This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAIT.2020.2991139, IEEE Journal on Selected Areas in Information Theory

Tightening Mutual Information Based Bounds on Generalization Error

Yuheng Bu, *Member, IEEE*, Shaofeng Zou, *Member, IEEE*, and Venugopal V. Veeravalli *Fellow, IEEE*

***Abstract*—An information-theoretic upper bound on the generalization error of supervised learning algorithms is derived. The bound is constructed in terms of the mutual information between each individual training sample and the output of the learning algorithm. The bound is derived under more general conditions on the loss function than in existing studies; nevertheless, it provides a tighter characterization of the generalization error. Examples of learning algorithms are provided to demonstrate the the tightness of the bound, and to show that it has a broad range of applicability. Application to noisy and iterative algorithms, e.g., stochastic gradient Langevin dynamics (SGLD), is also studied, where the constructed bound provides a tighter characterization of the generalization error than existing results. Finally, it is demonstrated that, unlike existing bounds, which are difficult to compute and evaluate empirically, the proposed bound can be estimated easily in practice.**

***Index Terms*—Cumulant generating function, generalization error, information-theoretic bounds, stochastic gradient Langevin dynamics**

I. Introduction

Recent success of deep learning algorithms [2] has dramatically boosted their applications in various engineering and science domains, e.g., computer vision [3], natural language processing [4], autonomous driving [5], and health care [6]. A deep neural network trained using a sufficiently large amount of training data can achieve a small training error, while simultaneously performing well on unseen data, i.e., it generalizes well. However, we have yet to develop a satisfactory understanding of why deep learning algorithms generalize well.

Classical statistical learning approaches for analyzing the generalization capability of supervised learning algorithms can be mainly categorized into two groups. The first set of methods are based on measures of the complexity of the output hypothesis space, e.g., VC dimension and Rademacher complexity [7], [8]. However, these complexity measures usually scale exponentially with the depth of deep neural networks [9]. Moreover, these approaches do not take into consideration the regularization implicitly imposed by the algorithms used to train the neural networks, e.g., stochastic gradient descent [10], [11]. Thus, the generalization error bounds based on these complexity measures tend to be loose and do not explain why deep neural networks generalize well in practice. The second set of methods are based on exploiting properties of the learning algorithm, e.g., PAC-Bayesian bounds [12], uniform stability [13], [14], and compression bounds [15]. However, as discussed in [11], [16], these approaches do not exploit the fact that the generalization error depends strongly on the underlying true data-generating distribution. For example, if the labels are irrelevant to the input features, then the generalization error will be large for a deep neural network, since training error is usually small due to the large capacity of the network, but test error will be large due to the fact that there is no relationship between the input features and the label [11].

Recently, it was proposed in [17] and further studied in [18] and [19] that the metric of mutual information can be used to develop upper bounds on the generalization error of learning algorithms. Such an information-theoretic framework can handle a broader range of problems, and it could also address the aforementioned challenges of implicit regularization and dependence on data generating distribution. More importantly, it offers an information-theoretic point of view on how to improve the generalization capability of a learning algorithm, and this new perspective provides us with a better understanding of the generalization behavior of deep neural networks.

In this paper, we follow the information-theoretic framework proposed in [17]–[19]. Our main contribution is a tighter upper bound on the generalization error using the mutual information between an individual training sample and the output hypothesis of the learning algorithm. We show that compared to existing studies, our bound has a broader range of applicability, and can be considerably tighter.

We consider an instance space $\mathcal{Z}$, a hypothesis space $\mathcal{W}$, and a nonnegative loss function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}^+$. A training dataset $S = \{Z_1, \cdots, Z_n\}$ that consists of n i.i.d samples $Z_i \in \mathcal{Z}$ drawn from an unknown distribution μ is available. The goal of a supervised learning algorithm is to find an output hypothesis $w \in \mathcal{W}$ that minimizes the *population risk*:

$$L_\mu(w) \triangleq \mathbb{E}_{Z\sim\mu}[\ell(w, Z)]. \tag{1}$$

In practice, μ is unknown, and thus $L_\mu(w)$ cannot be computed directly. Instead, the *empirical risk* of w on a training dataset

This work was presented in part at the IEEE International Symposium on Information Theory (ISIT), Paris, France, 2019 [1].

