Reasoning About Generalization via Conditional Mutual Information

Thomas Steinke CMI@THOMAS-STEINKE.NET
IBM Research – Almaden

Lydia Zakynthinou ZAKYNTHINOU.L@NORTHEASTERN.EDU
Khoury College of Computer Sciences, Northeastern University

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We provide an information-theoretic framework for studying the generalization properties of machine learning algorithms. Our framework ties together existing approaches, including uniform convergence bounds and recent methods for adaptive data analysis.

Specifically, we use Conditional Mutual Information (CMI) to quantify how well the input (i.e., the training data) can be recognized given the output (i.e., the trained model) of the algorithm. We show that bounds on CMI can be obtained from VC dimension, compression schemes, differential privacy, and other methods. We then show that bounded CMI implies various forms of generalization.

Keywords: Generalization, stability in data analysis, conditional mutual information

1. Introduction

How can we ensure that a machine learning system produces an output that generalizes to the underlying distribution, rather than overfitting its training data? That is, how can we ensure that the hypotheses or models that are produced are reflective of the underlying population the training data was drawn from, rather than patterns that occur only by chance in the training data? This is perhaps the fundamental question for the science of statistical machine learning.

A vast array of methods have been proposed to answer this question. Most notably, the theory of uniform convergence shows that, if the output is sufficiently "simple," then it cannot overfit too much. A more recent line of work has used distributional stability (in the form of differential privacy) to provide generalization guarantees that compose adaptively – that is, statistical validity is preserved even when a dataset is reused multiple times with each analysis being influenced by prior outcomes. Other methods for proving generalization include compression schemes and uniform stability.

Unfortunately, these different methods for providing generalization guarantees are largely disconnected from one another; it is, in general, not possible to compare or combine techniques. In this paper, we provide a framework to reason about many of these these differing approaches using the unifying language of information theory.

1.1. Background: Generalization

We consider the standard setting of statistical learning (Valiant, 1984; Haussler, 1992; Kearns et al., 1994). There is an unknown probability distribution $\mathcal{D}$ over some known set $\mathcal{Z}$. We have access to a

Extended abstract. Full version appears as [Steinke and Zakynthinou 2020, v3].

sample $Z \in \mathcal{Z}^n$ consisting of n independent draws from $\mathcal{D}$. Informally, our goal is to learn something about the underlying distribution $\mathcal{D}$ from the dataset Z. Formally, we have a function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$ and our goal is to find some $w_* \in \mathcal{W}$ that approximately minimizes $\ell(w_*, \mathcal{D}) := \underset{Z' \leftarrow \mathcal{D}}{\mathbb{E}} [\ell(w_*, Z')]$.[1]

Intuitively, w_* represents some hypothesis and $\ell(w_*, \mathcal{D})$ measures the veracity or quality of w_*. In supervised machine learning, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ represents pairs of feature vectors and labels and w_* represents a function $f_{w_*} : \mathcal{X} \to \mathcal{Y}$ that predicts the label given the features. Then ℓ is a "loss function." For example, the 0-1 loss measures the error rate of the predictor: $\ell(w_*, \mathcal{D}) = \underset{(X,Y) \leftarrow \mathcal{D}}{\mathbb{P}} [f_{w_*}(X) \neq Y]$, so minimizing $\ell(w_*, \mathcal{D})$ corresponds to finding the most accurate predictor.

However, we cannot evaluate the true loss (a.k.a. "population loss" or "risk") $\ell(w_*, \mathcal{D})$ since the distribution $\mathcal{D}$ is unknown. Instead we can compute the empirical loss (a.k.a. "empirical risk") $\ell(w_*, Z) := \frac{1}{n} \sum_{i=1}^n \ell(w_*, Z_i)$ using the sample Z. A natural learning strategy is "Empirical Risk Minimization (ERM)" – i.e., $w_* = \arg\min_{w \in \mathcal{W}} \ell(w, Z)$. The question of generalization is thus: How can we ensure that $\ell(w_*, Z) \approx \ell(w_*, \mathcal{D})$?

The classical theory of uniform convergence (Vapnik and Chervonenkis, 1971) approaches this problem by studying the class of functions $\mathcal{F} := \{\ell(w, \cdot) : w \in \mathcal{W}\}$. If we can show that $\sup_{w \in \mathcal{W}} |\ell(w, Z) - \ell(w, \mathcal{D})|$ is small with high probability for a random $Z \leftarrow \mathcal{D}^n$, then the question of generalization is answered. Such bounds can be obtained from combinatorial properties of $\mathcal{F}$, such as its Vapnik-Chervonenkis (VC) dimension (Vapnik and Chervonenkis, 1971; Talagrand, 1994; Alon et al., 1997) or its fat-shattering dimension (Kearns and Schapire, 1994; Bartlett et al., 1996).

Uniform convergence makes no reference to the algorithm; it depends only on its range $\mathcal{F}$. An algorithm may generalize better than uniform convergence would suggest (Shalev-Shwartz et al., 2009; Feldman, 2016). For example, it is common to add a regularizer to the ERM – that is, $w_* = \arg\min_{w \in \mathcal{W}} \ell(w, Z) + \lambda \|w\|$, where $\lambda > 0$ is a parameter and $\|w\|$ is a measure of the complexity of w. Thus we explicitly consider generalization to be a property of the learning algorithm $A : \mathcal{Z}^n \to \mathcal{W}$, which may or may not be randomized.

There are several ways to show that a specific algorithm A generalizes. Algorithms whose output essentially only depends on a few of the input data points, as formalized by compression schemes (Littlestone and Warmuth, 1986), can be shown to generalize. Uniform stability (Bousquet and Elisseeff, 2002) entails strong generalization bounds for algorithms where changing a single input datum does not change the loss of the algorithm too much. Similarly, differential privacy (Dwork et al., 2006) – a distributional notion of stability – entails generalization bounds (Dwork et al., 2015b; Bassily et al., 2016; Jung et al., 2019). In general, these various methods for proving generalization are incompatible and incomparable. This raises the question of whether it is possible to provide a unifying framework or language to study generalization.

1.1.1. (UNCONDITIONAL) MUTUAL INFORMATION

A recent line of work has studied generalization using mutual information and related quantities (Russo and Zou, 2016; Raginsky et al., 2016; Alabdulmohsin, 2016; Feldman and Steinke, 2018; Bassily et al., 2018; Dwork et al., 2015a; Rogers et al., 2016; Smith, 2017; Xu and Raginsky, 2017; Nachum and Yehudayoff, 2018; Nachum et al., 2018; Esposito et al., 2019; Bu et al., 2019, etc.). For a (possibly randomized) algorithm $A : \mathcal{Z}^n \to \mathcal{W}$ and a dataset $Z \leftarrow \mathcal{D}^n$, we consider the quantity $I(A(Z); Z)$, which measures how much information the output $A(Z)$ contains about its input Z.

[1] For simplicity, in this introduction we only consider ℓ to be a linear function (that is, only taking in a single element of $\mathcal{Z}$). Our methods readily extend to the more general case where $\ell : \mathcal{W} \times \mathcal{Z}^m \to \mathbb{R}$.

