THESIS FOR THE DEGREE OF LICENTIATE OF ENGINEERING

Guaranteeing Generalization via Measures of Information

Fredrik Hellström

CHALMERS
UNIVERSITY OF TECHNOLOGY

Communication Systems Group
Department of Electrical Engineering
Chalmers University of Technology
Göteborg, Sweden, 2020

Guaranteeing Generalization via Measures of Information

Fredrik Hellström

This thesis has been prepared using LaTeX and Tikz.

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

Printed by Chalmers Reproservice
Göteborg, Sweden, November 2020

Abstract

During the past decade, machine learning techniques have achieved impressive results in a number of domains. Many of the success stories have made use of deep neural networks, a class of functions that boasts high complexity. Classical results that mathematically guarantee that a learning algorithm generalizes, i.e., performs as well on unseen data as on training data, typically rely on bounding the complexity and expressiveness of the functions that are used. As a consequence of this, they yield overly pessimistic results when applied to modern machine learning algorithms, and fail to explain why they generalize.

This discrepancy between theoretical explanations and practical success has spurred a flurry of research activity into new generalization guarantees. For such guarantees to be applicable for relevant cases such as deep neural networks, they must rely on some other aspect of learning than the complexity of the function class. One avenue that is showing promise is to use methods from information theory. Since information-theoretic quantities are concerned with properties of different data distributions and relations between them, such an approach enables generalization guarantees that rely on the properties of learning algorithms and data distributions.

In this thesis, we first introduce a framework to derive information-theoretic guarantees for generalization. Specifically, we derive an exponential inequality that can be used to obtain generalization guarantees not only in the average sense, but also tail bounds for the PAC-Bayesian and single-draw scenarios. This approach leads to novel generalization guarantees and provides a unified method for deriving several known generalization bounds that were originally discovered through the use of a number of different proof techniques. Furthermore, we extend this exponential-inequality approach to the recently introduced random-subset setting, in which the training data is randomly selected from a larger set of available data samples.

One limitation of the proposed framework is that it can only be used to derive generalization guarantees with a so-called slow rate with respect to the size of the training set. In light of this, we derive another exponential inequality for the random-subset setting which allows for the derivation of generalization guarantees with fast rates with respect to the size of the training set. We show how to evaluate the generalization guarantees obtained through this inequality, as well as their slow-rate counterparts, for overparameterized neural networks trained on MNIST and Fashion-MNIST. Numerical results illustrate that, for some settings, these bounds predict the true generalization capability fairly well, essentially matching the best available bounds in the literature.

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

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," to be published in *IEEE Journal on Selected Areas of Information Theory*.

[B] **F. Hellström**, G. Durisi, "Nonvacuous Loss Bounds with Fast Rates for Neural Networks via Conditional Information Measures," submitted to *International Conference on Learning Representations*, October 2020.

Publications by the author not included in the thesis:

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

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

Acknowledgements

I would like to thank my supervisor Prof. Giuseppe Durisi for his guidance and support throughout the development of this thesis. Your judicious eye and helpful advice have taught me a great deal about research. I would also like to thank Prof. Fredrik Kahl and Prof. Cristopher Zach for discussions that have provided me with a wider perspective.

I appreciate all of my colleagues in the CommSys group and the rest of the E2 department, as well as my fellow students in the WASP Graduate School, for making work a better place. In this period of working from home, I am also heavily indebted to my lamp and my couch for providing me with comfort.

My profound gratitude goes to my family, for their continual encouragement throughout my studies. Finally, I wish to express an appreciation for Berna that is beyond measure. Your support (and general existence) has upper-bounded my stress levels and provided lower bounds for my happiness.

Göteborg, 2020

Financial Support

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

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 translation, medical diagnosis, and chess [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 an *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. Most of the success stories of recent years make use of *deep neural networks* (DNNs), which are 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.

In this thesis, we take some steps toward explaining generalization for randomized learning algorithms, and in particular, we present new results for DNNs. In Paper 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 an exponential inequality, from which generalization bounds follow from simple manipulations. We combine this framework with the random-subset setting introduced

by Steinke and Zakynthinou [7], where we can derive even tighter bounds. In Paper B, we strengthen the previously obtained random-subset bounds even further, improving their dependence on the size of the training data set. We demonstrate how to evaluate the bounds both from Paper A and Paper 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.

