Information-theoretic generalization bounds for black-box learning algorithms

Anonymous Author(s)
Affiliation
Address
email

Abstract

We derive information-theoretic generalization bounds for supervised learning algorithms based on the information contained in predictions rather that in the output of the training algorithm. These bounds improve over the existing information-theoretic bounds, are applicable to a wider range of algorithms, and solve two key challenges: (a) they give meaningful results for deterministic algorithms and (b) they are significantly easier to estimate. We show experimentally that the proposed bounds closely follow the generalization gap in practical scenarios for deep learning.

1 Introduction

Large neural networks trained with variants of stochastic gradient descent have excellent generalization capabilities, even in regimes where the number of parameters is much larger than the number of training examples. Zhang et al. [36] showed that classical generalization bounds based on various notions of complexity of hypothesis set fail to explain this phenomenon, as the same neural network can generalize well for one choice of training data and memorize completely for another one. This observation has spurred a tenacious search for algorithm-dependent and data-dependent generalization bounds that give meaningful results in practical settings for deep learning [14].

One line of attack bounds generalization error based on the information about training dataset stored in the weights [34, 2, 19, 4, 28, 11, 20, 23]. The main idea is that when the training and testing performance of a neural network are different, the network weights necessarily capture some information about the training dataset. However, the opposite might not be true: A neural network can store significant portions of training set in its weights and still generalize well [27, 35, 18]. Furthermore, because of their information-theoretic nature, these generalization bounds become infinite or produce trivial bounds for deterministic algorithms. When such bounds are not infinite, they are notoriously hard to estimate, due to the challenges arising in estimation of Shannon mutual information between two high-dimensional variables (e.g., weights of a ResNet and a training dataset).

This work addresses the aforementioned challenges. We first improve some of the existing information-theoretic generalization bounds, providing a unified view and derivation of them (Sec. 2). We then derive novel generalization bounds that measure information with predictions, rather than with the output of the training algorithm (Sec. 3). These bounds are applicable to a wide range of methods, including neural networks, Bayesian algorithms, ensembling algorithms, and non-parametric approaches. In the case of neural networks, the proposed bounds improve over the existing weight-based bounds, partly because they avoid a counter-productive property of weight-based bounds that information stored in *unused* weights affects generalization bounds, even though it has no effect on generalization. The proposed bounds produce meaningful results for deterministic algorithms and are significantly easier to estimate. For example, in case of classification, computing our most efficient bound involves estimating mutual information between a pair of predictions and a binary variable.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

We apply the proposed bounds to ensembling algorithms, binary classification algorithms with finite VC dimension hypothesis classes, and to stable learning algorithms (Sec. 4). We compute our most efficient bound on realistic classification problems involving neural networks, and show that the bound closely follows the generalization error, even in situations when a neural network with 3M parameters is trained deterministically on 4000 examples, achieving 1% generalization error.

2 Weight-based generalization bounds

We start by describing the necessary notation and definitions, after which we provide generalization of existing weight-based information-theoretic generalization bounds. All proofs are presented in Appendix A.

Preliminaries. We use capital letters for random variables, corresponding lowercase letters for their values, and calligraphic letters for their domains. If X is a random variable, $\bar{X}$ denotes an independent copy of X. For example, if (X, Y) is a pair of random variables with joint distribution $P_{X,Y}$, then the joint distribution of $(\bar{X}, \bar{Y})$ will be $P_{\bar{X},\bar{Y}} = P_{\bar{X}} \otimes P_{\bar{Y}} = P_X \otimes P_Y$. A random variable X is called σ-subgaussian if $\mathbb{E}\exp(t(X - \mathbb{E}\,X)) \leq \exp(\sigma^2 t^2/2),\ \forall t \in \mathbb{R}$. For example, a random variable that takes values in $[a, b]$ almost surely, is $(b-a)/2$-subgaussian. Given probability measures P and Q defined on the same measurable space, such that P is absolutely continuous with respect to Q, the Kullback–Leibler divergence from P to Q is defined as $\mathrm{KL}\left(P \parallel Q\right) = \int \log \frac{\mathrm{d}P}{\mathrm{d}Q} \mathrm{d}P$, where $\frac{\mathrm{d}P}{\mathrm{d}Q}$ is the Radon-Nikodym derivative of P with respect to Q. If X and Y are random variables defined on the same probability space, then $\mathrm{KL}\left(X \parallel Y\right)$ denotes $\mathrm{KL}\left(P_X \parallel P_Y\right)$. The Shannon mutual information between random variables X and Y is $I(X;Y) = \mathrm{KL}\left(P_{X,Y} \parallel P_X \otimes P_Y\right)$. In this paper, all information-theoretic quantities are measured in nats, instead of bits. Throughout the paper $[n]$ denotes the set $\{1, 2, \ldots, n\}$. Finally, if $A = (a_1, \ldots, a_n)$ is a collection, then $A_{-i} \triangleq (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n)$.

Theorems proved in the subsequent sections will be relying on the following lemma.

Lemma 1. *Let (Φ, Ψ) be a pair of random variables with joint distribution $P_{\Psi,\Phi}$. If $g(\phi, \psi)$ is a measurable function such that $\mathbb{E}_{\Phi,\Psi}\left[g(\Phi, \Psi)\right]$ exists and $g(\bar{\Phi}, \bar{\Psi})$ is σ-subgaussian, then*

$$\left|\mathbb{E}_{\Phi,\Psi}\left[g(\Phi, \Psi)\right] - \mathbb{E}_{\Phi,\bar{\Psi}}\left[g(\Phi, \bar{\Psi})\right]\right| \leq \sqrt{2\sigma^2 I(\Phi; \Psi)}. \tag{1}$$

Furthermore, if $g(\phi, \bar{\Psi})$ is σ-subgaussian for each ϕ and the expectation below exists, then

$$\mathbb{E}_{\Phi,\Psi}\left[\left(g(\Phi, \Psi) - \mathbb{E}_{\bar{\Psi}}\, g(\Phi, \bar{\Psi})\right)^2\right] \leq 4\sigma^2(I(\Phi; \Psi) + \log 3), \tag{2}$$

and

$$P\left(\left|g(\Phi, \Psi) - \mathbb{E}_{\bar{\Psi}}\, g(\Phi, \bar{\Psi})\right| \geq \epsilon\right) \leq \frac{4\sigma^2(I(\Phi; \Psi) + \log 3)}{\epsilon^2}, \quad \forall \epsilon > 0. \tag{3}$$

The first part of this lemma is equivalent to Lemma 1 of Xu and Raginsky [34], which in turn has its roots in Russo and Zou [24]. The second part generalizes Lemma 2 of Hafez-Kolahi et al. [10] by also providing bounds on the expected squared difference.