Bounded mutual information implies generalization: If $\ell : \mathcal{W} \times \mathcal{Z} \to [0,1]$, $A : \mathcal{Z}^n \to \mathcal{W}$ and $Z \leftarrow \mathcal{D}^n$, then $\left|\mathbb{E}\left[\ell(A(Z), Z) - \ell(A(Z), \mathcal{D})\right]\right| \le \sqrt{\frac{2}{n} \cdot I(A(Z); Z)}$ (Russo and Zou, 2016; Xu and Raginsky, 2017). Bounds on mutual information can be obtained from differential privacy or from bounds on the entropy of the output of A. Specifically, if A is ε-differentially private, then $I(A(Z); Z) \le \frac{1}{2}\varepsilon^2 n$ (McGregor et al., 2010; Bun and Steinke, 2016). And we have the generic bound $I(A(Z); Z) \le H(A(Z)) \le \log |\mathcal{W}|$.

Unfortunately, mutual information can easily be infinite even in settings where generalization is easy to prove. Bassily, Moran, Nachum, Shafer, and Yehudayoff (Bassily et al., 2018; Nachum et al., 2018) showed that *any* proper and consistent learner A for threshold functions must have $I(A(Z); Z) \ge \Omega\left(\frac{\log\log|\mathcal{Z}|}{n^2}\right)$ when $Z \leftarrow \mathcal{D}^n$ for some worst-case distribution $\mathcal{D}$. The dependence on the size of the domain $\mathcal{Z} \subset \mathbb{R}$ is mild, but, if the domain is infinite, then the mutual information is unbounded. In contrast, the VC dimension of threshold functions is 1, which implies strong uniform convergence bounds even for infinite domains.

We remark that thresholds can be "embedded" into larger classes, such as higher-dimensional linear thresholds (halfspaces) or even neural networks. Thus these negative results for unconditional mutual information extend to those classes too. This strong negative result shows that *any* proper empirical risk minimizer for thresholds must have unbounded mutual information; it is easier to show that many *specific* natural algorithms and natural distributions have unbounded mutual information: Linear regression has unbounded mutual information (even in dimension 0 with Gaussian data, which is simply outputting the mean (Bu et al., 2019)). The most natural algorithms for thresholds have infinite mutual information for *any* continuous data distribution.

The fundamental issue with the mutual information approach is that even a single data point has infinite information content if the distribution is continuous. Meanwhile, an algorithm revealing a single data point is not an issue for generalization.

We address the shortcomings of the mutual information approach by moving to conditional mutual information. Our conditioning approach can be viewed as "normalizing" the information content of each data point to one bit. That is, an algorithm that reveals one data point only has conditional mutual information of one bit, even if the unconditional mutual information is infinite.

1.2. Our Contributions: Conditional Mutual Information (CMI)

We introduce the conditional mutual information (CMI) framework for reasoning about the generalization properties of machine learning algorithms. CMI is a quantitative property of an algorithm A and a distribution $\mathcal{D}$. (Note that it does *not* depend on the loss function of interest.)

Intuitively, CMI measures how well we can "recognize" the input (i.e., training data) given the output (i.e., trained model) of the algorithm. Recognizing the input is formalized by considering a "supersample" consisting of $2n$ independent draws from the distribution – namely the n input data points mixed with n "ghost" data points – and measuring how well it is possible to distinguish the true inputs from their ghosts.[2][3] (Note that the ghost samples are entirely hypothetical – they only exist in the analysis.) The supersample is randomly partitioned into the input and the ghost

[2] The so-called "ghost samples" symmetrization technique has been used to prove generalization and Rademacher complexity bounds from VC bounds since its inception Vapnik and Chervonenkis (1971). (The name is attributed to Luc Devroye.) This technique is an inspiration for our definition and terminology.

[3] This intuition for CMI should be contrasted with that for (unconditional) mutual information, which asks how much of the input we could reconstruct from the output without the prompt of a supersample.

samples. We then measure how much information the output reveals about this partition using mutual information, where we take the supersample to be known (i.e., we condition on the supersample and the unknown information is how it is partitioned).

We now state the formal definition of CMI:

Definition 1 (Conditional Mutual Information (CMI) of an Algorithm) *Let $A : \mathcal{Z}^n \to \mathcal{W}$ be a randomized or deterministic algorithm. Let $\mathcal{D}$ be a probability distribution on $\mathcal{Z}$ and let $\tilde{Z} \in \mathcal{Z}^{n\times 2}$ consist of $2n$ samples drawn independently from $\mathcal{D}$. Let $S \in \{0,1\}^n$ be uniformly random and independent from $\tilde{Z}$ and the randomness of A. Define $\tilde{Z}_S \in \mathcal{Z}^n$ by $(\tilde{Z}_S)_i = \tilde{Z}_{i,S_i+1}$ for all $i \in [n]$ – that is, $\tilde{Z}_S$ is the subset of $\tilde{Z}$ indexed by S.*

The conditional mutual information (CMI) of A with respect to $\mathcal{D}$ *is*

$$\mathsf{CMI}_{\mathcal{D}}\left(A\right) := I(A(\tilde{Z}_S); S|\tilde{Z}).$$

We remark on some basic properties of CMI. Firstly, $0 \leq \mathsf{CMI}_{\mathcal{D}}\left(A\right) \leq n \cdot \log 2$ for any A and any $\mathcal{D}$.[4] The case $\mathsf{CMI}_{\mathcal{D}}\left(A\right) = 0$ corresponds to the output of A being independent from its input, such as when A is a constant function. The other extreme, $\mathsf{CMI}_{\mathcal{D}}\left(A\right) = n \cdot \log 2$, corresponds to an algorithm that reveals all of its input, allowing arbitrary overfitting. Note that the CMI is always finite, which is in stark contrast with unconditional mutual information. Essentially, the conditioning normalizes the information content of each datum to one bit – an algorithm that reveals k of its input points and reveals nothing about the other $n-k$ inputs has a CMI of k bits.

For further intuition about the scale or units of CMI, we briefly mention how it relates to generalization error and other notions: Our generalization bounds become non-vacuous as soon as the CMI drops below $n/2$ nats. In terms of asymptotics, we obtain meaningful generalization bounds whenever the CMI is $o(n)$. More precisely, $\mathsf{CMI}_{\mathcal{D}}\left(A\right) = \varepsilon^2 n$ roughly corresponds to generalization error ε and is roughly a consequence of ε-differential privacy. We also have, for any $A : \mathcal{Z}^n \to \mathcal{W}$ and any $\mathcal{D}$, that $\mathsf{CMI}_{\mathcal{D}}\left(A\right) \leq H(A(Z)) \leq \log |\mathcal{W}|$, where $H(A(Z))$ is the Shannon entropy (Cover and Thomas, 2006) of the output of A on an input Z consisting of n i.i.d. draws from $\mathcal{D}$.