1.0.1 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 consists of appended papers, which form the research contribution upon which this thesis is based.

Part I is organized as follows. In Chapter 1, we first give an informal overview of the field, to set the more specific problems in a wider context, before introducing necessary notation. Then, in Chapter 3, we present the statistical learning setting and review some of the classical generalization guarantees.

Next, in Chapter 3, we turn to the more recent information-theoretic generalization guarantees. After establishing the necessary toolbox that will be used throughout the chapter, we give an overview of PAC-Bayesian generalization guarantees, back to which many ideas of the information-theoretic bounds can be traced. Next, we give an overview of the information-theoretic bounds available in the literature, including the recently introduced random-subset setting.

In Chapter 4, we conclude the first part of the thesis by detailing the contributions made in the appended papers and discussing possible future directions to investigate.

1.0.2 Notation

Throughout the first part of this thesis, we use capital letters Z to denote random variables, and lower-case letters z to denote their realizations. Similarly, random vectors are denoted by bold capital letters $\boldsymbol{Z}$ and their realizations by lower-case bold letters $\boldsymbol{z}$. If P_Z is a probability measure, we denote the probability operator under it as $P_Z[\,\cdot\,]$, and we denote the expectation operator as $\mathbb{E}_{P_Z}[\,\cdot\,]$. 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 setup that we consider throughout the thesis. We then discuss the various flavors of generalization guarantees that will be discussed, before presenting the classical generalization guarantees that are based on the VC dimension and the Rademacher complexity. In Chapter 3, these classical results will be contrasted with more recently obtained information-theoretic guarantees.

2.1 The Learning Setup

We start by discussing the general ingredients that are common to all learning setups considered in this thesis, before giving some more 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 $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_n)$, consisting of n samples drawn i.i.d. 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 $\boldsymbol{z} = (z_1, \ldots, z_n)$ 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, $\mathrm{gen}(w, \boldsymbol{z}) = L_{P_Z}(w) - L_{\boldsymbol{z}}(w)$.

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

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

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) [8].

While the learning setting described in this section is quite general, it does not cover all possible settings of interest. For instance, the assumption that the training data $Z_1, Z_2, \ldots, Z_n$ are i.i.d. can be lifted [9]. In the setting of transfer learning, the training data are drawn from one distribution, while the population loss is computed with respect to a different one [10]. In meta-learning, one has access to several data sets from different, related tasks, drawn from a distribution over tasks. The goal is to learn hyperparameters, i.e., parameters that describe the within-task learning algortihms [11]. We will, however, restrict our attention to problems that can be seen as special cases of the setting described in this section.

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 i.i.d. 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 achieves zero population loss, so that $\inf_{w \in W} L_{P_Z}(w) = 0$. It is important to note that the PAC formulation of generalization is focused on properties of the hypothesis class $\mathcal{W}$ itself.

Average guarantee: 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. Having a bound on the average loss does not necessarily to imply any good guarantees on the tail of the loss distribution with respect to the data.

PAC-Bayesian guarantee: The PAC-Bayesian setting was introduced by McAllester [12] 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 [13].

Single-draw guarantee: 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 is typically applied 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] Thus, 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

[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.

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 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

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

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

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

The VC dimension d of $\mathcal{W}$ equals 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].

Theorem 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 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. *(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 [14, 15].