This work was supported by Army Research Laboratory under Cooperative Agreement W911NF-17-2-0196, through the University of Illinois at Urbana-Champaign.

Y. Bu was with the ECE Department and the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA. He is now with the Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (email: buyuheng@mit.edu).

S. Zou is with the Department of Electrical Engineering, University at Buffalo, The State University of New York, Buffalo, NY 14228 USA (email: szou3@buffalo.edu).

V. V. Veeravalli is with the ECE Department and the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (email: vvv@illinois.edu).

S is studied, which is defined as

$$L_S(w) \triangleq \frac{1}{n}\sum_{i=1}^{n}\ell(w, Z_i). \tag{2}$$

A learning algorithm can be characterized by a randomized mapping from the training dataset S to a hypothesis W according to a conditional distribution $P_{W|S}$.

In statistical learning theory, the *(mean) generalization error*[1] of a supervised learning algorithm is the expected difference between the population risk of the output hypothesis and its empirical risk on the training dataset:

$$\mathrm{gen}(\mu, P_{W|S}) \triangleq \mathbb{E}_{W,S}[L_\mu(W) - L_S(W)], \tag{3}$$

where the expectation is taken over the joint distribution $P_{S,W} = P_S \otimes P_{W|S}$. Note that $P_{W|S}$ will become degenerate if W is a deterministic function of S. The generalization error is used to measure the extent to which the learning algorithm overfits the training data.

Main Contributions and Related Works

We first review the following lemma from [18], which provides an upper bound on the generalization error using the mutual information $I(S;W)$ between the training dataset S and the output hypothesis W.

Lemma 1 ([18, Theorem 1])**.** *Suppose $\ell(w, Z)$ is R-sub-Gaussian*[2] *under $Z \sim \mu$ for all $w \in \mathcal{W}$, then*

$$|\mathrm{gen}(\mu, P_{W|S})| \leq \sqrt{\frac{2R^2}{n} I(S;W)}. \tag{4}$$

This mutual information based bound in (4) is related to the "on-average" stability (see, e.g., [20]), since it quantifies the overall dependence between the output of the learning algorithm and all the input training samples via $I(S;W)$. Note that $I(S;W)$ depends on the main components of a supervised learning problem, i.e., the hypothesis space $\mathcal{W}$, the learning algorithm $P_{W|S}$, and the data generating distribution μ, in contrast to the traditional bounds based on VC dimension or the uniform stability, which only depend on one aspect of the learning problem. We also note that there is a connection between the mutual information based generalization bound and the PAC-Bayesian bound in [21], since both methods adopt the variational representation of relative entropy to establish the decoupling lemma. By further exploiting the structure of the hypothesis space and the dependency between the algorithm input and output, the authors of [19], [22] combined the chaining and mutual information methods, and obtained a tighter bound on the generalization error.

However, the bound in Lemma 1 and the chaining mutual information (CMI) bound in [19] both suffer from the following two shortcomings. First, for empirical risk minimization (ERM), if W is the unique minimizer of $L_S(w)$ in $\mathcal{W}$, then W is a deterministic function of S and the mutual information $I(S;W) = \infty$. It can be shown that both bounds are not tight in this case. Second, both bounds assume that $\ell(w, Z)$ has a bounded cumulant generating function (CGF) under $Z \sim \mu$ *for all* $w \in \mathcal{W}$, which may not hold in many cases (see Section IV).

There has been some recent work on addressing these shortcomings of mutual information based bounds on generalization error by using other information-theoretical measures, e.g., Wasserstein distance [23]–[25], maximal leakage [26], [27] and total variation [28] to bound the generalization error. But the measures proposed in these papers are difficult to evaluate both analytically and empirically as we discuss in Section VI, which significantly undermines the usefulness of these results in practice.