Finally, we note that CMI composes non-adaptively, i.e., if $A_1, A_2 : \mathcal{Z}^n \to \mathcal{W}$ are algorithms (whose internal sources of randomness are independent) then $\mathsf{CMI}_{\mathcal{D}}\left((A_1, A_2)\right) \leq \mathsf{CMI}_{\mathcal{D}}\left(A_1\right) + \mathsf{CMI}_{\mathcal{D}}\left(A_2\right)$ for all distributions $\mathcal{D}$. Moreover, CMI has the postprocessing property (as an immediate consequence of the data processing inequality for conditional mutual information). Namely, if $A : \mathcal{Z}^n \to \mathcal{W}$ and $B : \mathcal{W} \to \mathcal{W}'$ are algorithms (with independent internal sources of randomness), then $\mathsf{CMI}_{\mathcal{D}}\left(B(A(\cdot))\right) \leq \mathsf{CMI}_{\mathcal{D}}\left(A\right)$ for all distributions $\mathcal{D}$. This is an important robustness property (closely related to post-hoc generalization (Cummings et al., 2016; Nissim et al., 2018)).

1.2.1. GENERALIZATION FROM CMI

The key property of CMI is, of course, that it implies generalization. Since there is no single definition of generalization, we prove several consequences of CMI bounds.

The following theorem gives several consequences for bounded linear loss functions.

[4] We take log to denote the natural logarithm and, correspondingly, the units for information-theoretic quantitites are *nats*, instead of *bits*, where 1 bit equals $\log 2 \approx 0.7$ nats.

Theorem 2 *Let $A : \mathcal{Z}^n \to \mathcal{W}$ and $\ell : \mathcal{W} \times \mathcal{Z} \to [0,1]$. Let $\mathcal{D}$ be a distribution on $\mathcal{Z}$ and define $\ell(w,\mathcal{D}) = \underset{Z \leftarrow \mathcal{D}}{\mathbb{E}}[\ell(w,Z)]$ and $\ell(w,z) = \frac{1}{n}\sum_{i=1}^n \ell(w,z_i)$ for all $w \in \mathcal{W}$ and $z \in \mathcal{Z}^n$. Then*

$$\left|\underset{Z \leftarrow \mathcal{D}^n, A}{\mathbb{E}}\left[\ell(A(Z),Z) - \ell(A(Z),\mathcal{D})\right]\right| \le \sqrt{\frac{2}{n} \cdot \mathsf{CMI}_{\mathcal{D}}(A)}, \tag{1}$$

$$\underset{Z \leftarrow \mathcal{D}^n, A}{\mathbb{E}}\left[\left(\ell(A(Z),Z) - \ell(A(Z),\mathcal{D})\right)^2\right] \le \frac{3 \cdot \mathsf{CMI}_{\mathcal{D}}(A) + 2}{n}, \tag{2}$$

$$\underset{Z \leftarrow \mathcal{D}^n, A}{\mathbb{E}}\left[\ell(A(Z),\mathcal{D})\right] \le 2 \cdot \underset{Z \leftarrow \mathcal{D}^n, A}{\mathbb{E}}\left[\ell(A(Z),Z)\right] + \frac{3}{n} \cdot \mathsf{CMI}_{\mathcal{D}}(A). \tag{3}$$

The first part of the theorem (1) is the simplest bound; it relates the expected empirical loss to the expected true loss. The second part (2) gives a bound on the expected squared difference between these quantities; this bound is qualitatively strictly stronger, but quantitatively weaker by (small) constants. The final part of the theorem deals with the realizeable (or overfitted) case where the empirical loss is zero or close to zero (sometimes this referred to as the interpolating setting); when $\underset{Z \leftarrow \mathcal{D}^n, A}{\mathbb{E}}[\ell(A(Z),Z)] \approx 0$ this yields a bound that is quadratically sharper than the other bounds.

We also have a result for unbounded loss functions:

Theorem 3 *Let $A : \mathcal{Z}^n \to \mathcal{W}$ and $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$. Let $\mathcal{D}$ be a distribution on $\mathcal{Z}$. Then*

$$\left|\underset{Z \leftarrow \mathcal{D}^n, A}{\mathbb{E}}\left[\ell(A(Z),Z) - \ell(A(Z),\mathcal{D})\right]\right| \le \sqrt{\frac{8}{n} \cdot \mathsf{CMI}_{\mathcal{D}}(A) \cdot \underset{Z' \leftarrow \mathcal{D}}{\mathbb{E}}\left[\sup_{w \in \mathcal{W}} \left(\ell(w,Z')\right)^2\right]}. \tag{4}$$

The final term in the bound (4) gives some scale for the loss function. It is necessary to make some kind of assumption on the losses, such as bounded moments. As an application, this allows us to derive generalization bounds for squared loss (i.e., mean squared error) or hinge loss.

In our proofs, we make heavy use of the following lemma.

Lemma 4 (Gray 2011, Thm. 5.2.1,van Handel 2014, Lem. 4.10) *Let X and Y be random variables on Ω (with X absolutely continuous with respect to Y) and $f : \Omega \to \mathbb{R}$ a (measurable) function. Then*

$$\mathbb{E}[f(X)] \le \mathrm{D}(X\|Y) + \log \mathbb{E}\left[e^{f(Y)}\right].$$

We mostly use the lemma above as follows.

Corollary 5 *Let S, S', and Z be independent random variables where S and S' have identical distributions. Let A be a random function whose randomness is independent from S, S', and Z. Let g be a fixed function. Then*

$$\underset{A,S,Z}{\mathbb{E}}\left[g(A(S,Z),S,Z)\right] \le \inf_{t>0} \frac{I(A(S,Z);S|Z) + \underset{Z}{\mathbb{E}}\left[\log \underset{A,S,S',Z}{\mathbb{E}}\left[e^{t \cdot g(A(S,Z),S',Z)}\right]\right]}{t}.$$

This follows from Lemma 4 by setting $X = (A(S,Z),S,Z)$, $Y = (A(S,Z),S',Z)$, $f((y,s,z)) = t \cdot g(y,s,z)$ and by the definition

$$I(A(S,Z);S|Z) = \underset{Z}{\mathbb{E}}\left[I(A(S,Z);S)\right] = \underset{Z}{\mathbb{E}}\left[\mathrm{D}\left((A(S,Z),S)\middle\|(A(S,Z),S')\right)\right].$$

This allows us to bound $\mathbb{E}\left[g(A(S,Z),S,Z)\right]$ – the object of interest – in terms of the conditional mutual information $I(A(S,Z);S|Z)$ and the moment generating function $\mathbb{E}\left[e^{t\cdot g(A(S,Z),S',Z)}\right]$. Here, we prove the first part of Theorem 2 (1), which illustrates the key steps in all of these proofs.