2.1 Generalization bounds with input-output mutual information

Let $S = (Z_1, Z_2, \ldots, Z_n) \sim \mathcal{D}^n$ be a dataset of n i.i.d. examples, $R \in \mathcal{R}$ be a source of randomness (a random variable independent of S) and $A : \mathcal{Z}^N \times \mathcal{R} \to \mathcal{W}$ be a training algorithm. Let $W = A(S, R)$ be the output of the training algorithm applied on the dataset S with randomness R. Given a loss function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$, the empirical risk is $\mathcal{L}_{\mathrm{emp}}(A, S, R) = \frac{1}{n}\sum_{i=1}^{n} \ell(W, Z_i)$ and the population risk is $\mathcal{L}(A, S, R) = \mathbb{E}_{Z' \sim D}\, \ell(W, Z')$, where Z' is a test example independent from S and R. The generalization gap, also call generalization error, is $\mathcal{L}(A, S, R) - \mathcal{L}_{\mathrm{emp}}(A, S, R)$. In this setting, Xu and Raginsky [34] establish the following information-theoretic bound on the absolute value of the expected generalization gap.

Theorem 2.1 (Thm. 1 of Xu and Raginsky [34])**.** *If $\ell(w, Z')$, where $Z' \sim \mathcal{D}$, is σ-subgaussian for all $w \in \mathcal{W}$, then*

$$\left|\mathbb{E}_{S,R}\left[\mathcal{L}(A, S, R) - \mathcal{L}_{\mathrm{emp}}(A, S, R)\right]\right| \leq \sqrt{\frac{2\sigma^2 I(W; S)}{n}}. \tag{4}$$

We generalize this result by showing that instead of measuring information with the entire dataset, one can measure information with a subset of size m chosen uniformly at random. For brevity, hereafter we call subsets chosen uniformly at random just "random subsets".

Theorem 2.2. *Let U be a random subset of $[n]$ with size m, independent of S and R. If $\ell(w, Z')$, where $Z' \sim \mathcal{D}$, is σ-subgaussian for all $w \in \mathcal{W}$, then*

$$|\mathbb{E}_{S,R}\left[\mathcal{L}(A,S,R) - \mathcal{L}_{\text{emp}}(A,S,R)\right]| \leq \mathbb{E}_{u\sim U}\sqrt{\frac{2\sigma^2}{m} I(W;S_u)}, \tag{5}$$

and

$$\mathbb{E}_{S,R}\left(\mathcal{L}(A,S,R) - \mathcal{L}_{\text{emp}}(A,S,R)\right)^2 \leq \frac{4\sigma^2}{n}\left(I(W;S) + \log 3\right). \tag{6}$$

With a simple application of Markov's inequality one can get tail bounds from the second part of the theorem. Note the bound on the squared generalization gap is written only for the case of $m = n$. It is possible to derive squared generalization gap bounds of form $\frac{4\sigma^2}{m}(\mathbb{E}_{u\sim U}\, I(W;S_u) + \log 3)$. Unfortunately, for small m the $\log 3$ constant starts to dominate, resulting in vacuous bounds.

Picking a small m decreases the mutual information term in (5), however, it also decreases the denominator. When setting $m = n$, we get the bound of Xu and Raginsky [34] (Thm. 2.1). When $m = 1$, the bound of (5) becomes $\frac{1}{n}\sum_{i=1}^{n}\sqrt{2\sigma^2 I(W;Z_i)}$, matching the result of Bu et al. [4] (Proposition 1). Bu et al. [4] prove that this bound is tighter than the bound with $I(W;S)$ (i.e., $m = n$). We generalize this result by proving that the bound of (5) is non-descreasing in m.

Proposition 1. *Let $m \in [n-1]$, U be a random subset of $[n]$ of size m, U' be a random subset of size $m+1$, and $\phi : \mathbb{R} \to \mathbb{R}$ be any non-decreasing concave function. Then*

$$\mathbb{E}_U\, \phi\left(\frac{1}{m} I(W;S_u)\right) \leq \mathbb{E}_{U'}\, \phi\left(\frac{1}{m+1} I(W;S_{u'})\right). \tag{7}$$

When $\phi(x) = \sqrt{x}$, this result proves that the optimal value for m in (5) is 1. Furthermore, when we use Jensen's inequality to move expectation over U inside the square root in (5), then the resulting bound becomes $\sqrt{\frac{2\sigma^2}{m}\,\mathbb{E}_{u\sim U}\, I(W;S_u)}$ and matches the result of Negrea et al. [19] (Thm. 2.3). These bounds are also non-decreasing with respect to m (using Proposition 1 with $\phi(x) = x$).

Thm. 2.1 can be used to derive generalization bounds that depend on the information between W and a single example Z_i conditioned on the remaining examples $Z_{-i} = (Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n)$.

Theorem 2.3. *If $\ell(w, Z')$, where $Z' \sim \mathcal{D}$, is σ-subgaussian for all $w \in \mathcal{W}$, then*

$$|\mathbb{E}_{S,R}\left[\mathcal{L}(A,S,R) - \mathcal{L}_{\text{emp}}(A,S,R)\right]| \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2\sigma^2 I(W;Z_i \mid Z_{-i})}, \tag{8}$$

and

$$\mathbb{E}_{S,R}\left(\mathcal{L}(A,S,R) - \mathcal{L}_{\text{emp}}(A,S,R)\right)^2 \leq \frac{4\sigma^2}{n}\left(\sum_{i=1}^{n} I(W;Z_i \mid Z_{-i}) + \log 3\right). \tag{9}$$

This theorem is a simple corollary of Thm. 2.2, using the facts that $I(W;Z_i) \leq I(W;Z_i \mid Z_{-i})$ and that $I(W;S)$ is upper bounded by $\sum_{i=1}^{n} I(W;Z_i \mid Z_{-i})$, which is also known as erasure information [30]. The first part of it improves the result of Raginsky et al. [22] (Thm. 2), as the averaging over i is outside of the square root. While these bounds are worse that the corresponding bounds of Thm. 2.2, it is sometimes easier to manipulate them analytically.