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 PAC learnability is not 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. *Rademacher complexity:*
Let $Z^n \in \mathcal{Z}^n$ be a set of data samples and let $\ell(\,\cdot\,,\,\cdot\,) : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}^+$ be a loss function. Let σ_i for $i = 1, \ldots, n$ 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 X^n *and* $\ell(\,\cdot\,,\,\cdot\,)$ *is given by*

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

One way to understand the Rademacher complexity is to think of randomly splitting the data set Z^n 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 3. *Generalization guarantee from Rademacher complexity:*
Assume that, for all $z \in \mathcal{Z}$ *and all* $w \in \mathcal{W}$, $|\ell(w,z)| \leq c$. *With probability at least* $1-\delta$ *over* $P_{\boldsymbol{Z}}$, *for all* $w \in \mathcal{W}$,

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

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 (SVMs), 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 [16], 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-Theoretic Generalization Guarantees

In this chapter, we overview the information-theoretic generalization guarantees that are available in the literature. It is in the context of these results that the contributions made in the appended papers is best understood. We start by motivating the need for new generalization guarantees, beyond the classical results discussed in Chapter 2, and discuss why information-theoretic methods constitute a good candidate approach. We then present the toolbox that is used in a large part of the literature, before describing the main results that have been obtained. We end by presenting the random-subset setting, where the training data is randomly selected from a larger set of data samples. This settings plays an important role in the appended papers.

3.1 Motivation

The celebrated fundamental theorem of statistical learning [5] shows that the VC dimension completely characterizes PAC learnability. However, the 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.

3.2 The Toolbox

We now introduce some tools that are used throughout the literature on information-theoretic generalization guarantees. We begin by defining some quantities of interest. First, the Radon-Nikodym derivative of a probability measure P with respect to a probability measure Q is denoted by $\mathrm{d}P/\,\mathrm{d}Q$. It is well defined if P is absolutely continuous with respect to Q, denoted $P \ll Q$, meaning that if P assigns non-zero probability to a set, then Q does as well. When these distributions are the joint distribution $P_{W\boldsymbol{Z}}$ and the product of marginals $P_W P_{\boldsymbol{Z}}$ respectively, the logarithm of the Radon-Nikodym derivative is the information density

$$\imath(W, \boldsymbol{Z}) = \log \frac{\mathrm{d}P_{W\boldsymbol{Z}}}{\mathrm{d}P_W P_{\boldsymbol{Z}}}. \tag{3.1}$$

The KL divergence, also known as the relative entropy, between P and Q is given by $D(P\,||\,Q) = \mathbb{E}_P[\log \mathrm{d}P/\,\mathrm{d}Q]$, and when $P = P_{W\boldsymbol{Z}}$ and $Q = P_W P_{\boldsymbol{Z}}$, this becomes the mutual information $I(W;\boldsymbol{Z})$.

The generalization error is the difference in loss value that arises when the data samples and hypothesis are jointly distributed or independently distributed according to their marginals. Thus, we want to compare the value of a function under one distribution to its value under another distribution. A widely used tool for doing precisely this is the Donsker-Varadhan variational representation of the KL divergence.

Lemma 1. *(Donsker-Varadhan variational representation of the KL divergence)*
Let P_X and Q_X be probability measures on a space $\mathcal{X}$ such that P_X is absolutely continuous with respect to Q_X, and let $f : \mathcal{X} \to \mathbb{R}$ be a measurable function. Then,

$$D(P_X\,||\,Q_X) \geq \mathbb{E}_{P_X}[f(X)] - \log \mathbb{E}_{Q_X}\left[e^{f(X)}\right]. \tag{3.2}$$

To derive generalization guarantees, we will also need some restrictions on the loss functions that we consider. Commonly, it is assumed that $\ell(w, Z)$ is a sub-Gaussian random variable under P_Z for all w. We now define sub-Gaussianity.

Definition 3. *(Sub-Gaussian random variables)*
A random variable X is σ-sub-Gaussian if, for all $\lambda \in \mathbb{R}$,

$$\log \mathbb{E}[\exp(\lambda X - \mathbb{E}[X])] \leq \frac{\lambda^2 \sigma^2}{2}. \tag{3.3}$$

The left-hand side of (3.3) is referred to as the cumulant generative function (CGF) of λ. Importantly, any random variable that is almost surely bounded to $[a, b]$ is $(b-a)/2$-sub-Gaussian [17, Chap. 2]. Furthermore, if X_i for $i = 1, \ldots, n$ are independent σ-sub-Gaussian random variables, the average $\frac{1}{n}\sum_{i=1}^{n} X_i$ is $\sigma/\sqrt{n}$-sub-Gaussian. The CGF of some random variables that are not sub-Gaussian can still be similarly bounded by using the Legendre dual [18]. For simplicity, we will restrict our attention to sub-Gaussian random variables.