Proof [Proof of Theorem 2(1)] Let $f_{\tilde{z}}(w,s) = \ell(w,\tilde{z}_s) - \ell(w,\tilde{z}_{\overline{s}})$, where $\overline{s}$ denotes the complement of s so that $\tilde{z}_{\overline{s}}$ is the elements of $\tilde{z}$ not selected in $\tilde{z}_s$. Let $W = A(\tilde{Z}_S)$. Let S' be an independent copy of S. Then

$$\begin{aligned}
&\underset{Z\leftarrow\mathcal{D}^n,A}{\mathbb{E}}\left[\ell(A(Z),Z) - \ell(A(Z),\mathcal{D})\right] = \underset{\tilde{Z},S,A}{\mathbb{E}}\left[\ell(A(\tilde{Z}_S),\tilde{Z}_S) - \ell(A(\tilde{Z}_S),\mathcal{D})\right] \\
&= \underset{\tilde{Z},S,A}{\mathbb{E}}\left[\ell(A(\tilde{Z}_S),\tilde{Z}_S) - \ell(A(\tilde{Z}_S),\tilde{Z}_{\overline{S}})\right] = \underset{\tilde{Z},S,A}{\mathbb{E}}\left[f_{\tilde{Z}}(A(\tilde{Z}_S),S)\right] \\
&\leq \inf_{t>0} \frac{I(A(\tilde{Z}_S);S|Z) + \underset{\tilde{Z}}{\mathbb{E}}\left[\log \underset{W,S'}{\mathbb{E}}\left[e^{tf_{\tilde{Z}}(W,S')}\right]\right]}{t} && \text{(by Corollary 5)} \\
&= \inf_{t>0} \frac{\mathsf{CMI}_{\mathcal{D}}(A) + \underset{\tilde{Z}}{\mathbb{E}}\left[\log \underset{W}{\mathbb{E}}\left[\prod_{i=1}^n \underset{S'_i}{\mathbb{E}}\left[e^{\frac{t}{n}(\ell(W,(\tilde{Z}_{S'})_i)-\ell(W,(\tilde{Z}_{\overline{S'}})_i))}\right]\right]\right]}{t} && \text{(by independence)} \\
&= \inf_{t>0} \frac{\mathsf{CMI}_{\mathcal{D}}(A) + \underset{\tilde{Z}}{\mathbb{E}}\left[\log \underset{W}{\mathbb{E}}\left[\prod_{i=1}^n \underset{S'_i}{\mathbb{E}}\left[e^{\frac{t}{n}(1-2S'_i)(\ell(W,\tilde{Z}_{i,1})-\ell(W,\tilde{Z}_{i,2}))}\right]\right]\right]}{t} \\
&\leq \inf_{t>0} \frac{\mathsf{CMI}_{\mathcal{D}}(A) + \underset{\tilde{Z}}{\mathbb{E}}\left[\log \underset{W}{\mathbb{E}}\left[\prod_{i=1}^n e^{\frac{t^2}{2n^2}(\ell(W,\tilde{Z}_{i,1})-\ell(W,\tilde{Z}_{i,2}))^2}\right]\right]}{t} && \text{(by Hoeffding's Lemma)} \\
&\leq \inf_{t>0} \frac{\mathsf{CMI}_{\mathcal{D}}(A) + \underset{\tilde{Z}}{\mathbb{E}}\left[\log \underset{W}{\mathbb{E}}\left[\prod_{i=1}^n e^{\frac{t^2}{2n^2}}\right]\right]}{t} && \text{(since } \ell \text{ is bounded in } [0,1]) \\
&= \inf_{t>0} \frac{\mathsf{CMI}_{\mathcal{D}}(A) + \frac{t^2}{2n}}{t} = \sqrt{\frac{2}{n}\cdot\mathsf{CMI}_{\mathcal{D}}(A)}.
\end{aligned}$$

■

Furthermore, we are able to extend our generalization results to non-linear loss functions. As an example application, we derive the following generalization bound for the Area Under the ROC Curve (AUC/AUROC) statistic, which is a commonly-used non-linear statistic for measuring the performance of a classifier. Specifically, for a classifier $f : \mathcal{Z} \to \mathbb{R}$ that produces a numerical score or probability (rather than just a binary label), the AUROC is the probability that a random positive example has a higher score than a random negative example – i.e., $\mathsf{AUROC}(f,\mathcal{D}) := \underset{(Z_+,Z_-)\leftarrow\mathcal{D}^2}{\mathbb{P}}\left[f(Z_+) > f(Z_-)|Z_+ \in \mathcal{Z}_+, Z_- \notin \mathcal{Z}_+\right]$, where $\mathcal{Z}_+$ is the set of positive examples.

Theorem 6 *Let $\mathcal{D}$ be a distribution on $\mathcal{Z}$. Let $\mathcal{Z}_+ \subseteq \mathcal{Z}$ be the set of positive examples and assume $0 < p := \underset{Z\leftarrow\mathcal{D}}{\mathbb{E}}\left[Z \in \mathcal{Z}_+\right] < 1$. Let $A : \mathcal{Z}^n \to \mathcal{W}$ be a randomized algorithm (whose randomness is independent from its input). If $n \geq O\left(\frac{1}{p(1-p)}\log\left(\frac{1}{p(1-p)}\right)\right)$, then, for any $\varepsilon \in (0,1)$,*

$$\underset{Z\leftarrow\mathcal{D}^n,A}{\mathbb{P}}\left[\left|\mathsf{AUROC}(A(Z),Z) - \mathsf{AUROC}(A(Z),\mathcal{D})\right| \leq \varepsilon\right] \geq 1 - O\left(\frac{\mathsf{CMI}_{\mathcal{D}}(A)}{\varepsilon^2 p(1-p)n}\right).$$

The above bounds illustrate how we are able to derive a great variety of generalization bounds from a single CMI bound. This versatility is a key strength of the CMI framework.

We note that although most stated bounds are on the expectation of the generalization error (or its square), these can be converted into probability bounds via Markov's inequality. For example, Theorem 2(2) implies $\underset{Z \leftarrow \mathcal{D}^n, A}{\mathbb{P}}[|\ell(A(Z), Z) - \ell(A(Z), \mathcal{D})| \ge \varepsilon] \le \frac{3 \cdot \mathsf{CMI}_{\mathcal{D}}(A)+2}{\varepsilon^2 n}$ for all $\varepsilon > 0$. However, this does not yield "high probability" bounds – that is, the failure probability decays polynomially with the desired error bound ε, rather than exponentially.

1.2.2. OBTAINING CMI BOUNDS

We show that a variety of known methods for proving generalization fit into our framework, by proving that they imply bounds on the CMI of the algorithm. Indeed, analysing these algorithms via CMI, versus a direct generalization analysis, yields essentially the same bound. These connections demonstrate the unifying nature of the CMI framework.

Compression Schemes First, we prove that, if an algorithm $A : \mathcal{Z}^n \to \mathcal{W}$ has a compression scheme of size k (Littlestone and Warmuth, 1986), then $\mathsf{CMI}_{\mathcal{D}}(A) \le O(k \cdot \log n)$. Intuitively, this is in agreement with the fact that an algorithm blatantly revealing k of the input points and nothing about the rest would have a CMI of k bits.