The bounds described above measure information with the output W of the training algorithm. In the case of prediction tasks with parametric methods, the parameters W might contain information about the training dataset, but not use it to make predictions. Partly for this reason, the main goal of this paper is to derive generalization bounds that measure information with the prediction function, rather than with the weights. In general, there is no straightforward way of encoding the prediction function into a random variable. However, when the domain $\mathcal{Z}$ is finite, we can encode the prediction function as the collection of predictions on all examples of $\mathcal{Z}$. This naturally leads us to the next setting (albeit with a different motivation), first considered by Steinke and Zakynthinou [28], where one first fixes a set of $2n$ examples, and then randomly selects n of them to form the training set. We use this setting to provide prediction-based generalization bounds in Sec. 3. Before describing these bounds we present the setting of Steinke and Zakynthinou [28] in detail and generalize some of the existing weight-based bounds in that setting.

2.2 Generalization bounds with conditional mutual information

Let $\tilde{Z} \in \mathcal{Z}^{n\times 2}$ be a collection of $2n$ i.i.d samples from $\mathcal{D}$, grouped in n pairs. The random variable $S \sim \text{Uniform}(\{0,1\}^n)$ specifies which example to select from each pair to form the training set $\tilde{Z}_S = (\tilde{Z}_{i,S_i})_{i=1}^n$. Let R be a random variable, independent of $\tilde{Z}$ and S, that captures the stochasticity of training. In this setting Steinke and Zakynthinou [28] defined condition mutual information (CMI) of algorithm A with respect to the data distribution $\mathcal{D}$ as

$$\text{CMI}_{\mathcal{D}}(A) = I(A(\tilde{Z}_S, R); S \mid \tilde{Z}) = \mathbb{E}_{\tilde{z}\sim\tilde{Z}}\, I(A(\tilde{z}_S, R); S), \tag{10}$$

and proved the following upper bound on expected generalization gap.

Theorem 2.4 (Thm. 2, Steinke and Zakynthinou [28])**.** *If the loss function $\ell(w,z) \in [0,1], \forall w \in \mathcal{W}, z \in \mathcal{Z}$, then the expected generalization gap can be bounded as follows:*

$$\left|\mathbb{E}_{\tilde{Z},S,R}\left[\mathcal{L}(A,\tilde{Z}_S,R) - \mathcal{L}_{\text{emp}}(A,\tilde{Z}_S,R)\right]\right| \leq \sqrt{\frac{2}{n}\text{CMI}_{\mathcal{D}}(A)}. \tag{11}$$

Haghifam et al. [11] improved this bound in two aspects. First, they provided bounds where expectation over $\tilde{Z}$ is outside of the square root. Second, they considered measuring information with subsets of S, as we did in the previous section.

Theorem 2.5 (Thm. 3.1 of Haghifam et al. [11])**.** *Let $m \in [n]$ and $U \subseteq [n]$ be a random subset of size m, independent from $R, \tilde{Z}$, and S. If the loss function $\ell(w,z) \in [0,1], \forall w \in \mathcal{W}, z \in \mathcal{Z}$, then*

$$\left|\mathbb{E}_{\tilde{Z},S,R}\left[\mathcal{L}(A,\tilde{Z}_S,R) - \mathcal{L}_{\text{emp}}(A,\tilde{Z}_S,R)\right]\right| \leq \mathbb{E}_{z\sim\tilde{Z}}\sqrt{\frac{2}{m}\,\mathbb{E}_{u\sim U}\, I(A(\tilde{z}_S,R);S_u)}. \tag{12}$$

Furthermore, for $m = 1$ they tighten the bound by showing that one can move the expectation over U outside of the squared root (Haghifam et al. [11], Thm 3.4). We generalize these results by showing that for all m expectation over U can be done outside of the square root. Furthermore, our proof closely follows the proof of Thm. 2.2.

Theorem 2.6. *Let $m \in [n]$ and $U \subseteq [n]$ be a random subset of size m, independent from $R, \tilde{Z}$, and S. If $\ell(w,z) \in [0,1], \forall w \in \mathcal{W}, z \in \mathcal{Z}$, then*

$$\left|\mathbb{E}_{\tilde{Z},S,R}\left[\mathcal{L}(A,\tilde{Z}_S,R) - \mathcal{L}_{\text{emp}}(A,\tilde{Z}_S,R)\right]\right| \leq \mathbb{E}_{\tilde{z}\sim\tilde{Z},u\sim U}\sqrt{\frac{2}{m}I(A(\tilde{z}_S,R);S_u)}, \tag{13}$$

and

$$\mathbb{E}_{\tilde{Z},S,R}\left(\mathcal{L}(A,\tilde{Z}_S,R) - \mathcal{L}_{\text{emp}}(A,\tilde{Z}_S,R)\right)^2 \leq \frac{8}{n}\left(\mathbb{E}_{\tilde{z}\sim\tilde{Z}}\, I(A(\tilde{z}_S,R);S) + 2\right). \tag{14}$$

The bound of (13) improves over the bound of Thm. 2.5 and matches the special result for $m = 1$. Haghifam et al. [11] showed that if one takes the expectations over $\tilde{Z}$ inside the square root in (12), then the resulting looser upper bounds become non-decreasing over m. Using this result they showed that their special case bound for $m = 1$ is the tightest. We generalize their results by showing that even without taking the expectations inside the squared root, the bounds of Thm. 2.5 are non-decreasing over m. We also show that the same holds for our tighter bounds of (13).

Proposition 2. *Let $m \in [n-1]$, U be a random subset of $[n]$ of size m, U' be a random subset of size $m+1$, $\tilde{z}$ be any fixed value of $\tilde{Z}$, and $\phi : \mathbb{R} \to \mathbb{R}$ be any non-decreasing concave function. Then*

$$\mathbb{E}_{u\sim U}\,\phi\left(\frac{1}{m}I(A(\tilde{z}_S,R);S_u)\right) \leq \mathbb{E}_{u'\sim U'}\,\phi\left(\frac{1}{m+1}I(A(\tilde{z}_S,R);S_{u'})\right). \tag{15}$$

By setting $\phi(x) = x$, taking square root of both sides of (15), and then taking expectation over $\tilde{z}$, we prove that bounds of (12) are non-decreasing over m. By setting $\phi(x) = \sqrt{x}$ and then taking expectation over $\tilde{z}$, we prove that bounds of (13) are non-decreasing with m.

