussed in this thesis may lead to tighter bounds. Furthermore, a variant of the CMI framework was recently presented by [61, 62]. In this variant, instead of choosing the n samples on the basis of a supersample with twice the size, the n samples are instead selected on the basis of a supersample of size $n + 1$. This may similarly be combined with the ideas in this thesis, in particular in relation to meta learning, to obtain improved results.

Finally, although this thesis addresses the expressiveness of information-theoretic generalization bounds to some extent, there are still questions remaining. While algorithmic stability, the VC dimension, and related complexity measures are studied in [47, 48], the relationship between information-theoretic bounds and Rademacher complexity, for instance, has not been established. Expanding the study of the information complexity of relevant algorithms would increase the applicability of information-theoretic generalization bounds, potentially giving rise to principled guidelines for practical algorithmic improvements.

Bibliography

[1] M. T. Banich, P. Dukes, and D. Caccamise, *Generalization of knowledge: Multidisciplinary perspectives.* New York, N.Y., USA: Psychology Press, 2010.

[2] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis, "A general reinforcement learning algorithm that masters chess, shogi, and go through self-play," *Science*, vol. 362, no. 6419, pp. 1140–1144, Dec. 2018.

[3] J. De Fauw, J. Ledsam, and B. e. a. Romera-Paredes, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nat. Med.*, vol. 24, pp. 1342–1350, Aug. 2018.

[4] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, Aug. 2021.

[5] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms.* Cambridge, U.K.: Cambridge Univ. Press, 2014.

[6] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Toulon, France, Apr. 2017.

[7] D. Narayanan, M. Shoeybi, J. Casper, P. LeGresley, M. Patwary, V. Korthikanti, D. Vainbrand, P. Kashinkunti, J. Bernauer, B. Catanzaro, A. Phanishayee, and M. Zaharia, "Efficient large-scale language model training on GPU clusters using Megatron-LM," in *Proc. Int. Conf. High Perf. Computing, Networking, Storage and Analysis (SC)*, St. Louis, MO, USA, Nov. 2021.

[8] T. Steinke and L. Zakynthinou, “Reasoning about generalization via conditional mutual information,” in *Conf. Learning Theory (COLT)*, Graz, Austria, July 2020.

[9] A. R. Esposito, M. Gastpar, and I. Issa, “Generalization error bounds via Rényi-, f-divergences and maximal leakage,” *IEEE Transactions on Information Theory*, vol. 67, no. 8, pp. 4986–5004, Aug. 2021.

[10] Kaggle, “Cats vs dogs,” Retrieved Nov. 2020. [Online]. Available: https://www.kaggle.com/c/dogs-vs-cats

[11] M. Dundar, B. Krishnapuram, J. Bi, and R. Rao, “Learning classifiers when the training data is not iid.” in *Inter. Joint Conf. on Artif. Intell. (IJCAI)*, Hyderabad, India, Jan. 2007.

[12] B. Settles, “Active learning literature survey,” 2009.

[13] L. Schmarje, M. Santarossa, S.-M. Schröder, and R. Koch, “A survey on semi-, self- and unsupervised learning for image classification,” *IEEE Access*, vol. 9, pp. 82 146–82 168, May 2021.

[14] D. McAllester, “Some PAC-Bayesian theorems,” in *Proc. Conf. Learn. Theory (COLT)*, Madison, WI, USA, July 1998, pp. 230–234.

[15] B. Guedj, “A primer on PAC-Bayesian learning,” *arXiv*, Jan. 2019. [Online]. Available: http://arxiv.org/abs/1901.05353

[16] T. Van Erven, P. Grünwald, N. Mehta, M. Reid, and R. Williamson, “Fast rates in statistical and online learning,” *J. of Mach. Learn. Res.*, vol. 16, pp. 1793–1861, Sep. 2015.

[17] P. Grünwald and N. Mehta, “Fast rates for general unbounded loss functions: from ERM to generalized Bayes,” *J. of Mach. Learn. Res.*, vol. 83, pp. 1–80, Mar. 2020.

[18] S. Arora, S. Du, W. Hu, Z. Li, and R. Wang, “Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, Long Beach, CA, USA, Jun 2019.

[19] Y. Polyanskiy and Y. Wu, *Lecture Notes On Information Theory*, 2019. [Online]. Available: http://www.stat.yale.edu/%7Eyw562/teaching/itlectures.pdf

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

[21] T. Van Erven and P. Harrëmos, “Rényi divergence and Kullback-Leibler divergence,” *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3797–3820, July 2014.

[22] I. Issa, S. Kamath, and A. B. Wagner, “An operational approach to information leakage,” *IEEE Trans. Inf. Theory*, vol. 66, no. 3, pp. 1625–1657, Mar. 2020.

[23] S. Verdú, "α-mutual information," in *Proc. Inf. Theory Appl. Workshop (ITA)*, San Diego, CA, USA, Feb. 2015.

[24] M. Tomamichel and M. Hayashi, "Operational interpretation of Rényi information measures via composite hypothesis testing against product and Markov distributions," *IEEE Trans. Inf. Theory*, vol. 64, no. 2, pp. 1064–1082, Feb. 2018.

[25] A. R. Esposito, D. Wu, and M. C. Gastpar, "On conditional Sibson's α-mutual information," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Melbourne, Australia, July 2021.

[26] W. Rudin, *Real and Complex Analysis, 3rd Ed.* USA: McGraw-Hill, Inc., 1987.

[27] M. D. Donsker and S. R. S. Varadhan, "Asymptotic evaluation of certain Markov process expectations for large time, i," *Comm. Pure Appl. Math*, vol. 28, no. 1, pp. 1–47, Jan. 1975.