Theorem 7 *Let $A_1 : \mathcal{Z}^n \to \mathcal{Z}^k$ have the property that $A_1(z) \subset z$ for all z. Let $A_2 : \mathcal{Z}^k \to \mathcal{W}$ be arbitrary and let $A : \mathcal{Z}^n \to \mathcal{W}$ satisfy $A(z) = A_2(A_1(z))$ for all z. Then $\mathsf{CMI}_{\mathcal{D}}(A) \le O(k \log n)$ for all distributions $\mathcal{D}$.*

Proof Let $K = K(z) = \{i_1, \ldots, i_k\} \subset [n]$ denote the set of indices chosen by the compression algorithm A_1 on input z. We will slightly abuse notation and denote by $z_K \in \mathcal{Z}^k$ the subset of $z \in \mathcal{Z}^n$ given by the indices $K \subset [n]$. So $A(z) = z_{K(z)} = (z_{i_1}, z_{i_2}, \cdots, z_{i_k}) \subset z$. For $\tilde{Z} \in \mathcal{Z}^{n \times 2}$ consisting of $2n$ samples drawn independently from $\mathcal{D}$ and $S \in \{0,1\}^n$ uniformly random and independent from $\tilde{Z}$:

$$\mathsf{CMI}_{\mathcal{D}}(A) = I(A_2(A_1(\tilde{Z}_S)); S|\tilde{Z}) \le I((\tilde{Z}_S)_K; S|\tilde{Z}) \le H(K|\tilde{Z}) \le H(K) \le k \log(2n).$$

The first inequality follows from the data-processing inequality, the second by the definition of mutual information on distributions over discrete domains, and the last inequality holds since the number of possible distinct values of $(\tilde{Z}_S)_K$ given $\tilde{Z}$ is at most $\binom{2n}{k} \le (2n)^k$. ■

Uniform Convergence & VC Dimension Next, we show a connection between uniform convergence and CMI. We consider hypothesis classes $\mathcal{W}$ consisting of functions $h : \mathcal{X} \to \{0,1\}$ and we consider the standard 0-1 loss $\ell : \mathcal{W} \times (\mathcal{X} \times \{0,1\}) \to \{0,1\}$ with $\ell(h,(x,y)) = 0 \Leftrightarrow h(x) = y$. Bounded VC dimension is a necessary (Vapnik and Chervonenkis, 1971) and sufficient (Talagrand, 1994) condition for uniform convergence (for worst-case distributions) and is hence a sufficient condition for generalization. Note that CMI is a property which depends on the algorithm, whereas the VC dimension is a property of the output space; this appears to cause an incompatibility between the two methods. Nonetheless, we connect the two by proving that, for any VC hypothesis class, there always *exists* an empirical risk minimization algorithm $A : \mathcal{Z}^n \to \mathcal{W}$ with bounded CMI:

Theorem 8 *Let $\mathcal{Z} = \mathcal{X} \times \{0,1\}$ and let $\mathcal{H} = \{h : \mathcal{X} \to \{0,1\}\}$ be a hypothesis class with VC dimension d. Then, there exists an empirical risk minimizer $A : \mathcal{Z}^n \to \mathcal{H}$ for the 0-1 loss such that $\mathsf{CMI}_D\left(A\right) \le O(d \log n)$ for all distributions $\mathcal{D}$.*

We prove this theorem by showing that any algorithm satisfying a consistency property described next has bounded CMI and that there always exists an empirical risk minimizer with this consistency property. Intuitively, the consistency property we require says the following. Suppose the algorithm is run on some labelled dataset (x, y) to obtain an output hypothesis $h = A(x, y)$. If the dataset is relabelled to be perfectly consistent with h, then the algorithm should still output h – i.e., $A(x, h(x)) = h$. This should also hold if further examples are added to the dataset (where the additional examples are also consistent with h) – i.e., $A(x', h(x')) = h$ when $x \subset x'$. This is a very natural and reasonable consistency property.

Note that it is not true that *every* empirical risk minimizer for a class of bounded VC dimension has bounded CMI; if there are multiple minimizers to choose from, a pathological algorithm could encode superfluous information about the input in its output using this choice (thus violating our consistency property). We also remark that this bound is tight up to the $\log n$ term; combining Theorems 8 and 2 yields generalization bounds that are tight up to this term. It is natural to ask whether this logarithmic term can be removed. We conjecture that it can be removed by instead considering an *approximate* empirical risk minimizer.

Obtaining CMI bounds in the case of compression schemes and VC dimension mainly reduces to observing that these two conditions effectively restrict the output space – that is, conditioned on the supersample $\tilde{Z}$, there are few possible outputs $\mathcal{W}_{\tilde{Z}} := \{A(\tilde{Z}_s) : s \in \{0,1\}^n\}$ and we can use the worst-case entropy bound $\log |\mathcal{W}_{\tilde{Z}}|$. In both cases, this results in a multiplicative factor of $\log n$ in the CMI bound. This logarithmic factor could potentially be eliminated given more information about the structure of the problem. We demonstrate two specific cases where tighter bounds can be obtained by taking into account assumptions on the algorithm or the distribution $\mathcal{D}$. First, we prove that there exists an empirical risk minimizer which learns threshold functions in the realizable case and has *constant* CMI, whereas the general result gives a bound of $O(\log n)$. Second, we consider the problem of learning parity functions on $\{0,1\}^d$ when $\mathcal{D}$ is the uniform distribution. Intuitively, this uniformity assumption on $\mathcal{D}$ ensures that, as the number of samples n increases, with high probability there will be only a single consistent hypothesis. This allows us to prove that there exists an empirical risk minimizer whose CMI decreases to zero as n increases, namely $\mathsf{CMI}_{\mathcal{D}}\left(A\right) \le O(n \cdot 2^{d-n})$.

Distributional Stability & Differential Privacy Finally, we show that distributional stability implies CMI bounds. Differential privacy is the most well-known form of distributional stability and its generalization properties are well-established (Dwork et al., 2015b; Bassily et al., 2016; Jung et al., 2019).

Theorem 9 *Let $A : \mathcal{Z}^n \to \mathcal{W}$ be a randomized algorithm. Any one of the following conditions imply that $\mathsf{CMI}_{\mathcal{D}}\left(A\right) \le \varepsilon n$ for any distribution $\mathcal{D}$.*

(i) A is $\sqrt{2\varepsilon}$-differentially private (Dwork et al., 2006).

(ii) A satisfies ε-concentrated differential privacy (Bun and Steinke, 2016).

(iii) A satisfies ε-average leave-one-out KL stability (Feldman and Steinke, 2018).

(iv) A is ε-TV stable (Bassily et al., 2016).[5]

We remark that, with the exception of TV stability, all of the conditions in Theorem 9 are known to imply bounds on (unconditional) mutual information. However, TV stability does not imply any bounds on mutual information, so this sets CMI apart. In particular, approximate differential privacy (a.k.a. (ε, δ)-differential privacy) implies TV stability and hence CMI bounds. Here, we prove the last part of the theorem, that is, a bound on the CMI of TV-stable algorithms.