Similarly to the Thm. 2.3 of the previous section, Thm. A.1 presented in Appendix A establishes generalization bounds with information-theoretic stability quantities.

3 Functional conditional mutual information

The bounds in Sec. 2 leverage information in the output of the algorithm, W. In this section we consider supervised learning problems: $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. To encompass many types of approaches, we do not assume that the training algorithm has an output W, which is then used to make predictions. Instead, we assume that the learning method implements a function $f : \mathcal{Z}^n \times \mathcal{X} \times \mathcal{R} \to \mathcal{K}$ that takes a training set z, a test input x', an auxiliary argument r capturing the stochasticity of training and predictions, and outputs a prediction $f(z, x', r)$ on the test example. Note that the prediction domain $\mathcal{K}$ can be different from $\mathcal{Y}$. This setting includes non-parametric methods (for which W is the training dataset itself), parametric methods, Bayesian algorithms, and more. For example, in parametric methods, where a hypothesis set $\mathcal{H} = \{h_w : \mathcal{X} \to \mathcal{K} \mid w \in \mathcal{W}\}$ is defined, $f(z, x, r) = h_{A(z,r)}(x)$.

In this supervised setting, the loss function $\ell : \mathcal{K} \times \mathcal{Y} \to \mathbb{R}$ measures the discrepancy between a prediction and a label. As in the previous subsection, we assume that a collection of $2n$ i.i.d examples $\tilde{Z} \sim \mathcal{D}^{n\times 2}$ is given, grouped in n pairs, and the random variable $S \sim \text{Uniform}(\{0,1\}^n)$ specifies which example to select from each pair to form the training set $\tilde{Z}_S = (\tilde{Z}_{i,S_i})_{i=1}^n$. Let R be an auxiliary random variable, independent of $\tilde{Z}$ and S, that provides stochasticity for predictions (e.g., in neural networks R can be used to make the training stochastic). The empirical risk of learning method f trained on dataset $\tilde{Z}_S$ with randomness R is defined as $\mathcal{L}_{\text{emp}}(f, \tilde{Z}_S, R) = \frac{1}{n}\sum_{i=1}^n \ell(f(\tilde{Z}_S, X_i, R), Y_i)$. The population risk is defined as $\mathcal{L}(f, \tilde{Z}_S, R) = \mathbb{E}_{Z'\sim\mathcal{D}}\, \ell(f(\tilde{Z}_S, X', R), Y')$. Before moving forward we adopt two conventions. First, if z is a collection of examples, then x and y denote the collection of its inputs and labels respectively. Second, if x is a collection of inputs, then $f(z, x, r)$ denotes the collection of predictions on x after training on z with randomness r.

We define functional conditional mutual information (f-CMI).

Definition 3.1. *Let $\mathcal{D}$, f, R, $\tilde{Z}$, S be defined as above and let $u \subseteq [n]$ be a subset of size m. Then* pointwise functional conditional mutual information $f\text{-CMI}(f, \tilde{z}, u)$ *is defined as*

$$f\text{-CMI}(f, \tilde{z}, u) = I(f(\tilde{z}_S, \tilde{x}_u, R); S_u), \tag{16}$$

while functional conditional mutual information $f\text{-CMI}_{\mathcal{D}}(f, u)$ *is defined as*

$$f\text{-CMI}_{\mathcal{D}}(f, u) = \mathbb{E}_{\tilde{z}\sim\tilde{Z}}\, f\text{-CMI}(f, \tilde{z}, u). \tag{17}$$

When $u = [n]$ we will simply use the notations $f\text{-CMI}(f, \tilde{z})$ and $f\text{-CMI}_{\mathcal{D}}(f)$, instead of $f\text{-CMI}(f, \tilde{z}, [n])$ and $f\text{-CMI}_{\mathcal{D}}(f, [n])$, respectively.

Theorem 3.1. *Let U be a random subset of size m, independent of $\tilde{Z}$, S, and randomness of training algorithm f. If $\ell(\widehat{y}, y) \in [0, 1], \forall \widehat{y} \in \mathcal{K}, z \in \mathcal{Z}$, then*

$$\left|\mathbb{E}_{\tilde{Z},R,S}\left[\mathcal{L}(f, \tilde{Z}_S, R) - \mathcal{L}_{\text{emp}}(f, \tilde{Z}_S, R)\right]\right| \le \mathbb{E}_{\tilde{z}\sim\tilde{Z},u\sim U}\sqrt{\frac{2}{m} f\text{-CMI}(f, \tilde{z}, u)}, \tag{18}$$

and

$$\mathbb{E}_{\tilde{Z},R,S}\left(\mathcal{L}(f, \tilde{Z}_S, R) - \mathcal{L}_{\text{emp}}(f, \tilde{Z}_S, R)\right)^2 \le \frac{8}{n}\left(\mathbb{E}_{\tilde{z}\sim\tilde{Z}}\, f\text{-CMI}(f, \tilde{z}) + 2\right). \tag{19}$$

For parametric methods, the bound of (18) improves over the bound of (13), as the Markov chain S_u — $A(\tilde{z}_S, R)$ — $f(\tilde{z}_S, \tilde{x}_u, R)$ allows to use the data processing inequality $I(f(\tilde{z}_S, \tilde{x}_u, R); S_u) \le I(A(\tilde{z}_S, R); S_u)$. For deterministic algorithms $I(A(\tilde{z}_S); S_u)$ is often equal to $H(S_u) = m \log 2$, as most likely each choice of S produces a different $W = A(\tilde{z}_S)$. In such cases the bound with $I(W; S_u)$ is vacuous. In contrast, the proposed bounds with f-CMI (especially when $m = 1$) do not have this problem. Even when the algorithm is stochastic, information between W and S_u can be much larger than information between predictions and S_u, as having access to weights makes it easier to determine S_u (e.g., by using gradients). A similar phenomenon has been observed in the context of membership attacks, where having access to weights of a neural network allows constructing more successful membership attacks compared to having access to predictions only [18, 9].