An important result for sub-Gaussian random variables is Hoeffding's inequality.

Lemma 2. *(Hoeffding's inequality)*
Let X be a σ-sub-Gaussian random variable. Then, for all $\epsilon > 0$,

$$\Pr[|X - \mathbb{E}[X]| \geq \epsilon] \leq 2 \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right). \tag{3.4}$$

3.3 PAC-Bayesian Bounds

The genesis of information-theoretic approaches to generalization guarantees can be found within the PAC-Bayesian literature. The PAC-Bayesian approach found its start when McAllester [12] worked on developing PAC-style bounds to classifiers of a Bayesian flavor. These bounds rely on 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 [19–22].

Below, we give a somewhat more modern version of the basic PAC-Bayesian bound [13].

Theorem 4. *(Canonical PAC-Bayesian bound)*
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 $P_{\boldsymbol{Z}}$,

$$\mathbb{E}_{P_{W|\boldsymbol{Z}}}[L_{P_Z}(W)] \leq \mathbb{E}_{P_{W|\boldsymbol{Z}}}[L_{\boldsymbol{Z}}(W)] + \sqrt{\frac{2\sigma^2}{n}\left(D(P_{W|\boldsymbol{Z}} \,||\, Q_W) + \log\frac{1}{\delta}\right)}. \tag{3.5}$$

Proof. We begin by applying the Donsker-Varadhan variational representation (3.2) with $P_X = P_{W|\boldsymbol{Z}}$, $Q_X = Q_W$ and $f(X) = \lambda(L_{P_Z}(W) - L_{\boldsymbol{Z}(\boldsymbol{S})}(W))$, to see that for all λ,

$$\mathbb{E}_{P_{W|\boldsymbol{Z}}}[L_{P_Z}(W) - L_{\boldsymbol{Z}}(W)] \leq \frac{D(P_{W|\boldsymbol{Z}} \,||\, Q_W) + \log \mathbb{E}_{Q_W}\left[e^{\lambda(L_{P_Z}(W) - L_{\boldsymbol{Z}}(W))}\right]}{\lambda}. \tag{3.6}$$

We now apply Markov's inequality to obtain

$$\log \mathbb{E}_{Q_W}\left[e^{\lambda(L_{P_Z}(W) - L_{\boldsymbol{Z}}(W))}\right] \leq \log \mathbb{E}_{Q_W P_Z}\left[e^{\lambda(L_{P_Z}(W) - L_{\boldsymbol{Z}}(W))}\right] + \log\frac{1}{\delta}. \tag{3.7}$$

Since $\ell(w, Z)$ is σ-sub-Gaussian for all w, $L_{\boldsymbol{Z}}(w)$ is an average of n independent σ-sub-Gaussian variables, and is thus $\sigma/\sqrt{n}$-sub-Gaussian for all w. By applying the definition of sub-Gaussianity (3.3) with $X = L_{\boldsymbol{Z}(\boldsymbol{S})}(W)$, we see that the logarithm on the right-hand side of (3.7) is bounded by $\lambda^2\sigma^2/2$. Inserting this into (3.6) and setting $\lambda = \sqrt{2n(\log 1/\delta + D(P_{W|\boldsymbol{Z}} \,||\, Q_W))}/\sigma$ to minimize the bound, we obtain the desired result. □

We note that the dependence on n in (3.12) 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 (3.12) 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 [19], 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 [14, 15], the KL divergence $D(P_{W|\boldsymbol{Z}} \,||\, Q_W)$ must grow at most polylogarithmically in n.

Theorem 5. *(Fast-rate PAC-Bayesian bound)*
For all $\lambda \in (0, 1)$, the following holds with probability at least $1-\delta$ under $P_{\boldsymbol{Z}}$:

$$\mathbb{E}_{P_{W|\boldsymbol{Z}}}[L_{P_Z}(W)] \leq \frac{1}{\lambda}\left[\mathbb{E}_{P_{W|\boldsymbol{Z}}}[L_{\boldsymbol{Z}}(W)] + \frac{D(P_{W|\boldsymbol{Z}} \,||\, Q_W) + \log\frac{1}{\delta}}{2(1-\lambda)n}\right]. \tag{3.8}$$

[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 starkly different behaviors for different sample sizes.