Theorem 10 (Theorem 9(iv), CMI of TV stable algorithms) *An algorithm $A : \mathcal{Z}^n \to \mathcal{W}$ is δ-*TV stable *if, for any two data sets $z, z' \in \mathcal{Z}^n$ that differ in a single element,*

$$\mathrm{d}_{TV}(A(z), A(z')) := \sup_{W \subseteq \mathcal{W}} \mathbb{P}\left[A(z) \in W\right] - \mathbb{P}\left[A(z') \in W\right] \leq \delta.$$

If $A : \mathcal{Z}^n \to \mathcal{W}$ is a δ-TV stable algorithm then $\mathsf{CMI}_{\mathcal{D}}(A) \leq \delta n$ for any distribution $\mathcal{D}$ over $\mathcal{Z}$.

Proof Let $\tilde{z}^* = \mathrm{argmax}_{\tilde{z} \in \mathcal{Z}^{n \times 2}} I(A(\tilde{z}_S); S)$ and $S \leftarrow \mathcal{U}^n$. Let us denote $A(\tilde{z}^*_s)$ by $F(s)$ for $s \in \{0,1\}^n$. Then

$$\mathsf{CMI}(A) = I(A(\tilde{z}^*_S); S) = I(F(S); S)$$

and it suffices to prove that $I(F(S); S) \leq \delta n$ for $S \leftarrow \mathcal{U}^n$. Let us define $S_{<i} = (S_1, \ldots, S_{i-1})$, $S_{>i} = (S_{i+1}, \ldots, S_n)$, $S_{-i} = S_{<i} \circ S_{>i}$ and $S_{\leq i} = S_{<i} \circ S_i$, where $x \circ y$ denotes the concatenation of x with y. By the chain rule for mutual information and by induction,

$$I(F(S); S) = \sum_{i=1}^{n} I(F(S); S_i | S_{<i}). \tag{5}$$

By applying the chain rule on $I(F(S), S_{>i}; S_i | S_{<i})$, for a fixed $i \in [n]$, we get

$$\begin{aligned}
&I(F(S); S_i|S_{<i}) + I(S_{>i}; S_i|S_{<i}, F(S)) = I(S_{>i}; S_i|S_{<i}) + I(F(S); S_i|S_{<i}, S_{>i}) \\
&\Leftrightarrow I(F(S); S_i|S_{<i}) = I(S_{>i}; S_i|S_{<i}) + I(F(S); S_i|S_{-i}) - I(S_{>i}; S_i|S_{<i}, F(S)) \\
&\Leftrightarrow I(F(S); S_i|S_{<i}) = I(F(S); S_i|S_{-i}) - I(S_{>i}; S_i|S_{<i}, F(S)) \quad (S_i, S_{<i}, S_{>i} \text{ are independent}) \\
&\Rightarrow I(F(S); S_i|S_{<i}) \leq I(F(S); S_i|S_{-i}). \qquad (I(\cdot, \cdot) \geq 0)
\end{aligned}$$

By inequality (5), it follows that $I(F(S); S) \leq \sum_{i=1}^{n} I(F(S); S_i|S_{-i})$ and it suffices to prove that

$$\forall i \in [n] \quad I(F(S); S_i|S_{-i}) \leq \delta.$$

For any $i \in [n]$, let $s^*_{-i} = \mathrm{argmax}_{x \in \{0,1\}^{n-1}} I(F(S)|S_{-i} = x; S_i)$. Now, let us denote the random variable $F(S)|S_{-i} = s^*_{-i}$ by $F_i(S_i)$. Then, for all $i \in [n]$,

$$I(F(S); S_i|S_{-i}) = \underset{s_{-i} \leftarrow \mathcal{U}^{n-1}}{\mathbb{E}} \left[I(F(S)|S_{-i} = s_{-i}; S_i)\right] \leq I(F(S)|S_{-i} = s^*_{-i}; S_i) = I(F_i(S_i); S_i).$$

Therefore, it suffices to prove that, for uniformly random S,

$$\forall i \in [n] \quad I(F_i(S_i); S_i) \leq \delta.$$

[5] ε-TV stability is equivalent to $(0, \varepsilon)$-differential privacy.

For the rest of this proof we fix an arbitrary $i \in [n]$. The relevant property of F_i implied by TV stability is that $\mathrm{d}_{TV}(F_i(0), F_i(1)) \le \delta$.

Let us denote by P_0 and P_1 the probability distributions of $F_i(0)$ and $F_i(1)$, respectively. We denote their convex combination by $\frac{P_0+P_1}{2}$. By the definition of mutual information, we have that

$$\begin{aligned} I(F_i(S_i); S_i) &= \mathop{\mathbb{E}}_{r \leftarrow \mathcal{U}} \left[\mathrm{D}\left(F_i(r) \| F_i(S_i)\right)\right] = \frac{1}{2}\mathrm{D}\left(F_i(0) \| F_i(S_i)\right) + \frac{1}{2}\mathrm{D}\left(F_i(1) \| F_i(S_i)\right) \\ &= \frac{1}{2}\mathrm{D}\left(P_0 \middle\| \frac{P_0+P_1}{2}\right) + \frac{1}{2}\mathrm{D}\left(P_1 \middle\| \frac{P_0+P_1}{2}\right). \end{aligned}$$

The quantity $\frac{1}{2}\mathrm{D}\left(P_0 \middle\| \frac{P_0+P_1}{2}\right) + \frac{1}{2}\mathrm{D}\left(P_1 \middle\| \frac{P_0+P_1}{2}\right)$ is known as the Jensen-Shannon divergence, denoted by $\mathrm{JSD}(P_0 \| P_1)$ and it is known that it can be bounded by TV distance (Lin, 1991, Thm. 3):

$$\mathrm{JSD}(P_0 \| P_1) \le \mathrm{d}_{TV}(F_i(0), F_i(1)).$$

Now, recall that since $A : \mathcal{Z}^n \to \mathcal{W}$ is δ-TV stable, $F : \{0,1\}^n \leftarrow \mathcal{W}$ is also δ-TV stable, and for any $i \in [n]$ $F_i : \{0,1\} \to \mathcal{W}$ is δ-TV stable. Thus, for all $i \in [n]$, $I(F_i(S_i); S_i) \le \mathrm{d}_{TV}(F_i(0), F_i(1)) \le \delta$, which, as we argued, suffices to conclude that $\mathsf{CMI}(A) \le \delta n$. ■

1.3. Related Work, Limitations, & Further Work

Information Theory & Generalization Generalization is a very well-studied subject and several connections to information theory have been made. Some of these connections are orthogonal to our work; for example, the information bottleneck method (Tishby et al., 2000) considers the mutual information between the input/output of the classifier (rather than the training algorithm) and various intermediate representations internal to the classifier.

Various recent works have considered the mutual information between the input and output of the training algorithm to derive generalization bounds; see the discussion in Section 1.1.1. This line of work is the inspiration and starting point for our work. CMI extends this line of work. In particular, we are able to incorporate VC dimension into our framework, whereas prior works (Bassily et al., 2018; Nachum et al., 2018) showed that this was impossible for (unconditional) mutual information.

Other extensions of the mutual information approach have been proposed. Inspired by generic chaining (a stochastic process theory methodology closely related to uniform convergence), Asadi et al. (2018) consider the mutual information between the input of the algorithm and an *approximation* of its output (or, rather, a sequence of increasingly tight approximations of its output). This method provides tighter generalization bounds, but requires analysis of the geometry of the output space.