Corollary 1. *When $m = n$, the bound of (18) becomes*

$$\left|\mathbb{E}_{\tilde{Z},R,S}\left[\mathcal{L}(f, \tilde{Z}_S, R) - \mathcal{L}_{\text{emp}}(f, \tilde{Z}_S, R)\right]\right| \le \mathbb{E}_{\tilde{z}\sim\tilde{Z}}\sqrt{\frac{2}{n} f\text{-CMI}(f, \tilde{z})} \le \sqrt{\frac{2}{n} f\text{-CMI}_{\mathcal{D}}(f)}. \tag{20}$$

For parametric models, this improves over the CMI bound (Thm. 2.4), as by data processing inequality, $f\text{-CMI}_{\mathcal{D}}(f) = I(f(\tilde{Z}_S, \tilde{X}, R); S \mid \tilde{Z}) \le I(A(\tilde{Z}_S, R); S \mid \tilde{Z}) = \text{CMI}_{\mathcal{D}}(A)$.

Remark 1. Note that the collection of training and testing predictions $f(\tilde{Z}_S, \tilde{X}, R)$ cannot be replaced with only testing predictions $f(\tilde{Z}_S, \tilde{X}_{\text{neg}(S)}, R)$. As an example, consider an algorithm that memorizes the training examples and outputs a constant prediction on any other example. This algorithm will have non-zero generalization gap, but $f(\tilde{Z}_S, \tilde{X}_{\text{neg}(S)}, R)$ will be constant and will have zero information with S conditioned on any random variable. Moreover, if we replace $f(\tilde{Z}_S, \tilde{X}, R)$ with only training predictions $f(\tilde{Z}_S, \tilde{X}_S, R)$, the resulting bound can become too loose, as one can deduce S by comparing training set predictions with the labels $\tilde{Y}$.

Corollary 2. *When $m = 1$, the bound of (18) becomes*

$$\left|\mathbb{E}_{\tilde{Z},R,S}\left[\mathcal{L}(f, \tilde{Z}_S, R) - \mathcal{L}_{\text{emp}}(f, \tilde{Z}_S, R)\right]\right| \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\tilde{z}\sim\tilde{Z}} \sqrt{2I(f(\tilde{z}_S, \tilde{x}_i, R); S_i)}. \tag{21}$$

A great advantage of this bound compared to all other bounds described so far is that the mutual information term is computed between a relatively low-dimensional random variable $f(\tilde{z}_S, \tilde{x}_i, R)$ and a binary random variable S_i. For example, in the case of binary classification with $\mathcal{K} = \{0, 1\}$, $f(\tilde{z}_S, \tilde{x}_i, R)$ will be a pair of 2 binary variables.

As in the case of bounds presented in the previous section (Thm. 2.2 and Thm. 2.6), we prove that the bound of Thm. 3.1 is non-decreasing in m. This stays true even when we increase the upper bounds by moving the expectation over U or the expectation over $\tilde{Z}$ or both under the square root. The following proposition allows us to prove all these statements.

Proposition 3. *Let $m \in [n-1]$, U be a random subset of $[n]$ of size m, U' be a random subset of size $m+1$, $\tilde{z}$ be any fixed value of $\tilde{Z}$, and $\phi : \mathbb{R} \to \mathbb{R}$ be any non-decreasing concave function. Then*

$$\mathbb{E}_{u\sim U}\, \phi\left(\frac{1}{m} I(f(\tilde{z}_S, \tilde{x}_u, R); S_u)\right) \leq \mathbb{E}_{u'\sim U'}\, \phi\left(\frac{1}{m+1} I(f(\tilde{z}_S, \tilde{x}_{u'}, R); S_{u'})\right). \tag{22}$$

By setting $\phi(x) = \sqrt{x}$ and then taking expectation over $\tilde{z}$ and u, we prove that bounds of Thm. 3.1 are non-decreasing over m. By setting $\phi(x) = x$, taking expectation over $\tilde{z}$, and then taking square root of both sides of (22), we prove that bounds are non-decreasing in m when both expectations are under the square root. Proposition 3 proves that $m = 1$ is the optimal choice in Thm. 3.1. Notably, the bound that is the easiest to compute is also the tightest!

Analogously to Thm. A.1, we provide the following stability-based bounds.

Theorem 3.2. *If $\ell(\widehat{y}, y) \in [0, 1], \forall \widehat{y} \in \mathcal{K}, z \in \mathcal{Z}$, then*

$$\left|\mathbb{E}_{\tilde{Z},R,S}\left[\mathcal{L}(f, \tilde{Z}_S, R) - \mathcal{L}_{\text{emp}}(f, \tilde{Z}_S, R)\right]\right| \leq \mathbb{E}_{\tilde{z}\sim\tilde{Z}}\left[\frac{1}{n}\sum_{i=1}^{n}\sqrt{2I(f(\tilde{z}_S, \tilde{x}_i, R); S_i \mid S_{-i})}\right], \tag{23}$$

and

$$\mathbb{E}_{\tilde{Z},R,S}\left(\mathcal{L}(f, \tilde{Z}_S, R) - \mathcal{L}_{\text{emp}}(f, \tilde{Z}_S, R)\right)^2 \leq \frac{8}{n}\left(\mathbb{E}_{\tilde{z}\sim\tilde{Z}}\left[\sum_{i=1}^{n} I(f(\tilde{z}_S, \tilde{x}, R); S_i \mid S_{-i})\right] + 2\right).$$

Note that unlike (23), in Thm. 3.2 we measure information with predictions on all $2n$ pairs and S_i conditioned on S_{-i}. It is an open question whether $f(\tilde{z}_S, \tilde{x}, R)$ can be replaced with $f(\tilde{z}_S, \tilde{x}_i, R)$ – predictions only on the i-th pair.

4 Applications

In this section we describe 3 applications of the f-CMI-based generalization bounds.

4.1 Ensembling algorithms

Ensembling algorithms combine predictions of multiple learning algorithms to obtain better performance. Let us consider k learning algorithms, $f_1, f_2, \ldots, f_k$, each with its own independent randomness R_i, $i \in [k]$. Some ensembling algorithm can be seen as a possibly stochastic function

$g : \mathcal{K}^k \to \mathcal{K}$, that takes predictions of the k algorithms and combines them into a single prediction. Relating the generalization gap of the resulting ensembling algorithm to that of individual f_is can be challenging for complicated choices of g. However, it is easy to bound the generalization gap of $g(f_1, \ldots, f_k)$ in terms of f-CMIs of individual predictors. Let $\tilde{z}$ be a fixed value of $\tilde{Z}$ and x be an arbitrary collection of inputs. Denoting $F_i = f_i(\tilde{z}_S, x, R_i),\ i \in [k]$, we have that

$$\begin{aligned}
I(g(F_1, \ldots, F_k); S) &\leq I(F_1, \ldots, F_k; S) && \text{(data processing inequality)}\\
&= I(F_1; S) + I(F_2, \ldots, F_k; S) - I(F_1; F_2, \ldots, F_k) + I(F_1; F_2, \ldots, F_k \mid S) && \text{(chain rule)}\\
&\leq I(F_1; S) + I(F_2, \ldots, F_k; S) && \text{(as MI is nonnegative and } F_1 \perp\!\!\!\perp F_2, \ldots, F_k \mid S)\\
&\leq \ldots \leq I(F_1; S) + \cdots + I(F_k; S). && \text{(repeating the arguments above to separate all } F_i)
\end{aligned}$$

Unfortunately, the same derivation above does not work if we replace S with S_u, where u is a proper subset of $[n]$, as $I(F_1; F_2, \ldots, F_k \mid S_u)$ will not be zero in general.