3.4 Information-Theoretic Bounds

We now turn to more recent studies on generalization guarantees, where the information-theoretic connections are stated more clearly. Initial work on explicitly tying generalization guarantees to the mutual information (MI), a core quantity within information theory, was performed by Russo and Zou [23]. 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 [24] extended this to uncountable domains, and highlighted the connection to learning. We present the main result from Xu and Raginsky [24, Thm. 1] below.

Theorem 6. *(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 $P_{W|\boldsymbol{Z}} \ll P_W$. Then,

$$\mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{P_Z}(W)] \leq \mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{\boldsymbol{Z}}(W)] + \sqrt{\frac{2\sigma^2 I(W;\boldsymbol{Z})}{n}}. \tag{3.9}$$

The proof of this result essentially follows along the same lines as the proof of Theorem 4, but with $P_X = P_{W\boldsymbol{Z}}$ and $Q = P_W P_{\boldsymbol{Z}}$. Already having the expectation over $P_{\boldsymbol{Z}}$ makes the Markov step superfluous, but the proof is otherwise identical.

The big advantage of the 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, such a learner is not very interesting. A discussion of more relevant scenarios where the mutual information can be bounded, such as noisy empirical risk minimization, can be found in [24].

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 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 [18], who used the methods of Xu and Raginsky to derive a generalization guarantee in terms of the sample-wise mutual informatia, $I(W;Z_i)$ for $i = 1, \ldots, n$. Since W is typically undecided 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 7. *(Average bound in terms of sample-wise 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,

$$\mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{P_Z}(W)] \leq \mathbb{E}_{P_{W\boldsymbol{Z}}}\left[L_{\boldsymbol{Z}(\boldsymbol{S})}(W)\right] + \frac{1}{n}\sum_{i=1}^{n}\sqrt{2\sigma^2 I(W;Z_i)}. \tag{3.10}$$

This result relies on the decomposition

$$\mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{P_Z}(W)] - \mathbb{E}_{P_{W\boldsymbol{Z}}}\left[L_{\boldsymbol{Z}(\boldsymbol{S})}(W)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{P_{WZ_i}}[\ell(W,Z_i)] - \mathbb{E}_{P_{WZ_i}}[\ell(W,Z_i)]\,. \tag{3.11}$$

Applying the same arguments as were used to prove Theorem 6 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 sample-wise mutual information guarantee is always tighter than the original mutual information result [18, Prop. 1].

As observed by Bassily *et al.* [25], the PAC-Bayesian bound in (3.12) 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 $1/\delta$.

Theorem 8. *(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,

$$\mathbb{E}_{P_{W|\boldsymbol{Z}}}[L_{P_Z}(W)] \leq \mathbb{E}_{P_{W|\boldsymbol{Z}}}\left[L_{\boldsymbol{Z}(\boldsymbol{S})}(W)\right] + \sqrt{\frac{2\sigma^2}{n}\left(\frac{2I(W;\boldsymbol{Z})}{\delta} + \log\frac{2}{\delta}\right)}. \tag{3.12}$$

Here, the factor 2 multiplying the $1/\delta$ stems from the fact that we have to use a union bound to combine both uses of Markov's inequality. Of course, the same conversion can be performed in Theorem 5.

In [24, 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.* [25] also derive such a single-draw bound, but obtain better constants.