Another approach is to consider the mutual information between a single (but arbitrary) input datum and the output (Raginsky et al., 2016; Wang et al., 2016; Bu et al., 2019; Haghifam et al., 2020). If we consider the mutual information between a single datum and the output conditioned on the rest of the data (i.e., $I(A(Z); Z_i | Z_{-i})$, where $Z_{-i} = (Z_1, \cdots, Z_{i-1}, Z_{i+1}, \cdots, Z_n)$), then this implies bounds on the overall mutual information (i.e., $I(A(Z); Z) \le \sum_{i=1}^n I(A(Z); Z_i | Z_{-i})$) (Feldman and Steinke, 2018, Lem. 3.7). If we do not condition on the rest of the data, then the reverse inequality holds (i.e., $I(A(Z); Z) \ge \sum_{i=1}^n I(A(Z); Z_i)$) (Bu et al., 2019, Eq. 17) and it is possible to obtain sharper bounds than via the overall mutual information (Bu et al., 2019; Haghifam et al., 2020). We believe that further exploration in this direction is warranted (in particular, by combining this single-datum approach with our conditioning approach).

Negrea et al. (2019) study the mutual information between the output of an algorithm and a random subset of its input dataset. This is very similar to our CMI definition. This is used to provide generalization guarantees for Stochastic Gradient Langevin Dynamics (SGLD). Overall, their results are incomparable to ours, since they exploit the random subset method in a different manner – they consider the "disintegrated mutual information" (in essence this is a random variable whose expectation is the conditional mutual information and each realization is the mutual information conditioned on a fixed value of the subset). However, their techniques can be combined with ours to yield even tighter bounds Haghifam et al. (2020).

PAC-Bayesian bounds (McAllester, 1999) also relate information-theoretic quantities to generalization and are similar to the mutual information approach. These bounds are usually output-dependent – that is, they give a generalization bound for a particular output hypothesis or hypothesis distribution, rather than uniformly bounding the expected error of the algorithm. Such output-dependent bounds may be stronger and output-independent results can be obtained by averaging over outputs. PAC-Bayesian bounds can be used to analyze and interpret regularization. Hellström and Durisi (2020) extend our generalization bounds for bounded loss to the PAC-Bayesian setting, as an application of their unifying approach to deriving information-theoretic generalization bounds.

High Probability Generalization The generalization implied by CMI (Section 1.2.1) does not yield "high probability" guarantees – that is, to guarantee failure probability δ, the error tolerance must grow polynomially in $1/\delta$, whereas polylogarithmic growth would be desireable. This is an inherent limitation of the CMI framework – mutual information is an expectation and is thus not very sensitive to low-probability failures. In particular, an algorithm that does something "good" (e.g., output a fixed hypothesis) with probability $1 - p$ and something "bad" (e.g., output a hypothesis entirely overfitted to the dataset) with probability p has CMI $\approx pn$. Due to this sort of pathological example, CMI bounds cannot guarantee a failure probability lower than CMI$/n$.

An interesting direction for further work would be to extend the CMI framework so that it yields high probability bounds. This would require moving from conditional mutual information to something like approximate max information (Rogers et al., 2016) or Rényi mutual information (Esposito et al., 2020). However, we note that it is almost always possible to obtain high probability guarantees by repetition to amplify the success probability.

Loss Stability/Uniform Stability A long line of work (Rogers and Wagner, 1978; Devroye and Wagner, 1979; Bousquet and Elisseeff, 2002; Feldman and Vondrák, 2019; Dagan and Feldman, 2019, etc.) has proven generalization bounds by showing that various algorithms have the property that their loss changes very little if a single input datum is replaced or removed. This is a very beautiful and well-developed theory that provides a unifying framework for generalization. Uniform stability (one of the strongest and most well-studied variants of loss stability) has the advantage that it readily yields high probability generalization bounds.

However, one limitation of the loss stability approach is that the loss function is an integral part of the definition, whereas CMI and distributional stability notions do not depend on the loss function. Thus loss stability lacks the postprocessing robustness property and does not yield the same variety of generalization bounds as CMI does. Loss stability is typically defined for deterministic algorithms and randomized algorithms must be "derandomized" (such as by taking their expectation) to satisfy the definition. This is somewhat awkward and, arguably, the CMI framework is more elegant when handling randomized algorithms.

It is of course natural to ask whether loss stability and CMI can be unified in some way. We propose a variant of CMI (Evaluated CMI or eCMI) that takes the loss function into account and allows us to translate between the notions.

Adaptive Composition When a single dataset is analyzed multiple times and each analysis is informed by the outcome of earlier analyses, generalization may fail even if each individual step generalizes well. This phenomenon led to the study of generalization in adaptive data analysis (Hardt and Ullman, 2014; Dwork et al., 2015b; Steinke and Ullman, 2015; Dwork et al., 2015a; Bassily et al., 2016, etc.). In particular, differential privacy provides a method for guaranteeing generalization that composes adaptively – that is, running a sequence of algorithms, each of which is differentially private, on a single dataset results in a differentially private final output, even if each algorithm is given access to the output of previous algorithms. Also, the recent work of Ligett and Shenfeld (2019) introduced Local Statistical Stability, a notion based on the statistical distance between the prior distribution over the database elements and their posterior distribution conditioned on the output of the algorithm, which composes adaptively and yields high probability bounds.

Unfortunately, CMI does not compose adaptively. A challenge for further work is to fully integrate adaptive composition into some variant of the CMI framework. Towards this direction, we consider a variant of CMI (Universal CMI or uCMI) that does compose adaptively. This notion extends CMI by considering a worst-case supersample $\tilde{Z}$ and worst-case distribution over S but we can still show that we can obtain useful uCMI bounds from some of the notions we have tied into our framework.

More CMI bounds An immediate direction for further work is to improve our CMI bounds and to prove entirely new generalization bounds for algorithms such as stochastic convex optimization or even non-convex gradient methods. The value of the CMI framework will be demonstrated if it yields new insights, such as entirely new generalization results or simplifications of known bounds.

Notably, Haghifam et al. (2020) recently showed that the generalization bounds based on CMI are tighter than those based on mutual information. Moreover, by combining CMI with the idea of "disintegrated mutual information" (Negrea et al., 2019) and the single-datum mutual information approach (Bu et al., 2019), they give yet tighter bounds, which are later applied to yield improved generalization bounds for a Langevin dynamics algorithm. We believe that further exploration of the combination of these ideas with our conditioning approach is warranted.

Hellström and Durisi (2020) extended our average generalization bounds to the PAC-Bayes and single-draw settings, giving both data-independent and data-dependent bounds. These results are established as an application of their unifying approach to proving generalization bounds, which is based on an exponential inequality in terms of the information density between the output and the input. The authors also explore the effect of the conditioning approach to other measures, obtaining bounds in terms of the conditional versions of α-mutual information, Rényi divergence, and maximal leakage (the latter being tighter than the bound of its unconditional counterpart, in some cases).