4.2 Binary classification with finite VC dimension

Let us consider the case of binary classification: $\mathcal{Y} = \{0, 1\}$, where the learning method $f : \mathcal{Z}^n \times \mathcal{X} \times \mathcal{R} \to \{0, 1\}$ is implemented using a learning algorithm $A : \mathcal{Z}^n \times \mathcal{R} \to \mathcal{W}$ that selects a classifier from a hypothesis set $\mathcal{H} = \{h_w : \mathcal{X} \to \mathcal{Y}\}$. If $\mathcal{H}$ has finite VC dimension d [29], then for any algorithm f, the quantity $f\text{-CMI}(f, \tilde{z})$ can be bounded the following way.

Theorem 4.1. *Let $\mathcal{Z}$, $\mathcal{H}$, f be defined as above, and let $d < \infty$ be the VC dimension of $\mathcal{H}$. Then for any algorithm f and $\tilde{z} \in \mathcal{Z}^{n \times 2}$,*

$$f\text{-CMI}(f, \tilde{z}) \leq \max\left\{d \log(2d + 2), d \log\left(2en/d\right)\right\}. \tag{24}$$

This results shows that as $n \to \infty$ the expected generalization gap (for the 0-1 loss) will converge to zero for any learning algorithm A. Asymptotically the bound is $O\left(\sqrt{\log n}/\sqrt{n}\right)$, which is slightly weaker (by an extra $\sqrt{\log n}$ factor) than the classical uniform convergence bound [29]. This additional $\sqrt{\log n}$ factor can be removed in some cases [10].

Both Xu and Raginsky [34] and Steinke and Zakynthinou [28] prove similar information-theoretic bounds in the case of finite VC dimension classes, but their results holds for specific algorithms only. Even in the simple case of threshold functions: $\mathcal{X} = [0, 1]$ and $\mathcal{H} = \left\{h_w : x \mapsto \mathbf{1}_{\{x > w\}} \mid w \in [0, 1]\right\}$, all weight-based bounds described in Sec. 2 are vacuous if one uses a training algorithm that encodes the training set in insignificant bits of W, while still getting zero error on the training set and hence achieving low test error.

4.3 Stable deterministic or stochastic algorithms

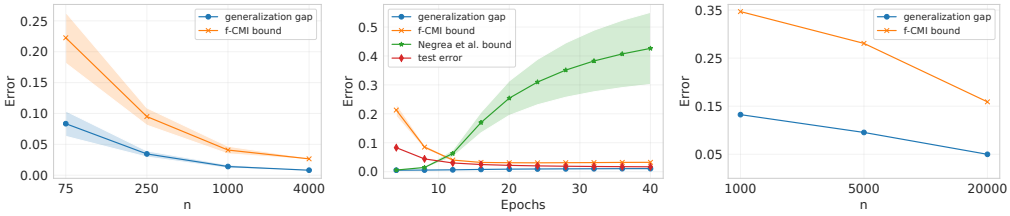
Theorems 2.3, A.1 and 3.2 provide generalization bounds involving information-theoretic stability measures, such as $I(W; Z_i \mid Z_{-i})$, $I(A(\tilde{z}_S, R); S \mid S_{-i})$ and $I(f(\tilde{z}_S, \tilde{x}, R); S_i \mid S_{-i})$. In this section we build upon the predication-based stability bounds of Thm. 3.2. First, we show that for any collection of examples x, the mutual information $I(f(\tilde{z}_S, x); S_i \mid S_{-i})$ can be bounded as follows.

Proposition 4. *Let $S^{i \leftarrow c}$ denote S with S_i set to c. Then for any $\tilde{z} \in \mathcal{Z}^{n \times 2}$ and $\tilde{x} \in \mathcal{X}^k$, the mutual information $I(f(\tilde{z}_S, x, R); S_i \mid S_{-i})$ is upper bounded by*

$$\frac{1}{4} KL\left(f(\tilde{z}_{S^{i \leftarrow 1}}, x, R) | S_{-i} \parallel f(\tilde{z}_{S^{i \leftarrow 0}}, x, R) | S_{-i}\right) + \frac{1}{4} KL\left(f(\tilde{z}_{S^{i \leftarrow 0}}, x, R) | S_{-i} \parallel f(\tilde{z}_{S^{i \leftarrow 1}}, x, R) | S_{-i}\right).$$

To compute the right-hand side of Proposition 4 one needs to know how much on-average the distribution of predictions on x changes after replacing the i-th example in the training dataset. The problem arises when we consider deterministic algorithms. In such cases, the right-hand side is infinite, while the left-hand side $I(f(\tilde{z}_S, x, R); S_i \mid S_{-i})$ is always finite and could be small. Therefore, for deterministic algorithms, directly applying the result of Proposition 4 will not give meaningful generalization bounds. Nevertheless, we show that we can add an optimal amount of noise to predictions, upper bound the generalization gap of the resulting noisy algorithm, and relate that to the generalization gap of the original deterministic algorithm.

Let us consider a deterministic algorithm $f : \mathcal{Z}^n \times \mathcal{X} \to \mathbb{R}^d$. For simplicity we assume that the algorithm is invariant to the order of training examples (called symmetric algorithms hereafter). Formally, for any $z = (z_1, \ldots, z_n) \in \mathcal{Z}^n$, $x \in \mathcal{X}$ and permutation π, $f((z_1, \ldots, z_n), x) = f((z_{\pi(1)}, \ldots, z_{\pi(n)}), x)$. We define the following notions of functional stability.