Theorem 9. *(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 $P_{W\boldsymbol{Z}}$,*

$$L_{P_Z}(W) \leq L_{\boldsymbol{Z}}(W) + \sqrt{\frac{2\sigma^2}{n}\left(\frac{I(W;\boldsymbol{Z}) + H_b(\delta)}{\delta}\right)}. \tag{3.13}$$

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. Then, the data processing inequality for the KL divergence implies that

$$D(P\,||\,Q) \geq d(p\,||\,q) \geq -H_b(p) + p\log\frac{1}{q}. \tag{3.14}$$

Here, $d(p\,||\,q)$ denotes the KL divergence between two Bernoulli distributions with parameters p and q respectively, while $H_b(p)$ denotes 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

$$\mathcal{E} = \{(w, \boldsymbol{z}) : L_{P_Z}(w) - L_{\boldsymbol{z}}(w) > \epsilon\}. \tag{3.15}$$

The σ-sub-Gaussianity of the loss function implies that [17, Eq. (2.9)]

$$P_{Z^n}[\mathcal{E} > \epsilon] \leq \exp\bigl(-n\epsilon^2/(2\sigma^2)\bigr)\,. \tag{3.16}$$

From this, it follows that

$$\log \frac{1}{q} \geq n\frac{\epsilon^2}{2\sigma^2} \tag{3.17}$$

which, substituted into (3.14), gives us

$$\epsilon \leq \sqrt{\frac{2\sigma^2}{n}\left(\frac{I(W;Z^n) + H_b(p)}{p}\right)}. \tag{3.18}$$

Since the right-hand side of (3.18) is monotonically decreasing in p, we conclude that the condition

$$\epsilon \geq \sqrt{\frac{2\sigma^2}{n}\left(\frac{I(W;Z^n) + H_b(\delta)}{\delta}\right)} \tag{3.19}$$

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.* [26] 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 10. *(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 $P_{W\boldsymbol{Z}}$,

$$L_{P_Z}(W) \leq L_{\boldsymbol{Z}}(W) + \sqrt{\frac{2\sigma^2}{n}\left[I_\alpha(W;\boldsymbol{Z}) + \frac{\alpha}{\alpha-1}\log\frac{1}{\delta}\right]}. \tag{3.20}$$

Here, $I_\alpha(\,\cdot\,,\,\cdot\,)$ is the α-mutual information

$$I_\alpha(W;Z^n) = \frac{\alpha}{\alpha-1}\log \mathbb{E}_{P_{Z^n}}\left[\mathbb{E}^{1/\alpha}_{P_W}\left[\left(\frac{\mathrm{d}P_{WZ^n}}{\mathrm{d}P_W P_{Z^n}}\right)^\alpha\right]\right]. \tag{3.21}$$

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 typically called *high-probability* bounds. 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 Paper 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. Furthermore, the information-theoretic bounds presented in this section depend on the data distribution $P_{\boldsymbol{Z}}$, which is unknown in most scenarios. This makes the bounds impossible to compute in any practical setting. In Paper A, we address this issue by noting that one can replace the marginal distribution P_W, which depends on the unknown data distribution, with a suitably chosen auxiliary distribution Q_W.

3.4.1 The Random-Subset Setting

Recently, Steinke and Zakynthinou [7] considered a setting with more structure, which we will refer to as the *random-subset setting*. In this setting, we have $2n$ training samples $\tilde{\boldsymbol{Z}} = (\tilde{Z}_1, \ldots, \tilde{Z}_{2n})$, referred to as a *supersample*. From this, the training set is randomly formed as follows: let $\boldsymbol{S} = (S_1, \ldots, S_n)$ 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 $\boldsymbol{Z}(\boldsymbol{S}) = (Z_1(S_1), \ldots, Z_n(S_n))$ 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 $\boldsymbol{Z}(\boldsymbol{S})$.

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 [24, Thm. 1], but given in terms of the *conditional* mutual information $I(W; \boldsymbol{S}|\tilde{\boldsymbol{Z}})$. We present this result below.