Acknowledgments

We thank the anonymous reviewers for suggestions that helped improve this manuscript and led to a simplification of the proof of Theorem 10. Part of this work was done when LZ was an intern at IBM Research – Almaden. Otherwise, LZ was supported by NSF grants CCF-1718088 and CCF-1750640.

References

Ibrahim Alabdulmohsin. Uniform generalization, concentration, and adaptive learning. *arXiv preprint arXiv:1608.06072*, 2016.

Noga Alon, Shai Ben-David, Nicolo Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM (JACM)*, 44(4):615–631, 1997.

Amir Asadi, Emmanuel Abbe, and Sergio Verdú. Chaining mutual information and tightening generalization bounds. In *Advances in Neural Information Processing Systems*, pages 7234–7243, 2018.

Peter L Bartlett, Philip M Long, and Robert C Williamson. Fat-shattering and the learnability of real-valued functions. *journal of computer and system sciences*, 52(3):434–452, 1996.

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing*, STOC '16, pages 1046–1059, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4132-5. doi: 10.1145/2897518.2897566. URL http://doi.acm.org/10.1145/2897518.2897566.

Raef Bassily, Shay Moran, Ido Nachum, Jonathan Shafer, and Amir Yehudayoff. Learners that use little information. In Firdaus Janoos, Mehryar Mohri, and Karthik Sridharan, editors, *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 25–55. PMLR, 07–09 Apr 2018. URL http://proceedings.mlr.press/v83/bassily18a.html.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

Yuheng Bu, Shaofeng Zou, and Venugopal V Veeravalli. Tightening mutual information based bounds on generalization error. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 587–591. IEEE, 2019.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, New York, NY, USA, 2006. ISBN 0471241954.

Rachel Cummings, Katrina Ligett, Kobbi Nissim, Aaron Roth, and Zhiwei Steven Wu. Adaptive learning with robust generalization guarantees. In *Conference on Learning Theory*, pages 772–814, 2016.

Yuval Dagan and Vitaly Feldman. Pac learning with stable and private predictions. *arXiv preprint arXiv:1911.10541*, 2019.

Luc Devroye and Terry Wagner. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604, 1979.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, pages 265–284, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3-540-32731-2, 978-3-540-32731-8. doi: 10.1007/11681878_14. URL http://dx.doi.org/10.1007/11681878_14.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, pages 2350–2358, 2015a.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 117–126. ACM, 2015b.

Amedeo Roberto Esposito, Michael Gastpar, and Ibrahim Issa. A new approach to adaptive data analysis and learning via maximal leakage. *arXiv preprint arXiv:1903.01777*, 2019.

Amedeo Roberto Esposito, Michael Gastpar, and Ibrahim Issa. Robust generalization via α-mutual information. *arXiv preprint arXiv:2001.06399*, 2020.

Vitaly Feldman. Generalization of erm in stochastic convex optimization: The dimension strikes back. In *Advances in Neural Information Processing Systems*, pages 3576–3584, 2016.

Vitaly Feldman and Thomas Steinke. Calibrating noise to variance in adaptive data analysis. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 535–544. PMLR, 06–09 Jul 2018. URL http://proceedings.mlr.press/v75/feldman18a.html.

Vitaly Feldman and Jan Vondrák. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. *arXiv preprint arXiv:1902.10710*, 2019.

Robert M Gray. *Entropy and information theory*. Springer, 2011. URL https://ee.stanford.edu/~gray/it.html.

Mahdi Haghifam, Jeffrey Negrea, Ashish Khisti, Daniel M. Roy, and Gintare Karolina Dziugaite. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms, 2020.

Moritz Hardt and Jonathan Ullman. Preventing false discovery in interactive data analysis is hard. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 454–463. IEEE, 2014.

David Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and computation*, 100(1):78–150, 1992.

Fredrik Hellström and Giuseppe Durisi. Generalization bounds via information density and conditional information density, 2020.

Christopher Jung, Katrina Ligett, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Moshe Shenfeld. A new analysis of differential privacy's generalization guarantees. *arXiv preprint arXiv:1909.03577*, 2019.

Michael J Kearns and Robert E Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3):464–497, 1994.

Michael J Kearns, Umesh Virkumar Vazirani, and Umesh Vazirani. *An introduction to computational learning theory*. MIT press, 1994.

Katrina Ligett and Moshe Shenfeld. A necessary and sufficient stability notion for adaptive generalization. In *Advances in Neural Information Processing Systems*, pages 11481–11490, 2019.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

Nick Littlestone and Manfred K. Warmuth. Relating data compression and learnability. Technical report, 1986.

David A McAllester. Some pac-bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

Andrew McGregor, Ilya Mironov, Toniann Pitassi, Omer Reingold, Kunal Talwar, and Salil Vadhan. The limits of two-party differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 81–90. IEEE, 2010.

Ido Nachum and Amir Yehudayoff. Average-case information complexity of learning. *arXiv preprint arXiv:1811.09923*, 2018.

Ido Nachum, Jonathan Shafer, and Amir Yehudayoff. A direct sum result for the information complexity of learning. *arXiv preprint arXiv:1804.05474*, 2018.

Jeffrey Negrea, Mahdi Haghifam, Gintare Karolina Dziugaite, Ashish Khisti, and Daniel M Roy. Information-theoretic generalization bounds for sgld via data-dependent estimates. In *Advances in Neural Information Processing Systems*, pages 11013–11023, 2019.

Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. The limits of post-selection generalization. In *Advances in Neural Information Processing Systems*, pages 6400–6409, 2018.

Maxim Raginsky, Alexander Rakhlin, Matthew Tsao, Yihong Wu, and Aolin Xu. Information-theoretic analysis of stability and bias of learning algorithms. *2016 IEEE Information Theory Workshop (ITW)*, pages 26–30, 2016.

Ryan Rogers, Aaron Roth, Adam Smith, and Om Thakkar. Max-information, differential privacy, and post-selection hypothesis testing. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 487–494. IEEE, 2016.

William H Rogers and Terry J Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, pages 506–514, 1978.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 1232–1240, Cadiz, Spain, 09–11 May 2016. PMLR. URL http://proceedings.mlr.press/v51/russo16.html.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *COLT*, 2009.

Adam Smith. Information, privacy and stability in adaptive data analysis. *arXiv preprint arXiv:1706.00820*, 2017.

Thomas Steinke and Jonathan Ullman. Interactive fingerprinting codes and the hardness of preventing false discovery. In *Conference on Learning Theory*, pages 1588–1628, 2015.

Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information, 2020. URL https://arxiv.org/abs/2001.09122.

Michel Talagrand. Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, 22(1):28–76, 1994.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Leslie G Valiant. A theory of the learnable. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 436–445. ACM, 1984.

Ramon van Handel. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014.

V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, January 1971. doi: 10.1137/1116025. URL https://doi.org/10.1137/1116025.

Yu-Xiang Wang, Jing Lei, and Stephen E Fienberg. On-average kl-privacy and its equivalence to generalization for max-entropy mechanisms. In *International Conference on Privacy in Statistical Databases*, pages 121–134. Springer, 2016.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2524–2533, 2017.