(a) (b) (c)

Figure 1: Comparison of expected generalization gap and f-CMI bound in 3 settings: (a) MNIST 4 vs 9 classification with a CNN trained using a deterministic algorithm, (b) MNIST 4 vs 9 with a CNN trained using SGLD, and (c) pretrained ResNet-50 finetuned on CIFAR-10.

Definition 4.1 (Functional stability)**.** *Let $S = (Z_1, \dots, Z_n) \sim \mathcal{D}^n$ be a collection of n i.i.d. samples and $Z_1' \sim \mathcal{D}$, $Z' \sim \mathcal{D}$ be two additional independent samples. Let $S' \triangleq (Z_1', Z_2, \dots, Z_n)$ be the collection constructed from S by replacing the first example with Z_1'. A deterministic symmetric algorithm $f : \mathcal{Z}^n \times \mathcal{X} \to \mathbb{R}^d$ is*

$$\text{a) } \beta \text{ self-stable if } \mathbb{E}_{S,Z_1',Z'} \left\| f(S, Z_1) - f(S', Z_1) \right\|^2 \leq \beta^2, \tag{25}$$

$$\text{b) } \beta_1 \text{ test-stable if } \mathbb{E}_{S,Z_1',Z'} \left\| f(S, Z') - f(S', Z') \right\|^2 \leq \beta_1^2, \tag{26}$$

$$\text{c) } \beta_2 \text{ train-stable if } \mathbb{E}_{S,Z_1',Z'} \left\| f(S, Z_2) - f(S', Z_2) \right\|^2 \leq \beta_2^2. \tag{27}$$

Theorem 4.2. *Let $\mathcal{Y} = \mathbb{R}^d$, $f : \mathcal{Z}^n \times \mathcal{X} \to \mathbb{R}^d$ be a deterministic symmetric algorithm that is β self-stable, and $\ell(\widehat{y}, y) \in [0, 1]$ be a loss function that is γ-Lipschitz in the first coordinate. Then*

$$\left| \mathbb{E}_{\tilde{Z},R,S} \left[\mathcal{L}(f, \tilde{Z}_S, R) - \mathcal{L}_{\text{emp}}(f, \tilde{Z}_S, R) \right] \right| \leq 2^{\frac{3}{2}} d^{\frac{1}{4}} \sqrt{\gamma\beta}. \tag{28}$$

Furthermore, if f is also β_1 train-stable and β_2 test-stable, then

$$\mathbb{E}_{\tilde{Z},R,S} \left(\mathcal{L}(f, \tilde{Z}_S, R) - \mathcal{L}_{\text{emp}}(f, \tilde{Z}_S, R) \right)^2 \leq \frac{32}{n} + 12^{\frac{3}{2}} \sqrt{d}\gamma \sqrt{2\beta^2 + n\beta_1^2 + n\beta_2^2}. \tag{29}$$

It is expected that β_1 is smaller than β and β_2. For example, in the case of neural networks interpolating the training data or in the case of empirical risk minimization in the realizable setting, β_1 will be zero. It is also expected that β is larger than β_2. However, the relation of β^2 and $n\beta_2^2$ is not trivial.

5 Experiments

As mentioned earlier, the expected generalization gap bound of Corollary 2 is significantly easier to compute compared to existing information-theoretic bounds, and does not give trivial results for deterministic algorithms. To understand how well the bound does in challenging situations, we consider cases when the algorithm generalizes well despite the high complexity of the function class and relatively small number of training examples. Due to space constraints we omit some experimental details and present them in Appendix B.

First, we consider the MNIST 4 vs 9 digit classification task [17] using a 4-layer convolutional neural network (CNN) that has approximately 200K parameters. We train the network using for 200 epochs using the ADAM algorithm [15] with 0.001 learning rate, $\beta_1 = 0.9$, and mini-batches of 128 examples. Importantly, we fix the random seed that controls the initialization of weights and the shuffling of training data, making the training algorithm deterministic. Fig. 1a plots the expected generalization gap and the f-CMI bound of (21). We see that the bound closely tracks the expected generalization gap and is not too far from it even when considering only 75 training examples. As shown in the Fig. 2a of Appendix B, if we increase the width of all layers 4 times, making the number of parameters approximately 3M, the results remain largely unchanged.

Note that the existing weight-based information-theoretic bounds would give either infinite or trivial bounds for the deterministic algorithm described above. Even when we make the training

algorithm stochastic by randomizing the seed, the quantities like $I(W;S)$ still remain infinite, while both the generalization gap and the f-CMI bound do not change significantly (see Fig. 2b of Appendix B). For this reason, we change the training algorithm to Stochastic Gradient Langevin Dynamics (SGLD) [7, 33] and compare the f-CMI-based bound against the specialized bound of Negrea et al. [19] (see eq. (6) of [19]). This bound (referred as SGLD bound here) is derived from a weight-based information-theoretic generalization bound, and depends on the the hyper-parameters of SGLD and on the variance of per-example gradients along the training trajectory. The SGLD algorithm is trained for 40 epochs, with learning rate and inverse temperature schedules described in Appendix B. Fig. 1b plots the expected generalization gap, the expected test error, the f-CMI bound and the SGLD bound. We see that the test accuracy plateaus after 16 epochs. At this time and afterwards, the f-CMI bound closely follows the generalization gap, while the SGLD bound increases to very high values. However, we see that the SGLD bound does better when the number of epochs is less than 12. Surprisingly, the f-CMI bound is large in the early epochs, despite of the extremely small generalization error. In fact, a similar trend, albeit with lesser extent, is visible in the experiments with the deterministic algorithm described above (see Fig. 2c of Appendix B). We hypothesize that predictions on small number of hard examples are revealing significant information about S: A correct prediction on a hard example after training for a few epochs, increases the probability of that example being in the training set.

Finally, we move away from binary classification and consider the CIFAR-10 classication task [16]. To construct a well-generalizing algorithm, we use the ResNet-50 [13] network pretrained on the ImageNet [5], and fine-tune it for 40 epochs using SGD with mini-batches of size 64, 0.01 learning rate, 0.9 momentum, and standard data augmentations. The results presented in Fig. 1c indicate that the f-CMI bound is always approximately 3 times larger than the expected generalization gap. In particular, when $n = 20000$, the expected generalization gap is 5%, while the bound predicts 16%.