[28] D. A. McAllester, "PAC-Bayesian stochastic model selection," *Mach. Learn.*, vol. 51, pp. 5–21, Apr. 2003.

[29] A. Banerjee, "On Bayesian bounds," in *Proc. Int. Conf. Mach. Learning (ICML)*, June 2006.

[30] M. Raginsky and I. Sason, "Concentration of measure inequalities in information theory, communications, and coding," *Foundations and Trends in Communications and Information Theory*, vol. 10, no. 1-2, pp. 1–246, 2013.

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

[32] M. J. Wainwright, *High-Dimensional Statistics: a Non-Asymptotic Viewpoint.* Cambridge, U.K.: Cambridge Univ. Press, 2019.

[33] A. Maurer, "A note on the PAC Bayesian theorem," Nov. 2004. [Online]. Available: https://arxiv.org/abs/cs/0411099

[34] D. McAllester, "A PAC-Bayesian tutorial with a dropout bound," *arXiv*, July 2013. [Online]. Available: http://arxiv.org/abs/1307.2118

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

[36] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, USA, Dec. 2017.

[37] Y. Bu, S. Zou, and V. V. Veeravalli, "Tightening mutual information based bounds on generalization error," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Paris, France, July 2019.

[38] J. Shawe-Taylor and R. C. Williamson, "A PAC analysis of a Bayesian estimator," in *Proc. Conf. Learn. Theory (COLT)*, Nashville, TN, USA, July 1997.

[39] G. Dziugaite and D. Roy, "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, Aug. 2017.

[40] W. Zhou, V. Veitch, M. Austern, R. Adams, and P. Orbanz, "Non-vacuous generalization bounds at the ImageNet scale: a PAC-Bayesian compression approach," in *Proc. Int. Conf. Learn. Representations (ICLR)*, New Orleans, LA, USA, May 2019.

[41] R. Bassily, S. Moran, I. Nachum, J. Shafer, and A. Yehudayoff, "Learners that use little information," *J. of Mach. Learn. Res.*, vol. 83, pp. 25–55, Apr. 2018.

[42] J. Langford and M. Seeger, "Bounds for averaging classifiers," *CMU Technical report*, vol. CMU-CS-01-102, 2001.

[43] P. Alquier, "User-friendly introduction to PAC-Bayes bounds," *arXiv*, Nov. 2021. [Online]. Available: https://arxiv.org/abs/2110.11216

[44] A. Y. K. Foong, W. P. Bruinsma, D. R. Burt, and R. E. Turner, "How tight can PAC-Bayes be in the small data regime?" June 2021. [Online]. Available: https://arxiv.org/abs/2106.03542

[45] M. Haghifam, J. Negrea, A. Khisti, D. Roy, and G. Dziugaite, "Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms," *arXiv*, Apr. 2020. [Online]. Available: http://arxiv.org/abs/2004.12983

[46] J. Negrea, M. Haghifam, G. Dziugaite, A. Khisti, and D. Roy, "Information-theoretic generalization bounds for SGLD via data-dependent estimates," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Canada, Dec. 2019.

[47] H. Harutyunyan, M. Raginsky, G. V. Steeg, and A. Galstyan, "Information-theoretic generalization bounds for black-box learning algorithms," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Virtual Conference, Dec. 2021.

[48] M. Haghifam, G. K. Dziugaite, S. Moran, and D. M. Roy, "Towards a unified information-theoretic framework for generalization," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Virtual Conference, Dec. 2021.

[49] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2012.

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2017.

[51] G. Dziugaite, K. Hsu, W. Gharbieh, and D. Roy, “On the role of data in PAC-Bayes bounds,” in *Proc. Artif. Intell. Statist. (AISTATS)*, Virtual conference, Apr. 2021.

[52] J. Baxter, “A model of inductive bias learning,” *J. Artif. Int. Res.*, vol. 12, no. 1, pp. 149–198, Mar. 2000.

[53] A. Maurer, M. Pontil, and B. Romera-Paredes, “The benefit of multitask representation learning,” *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2853–2884, Jan. 2016.

[54] N. Tripuraneni, M. Jordan, and C. Jin, “On the theory of transfer learning: The importance of task diversity,” in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Canada, Dec. 2020.

[55] S. T. Jose and O. Simeone, “Information-theoretic generalization bounds for meta-learning and applications,” *Entropy*, vol. 23, no. 1, Jan. 2021.

[56] A. Pentina and C. Lampert, “A PAC-Bayesian bound for lifelong learning,” in *Proc. Int. Conf. Mach. Learning (ICML)*, Bejing, China, June 2014.

[57] R. Amit and R. Meir, “Meta-learning by adjusting priors based on extended PAC-Bayes theory,” in *Proc. Int. Conf. Mach. Learning (ICML)*, Stockholm, Sweden, July 2018.

[58] Q. Chen, C. Shui, and M. Marchand, “Generalization bounds for meta-learning: An information-theoretic analysis,” in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Virtual Conference, Dec. 2021.

[59] A. Rezazadeh, S. T. Jose, G. Durisi, and O. Simeone, “Conditional mutual information-based generalization bound for meta learning,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Melbourne, Australia, July 2021.

[60] G. Lugosi and G. Neu, “Generalization bounds via convex analysis,” in *Conf. Learning Theory (COLT)*, London, UK, July 2022.

[61] M. R. Rammal, A. Achille, S. Diggavi, S. Soatto, and A. Golatkar, “On leave-one-out conditional mutual information for generalization,” in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, New Orleans, LA, USA, Dec. 2022.

[62] M. Haghifam, S. Moran, D. M. Roy, and G. Karolina Dziugiate, “Understanding generalization via leave-one-out conditional mutual information,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Espoo, Finland, June 2022.