Theorem 11. *(Slow-rate average bound in terms of conditional MI)*
Assume that $\ell(w, z) \in [0, 1]$ for all $w \in \mathcal{W}$ and $z \in \mathcal{Z}$. Then,

$$\mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{P_Z}(W)] \leq \mathbb{E}_{P_{W\boldsymbol{Z}}}\left[L_{\boldsymbol{Z}(\boldsymbol{S})}(W)\right] + \sqrt{\frac{2I(W; \boldsymbol{S}|\tilde{\boldsymbol{Z}})}{n}}. \tag{3.22}$$

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 11 improves upon Theorem 6, for the special case of a bounded loss function, because the information of each sample is normalized to 1 bit—indeed, the conditional MI can be upper-bounded as $I(W; \boldsymbol{S}|\tilde{\boldsymbol{Z}}) \leq H(\boldsymbol{S}) = n \log 2$. 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 11 and Theorem 6 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 [7, 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 12. *(Fast-rate average bound in terms of conditional MI)*
Assume that $\ell(w,z) \in [0,1]$ for all $w \in \mathcal{W}$ and $z \in \mathcal{Z}$. Then,

$$\mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{P_Z}(W)] \leq 2\,\mathbb{E}_{P_{W\boldsymbol{Z}}}\left[L_{\boldsymbol{Z}(\boldsymbol{S})}(W)\right] + \frac{3I(W;\boldsymbol{S}|\tilde{\boldsymbol{Z}})}{n}. \tag{3.23}$$

To achieve a fast rate, the boundedness of the loss function is used more directly than in the derivation of the slow-rate bound in Theorem 11. An alternative derivation can be found in Corollary 1 of Paper B.

Similar to the sample-wise extension of the average MI bound performed by Bu and Veeravalli [18], Haghifam et al. [27, Thm. 3.4] extended the CMI result to a sample-wise CMI bound, using the same decomposition as in Theorem 7. They also use the disintegration ideas introduced in [28] to pull the expectation over $P_{\tilde{\boldsymbol{Z}}}$ outside of the square root, which tightens the resulting bound.

Theorem 13. *(Slow-rate average bound in terms of sample-wise conditional MI)*
Assume that $\ell(w,z) \in [0,1]$ for all $w \in \mathcal{W}$ and $z \in \mathcal{Z}$. Then,

$$\mathbb{E}_{P_{W\boldsymbol{Z}}}[L_{P_Z}(W)] \leq \mathbb{E}_{P_{W\boldsymbol{Z}}}\left[L_{\boldsymbol{Z}(\boldsymbol{S})}(W)\right] + \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{P_{\tilde{\boldsymbol{Z}}}}\left[\sqrt{2D(P_{WS_i|\tilde{\boldsymbol{Z}}} \,||\, P_{W|\tilde{\boldsymbol{Z}}} P_{S_i})}\right]. \tag{3.24}$$

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 11.

In Proposition 1 in Paper B, we slightly extend this by showing how to also move the expectation over W outside the square root. We also use a similar decomposition to obtain a fast-rate version of the sample-wise CMI bound.

3.5 Applications to Neural Networks

As mentioned in the beginning of this section, one motivation for studying new types of generalization guarantees, beyond the classical ones, is that the performance of modern machine learning algorithms 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 [21], 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 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. [22] derived a PAC-Bayesian bound that applies to deterministic, pruned networks. To obtain such a network, you first train a large neural network, and then remove parameters that do not affect performance too much, and end up with a similarly well-performing network, the size of which is but a fraction of the original network size. An impressive aspect of [22] 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. [28] applied their disintegrated, sample-wise 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.

More recently, in [29], 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 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. These bounds are the tightest available in the literature for both CNNs and fully connected networks trained on MNIST and Fashion-MNIST, both for the scenario in which the network is trained using normal stochastic gradient descent and the scenario where the bound is directly optimized.

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. For instance, the tightest available bounds are obtained by artificially adding Gaussian noise to the outputs of stochastic gradient descent. Better modelling the noise inherent to neural network training, or obtaining bounds for the means of these distributions, would be a step towards bounds for a more realistic setting. 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.

CHAPTER 4

Contributions and Outlook

In this chapter, the contributions of the appended papers are summarized. Then, we overview some possible directions for future investigations emanating from the work contained in this thesis.

4.1 Contributions

"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 an exponential inequality. 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 random-subset setting introduced by Steinke and Zakynthinou [7], 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 [26, 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-dependent single-draw bounds that we derived through the exponential-inequality approach. We also extend this approach to the random-subset setting, 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.* [26] to the random-subset 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 (conditional) information measure. Due to the presence of the square root, these results are slow-rate bounds.