6 Related Work

This work is closely related to a rich literature of information-theoretic generalization bounds, some of which were discussed earlier [34, 2, 21, 19, 4, 28, 11, 10, 20, 23]. Most of these work derive generalization bounds that depend on a mutual information quantity measured between the output of the training algorithm and some quantity related to training data. Different from this major idea, Xu and Raginsky [34] and Russo and Zou [24] discussed the idea of bounding generalization gap with the information between the input and the vector of loss functions computed on training examples. This idea was later extended to the setting of conditional mutual information by Steinke and Zakynthinou [28]. This works are similar to ours in the sense that they move away from measuring information with weights, but they did not develop this line of reasoning enough to arrive to efficient bounds similar to Corollary 2. Additionally, we believe that measuring information with the prediction function allows better interpretation and is easier to work with analytically.

Another related line of research are the stability-based bounds [3, 22, 1, 6, 32, 23]. In Sec. 2 and Sec. 3 we improve existing generalization bounds that use information stability. In Sec. 4.3 we describe a technique of applying information stability bounds to deterministic algorithms. The main idea is to add noise to predictions, but only for analysis purposes. A similar idea, but in the context of measuring information content of an individual example, was suggested by Harutyunyan et al. [12]. In fact, our notion of test-stability defined in Sec. 4.3 comes very close to their definition of functional sample information. A similar idea was recently used by Neu [20] in analyzing generalization performance of SGD. More broadly this work is related to PAC-Bayes bounds and to classical generalization bounds. Please refer to the the survey by Jiang* et al. [14] for more information on these bounds.

Finally, our work has connections with attribute and membership inference attacks [27, 35, 18, 9]. Some of these works show that having a white-box access to models allows constructing better membership inference attacks, compared to having a black-box access. This is analogous to our observation that prediction-based bounds are better than weight-based bounds. Shokri et al. [27] and Yeom et al. [35] demonstrate that even in the case of *black-box access to a well-generalizing* model, sometimes it is still possible to construct successful membership attacks. This is in line with our observation that the f-CMI bound can be significantly large, despite of small generalization gap (see epoch 4 of Fig. 1b). This suggests a possible direction of improving the f-CMI-based bounds.

References

[1] R. Bassily, K. Nissim, A. Smith, T. Steinke, U. Stemmer, and J. Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1046–1059, 2016.

[2] R. Bassily, S. Moran, I. Nachum, J. Shafer, and A. Yehudayoff. Learners that use little information. In *Algorithmic Learning Theory*, pages 25–55. PMLR, 2018.

[3] O. Bousquet and A. Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

[4] Y. Bu, S. Zou, and V. V. Veeravalli. Tightening mutual information-based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 1(1):121–130, 2020.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] V. Feldman and T. Steinke. Calibrating noise to variance in adaptive data analysis. In *Conference On Learning Theory*, pages 535–544, 2018.

[7] S. B. Gelfand and S. K. Mitter. Recursive stochastic algorithms for global optimization in r^d. *SIAM Journal on Control and Optimization*, 29(5):999–1018, 1991.

[8] R. M. Gray. *Entropy and information theory*. Springer Science & Business Media, 2011.

[9] U. Gupta, D. Stripelis, P. K. Lam, P. Thompson, J. L. Ambite, and G. V. Steeg. Membership inference attacks on deep regression models for neuroimaging. In *Medical Imaging with Deep Learning*, 2021.

[10] H. Hafez-Kolahi, Z. Golgooni, S. Kasaei, and M. Soleymani. Conditioning and processing: Techniques to improve information-theoretic generalization bounds. *Advances in Neural Information Processing Systems*, 33, 2020.

[11] M. Haghifam, J. Negrea, A. Khisti, D. M. Roy, and G. K. Dziugaite. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9925–9935. Curran Associates, Inc., 2020.

[12] H. Harutyunyan, A. Achille, G. Paolini, O. Majumder, A. Ravichandran, R. Bhotika, and S. Soatto. Estimating informativeness of samples with smooth unique information. In *International Conference on Learning Representations*, 2021.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Y. Jiang*, B. Neyshabur*, H. Mobahi, D. Krishnan, and S. Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJgIPJBFvH.

[15] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[17] Y. LeCun, C. Cortes, and C. Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[18] M. Nasr, R. Shokri, and A. Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 739–753. IEEE, 2019.

[19] J. Negrea, M. Haghifam, G. K. Dziugaite, A. Khisti, and D. M. Roy. Information-theoretic generalization bounds for sgld via data-dependent estimates. In *Advances in Neural Information Processing Systems*, pages 11015–11025, 2019.

[20] G. Neu. Information-theoretic generalization bounds for stochastic gradient descent. *arXiv preprint arXiv:2102.00931*, 2021.

[21] A. Pensia, V. Jog, and P.-L. Loh. Generalization error bounds for noisy, iterative algorithms. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 546–550. IEEE, 2018.

[22] M. Raginsky, A. Rakhlin, M. Tsao, Y. Wu, and A. Xu. Information-theoretic analysis of stability and bias of learning algorithms. In *2016 IEEE Information Theory Workshop (ITW)*, pages 26–30. IEEE, 2016.

[23] M. Raginsky, A. Rakhlin, and A. Xu. 10 information-theoretic stability and generalization. *Information-Theoretic Methods in Data Science*, page 302, 2021.

[24] D. Russo and J. Zou. How much does your data exploration overfit? controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1):302–323, 2019.

[25] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1): 145–147, 1972.

[26] S. Shelah. A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41(1):247–261, 1972.

[27] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

[28] T. Steinke and L. Zakynthinou. Reasoning About Generalization via Conditional Mutual Information. In J. Abernethy and S. Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3437–3452. PMLR, 09–12 Jul 2020.

[29] V. Vapnik. Statistical learning theory new york. *NY: Wiley*, 1998.

[30] S. Verdu and T. Weissman. The information lost in erasures. *IEEE Transactions on Information Theory*, 54(11):5030–5058, 2008.

[31] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[32] Y.-X. Wang, J. Lei, and S. 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.

[33] M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

[34] A. Xu and M. Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2524–2533, 2017.

[35] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282. IEEE, 2018.

[36] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] In Sec. 3 we discuss some limitations of the proposed squared generalization bounds. In Sec. 5 and Sec. 6 we discuss cases where the proposed bounds are loose.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We do not see any potential societal impacts our work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Sections 2, 3, and A.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix A.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See the supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Sec. 5 and Appendix B.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Sec. 5 and Appendix B.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix B.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] The supplementary material includes the code for replicating experiments of this paper.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]