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

Building on the work of Steinke and Zakynthinou [7], we obtain fast-rate random-subset bounds on the population and test loss of a randomized learning algorithm, 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 increase as compared to the bounds in Paper 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 Paper 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 paper is how to apply the random-subset bounds in the context of neural networks. Following the approach taken in [21, 29], 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 [29], but unlike previous results, our bounds also apply to the single-draw setting.

4.2 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 at all exploit the structure of, for instance, neural networks, instead just treating the parameters as a generic vector that could potentially describe anything. Of course, there is a strength in such generality, but specializing the bounds to more concrete setups is needed to gain new insights. One straightforward improvement that could be performed is to incorporate the symmetries that are present in most neural network architectures. Examples of such symmetries include 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 result that can be obtained, and potentially provide new insights. However, as discussed in [21], 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. Perhaps more interesting would be to study systematically how design choices, such as the network architecture, learning rates, and other hyperparameters, affect both generalization performance and the estimates obtained by information-theoretic bounds. If one finds good agreement between these, this could prove to be a path to connecting design choices with information-theoretic generalization guarantees.

A more concrete promising avenue for future work is to combine the exponential-inequality approach that is used in both Paper A and Paper B with the data-holdout technique used in [29]. This approach can lead to a new exponential inequality that can be used to derive different types of generalization bounds in the same way as was done in both Paper A and Paper B. An advantage of such an approach would be that, within this framework, the posterior can be chosen based on the entirety of $\tilde{\boldsymbol{Z}}$, while the prior can be chosen only on the basis of $\boldsymbol{Z}(\boldsymbol{R})$, where $\boldsymbol{R}$ can be any random variable that selects a subset of $\tilde{\boldsymbol{Z}}$. This opens up many possibilities for adjusting the procedure by which the bound is evaluated. For instance, [29] exploited the random ordering of mini-batches in stochastic gradient descent to guarantee that the initial training epochs of both the prior and posterior were based on the same training samples. They also used this freedom to study a scenario in which the data-dependent prior was used as a regularizer term during optimization. Further exploiting the flexibility of this scenario is a promising direction, and connections to flatness-seeking optimization procedures such as sharpness-aware minimization [30] may be fruitful to explore further.

Bibliography

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

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," May 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[3] 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, "Mastering chess and shogi by self-play with a general reinforcement learning algorithm," Dec. 2017. [Online]. Available: https://arxiv.org/abs/1712.01815

[4] 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.

[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] T. Steinke and L. Zakynthinou, "Reasoning about generalization via conditional mutual information," *Conf. Learn Theory (COLT)*, vol. 125, pp. 1–16, July 2020.

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

[9] M. Dundar, B. Krishnapuram, J. Bi, and R. Rao, "Learning classifiers when the training data is not iid." in *IJCAI Inter. Joint Conf. on Artif. Intell.*, Jan. 2007, pp. 756–761.

[10] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," Jun. 2020. [Online]. Available: https://arxiv.org/abs/1911.02685

[11] H. Peng, "A comprehensive overview and survey of recent advances in meta-learning," Oct. 2020. [Online]. Available: https://arxiv.org/abs/2004.11149

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

[13] B. Guedj and L. Pujol, "Still no free lunches: the price to pay for tighter PAC-Bayes bounds," *arXiv*, Oct. 2019. [Online]. Available: http://arxiv.org/abs/1910.04460

[14] 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.

[15] 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.

[16] 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, Jun 2019.

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

[18] 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.

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

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

[21] 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.

[22] 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, May 2019.

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

[24] 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, Dec. 2017.

[25] 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.

[26] A. Esposito, M. Gastpar, and I. Issa, "Generalization error bounds via Rènyi f-divergences and maximal leakage," *arXiv*, Dec. 2019. [Online]. Available: http://arxiv.org/abs/1912.01439

[27] 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

[28] 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.

[29] G. Dziugaite, K. Hsu, W. Gharbieh, and D. Roy, "On the role of data in PAC-Bayes bounds," June 2020. [Online]. Available: https://arxiv.org/abs/2006.10929

[30] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," Oct. 2020. [Online]. Available: https://arxiv.org/abs/2010.01412