In this paper, we get around the aforementioned shortcomings by combining the idea of point-wise stability [14], [23] with the information-theoretic framework introduced in [18]. Specifically, an algorithm is said to be point-wise stable if the expectation of the loss function $\ell(W, Z_i)$ does not change too much with the replacement of any *individual* training sample Z_i, and if an algorithm is point-wise stable, then it generalizes well [14], [23]. Motivated by these facts, we tighten the mutual information based generalization error bound through a bound based on the individual sample mutual information (ISMI) $I(W;Z_i)$. Compared with the bound in Lemma 1, and the CMI bound in [19], the ISMI bound is derived under a more general condition on the CGF of the loss function, is applicable to a broader range of problems, and can provide a tighter characterization of the generalization error.

The rest of the paper is organized as follows. In Section II, we provide some preliminary definitions and results for our analysis. In Section III, we introduce the individual sample mutual information generalization bound. In Section IV, we apply our method to bound the generalization errors of two learning problems with infinite $I(S;W)$. We show in the second example that our ISMI bound can be tighter than the CMI bound in [19], while the bound in Lemma 1 is infinity. In Section V, we improve the generalization error bound in [29] for SGLD algorithm using our method, which demonstrates that the ISMI bound is applicable to the noisy, iterative algorithms discussed in [29]. In Section VI, we provide an example where the ISMI bound can be evaluated empirically from the samples, while other existing bounds are difficult to estimate due to prohibitive computational complexity.

II. Preliminaries

We use upper letters to denote random variables, and calligraphic upper letters to denote sets. For a random variable X generated from a distribution μ, we use $\mathbb{E}_{X\sim\mu}$ to denote the expectation taken over X with distribution μ. We write I_d to denote the d-dimensional identity matrix. All the logarithms are the natural ones, and all the information measure units

[1]The term "generalization error" of a learning algorithm is usually defined as the difference between the population risk and the training error *without* taking the expectation with respect to the randomness of the data and the learning algorithm. Here, we consider the expectation of the generalization error over the randomness of both the data and the learning algorithm. We will use the term "generalization error" throughout the paper, with the understanding that it is the "mean generalization error".

[2]A random variable X is R-sub-Gaussian if $\log \mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] \leq \frac{R^2\lambda^2}{2}$, $\forall \lambda \in \mathbb{R}$.

are nats. We use $\mu^{\otimes n}$ to denote the product distribution of n copies of μ.

Definition 1. *The cumulant generating function (CGF) of a random variable X is defined as*

$$\Lambda_X(\lambda) \triangleq \log \mathbb{E}[e^{\lambda(X-\mathbb{E}X)}]. \tag{5}$$

Assuming $\Lambda_X(\lambda)$ exists, it can be verified that $\Lambda_X(0) = \Lambda'_X(0) = 0$, and that it is convex.

Definition 2. *For a convex function ψ defined on the interval $[0, b)$, where $0 < b \leq \infty$, its Legendre dual ψ^* is defined as*

$$\psi^*(x) \triangleq \sup_{\lambda\in[0,b)} \big(\lambda x - \psi(\lambda)\big). \tag{6}$$

The following lemma characterizes a useful property of the Legendre dual and its inverse function.

Lemma 2 ([30, Lemma 2.4])**.** *Assume that $\psi(0) = \psi'(0) = 0$. Then $\psi^*(x)$ defined above is a non-negative convex and non-decreasing function on $[0, \infty)$ with $\psi^*(0) = 0$. Moreover, its inverse function $\psi^{*-1}(y) = \inf\{x \geq 0 : \psi^*(x) \geq y\}$ is concave, and can be written as*

$$\psi^{*-1}(y) = \inf_{\lambda\in(0,b)} \Big(\frac{y + \psi(\lambda)}{\lambda}\Big). \tag{7}$$

For an R-sub-Gaussian random variable X, $\psi(\lambda) = \frac{R^2\lambda^2}{2}$ is an upper bound on $\Lambda_X(\lambda)$. Then by Lemma 2, $\psi^{*-1}(y) = \sqrt{2R^2 y}$.

III. Bounding Generalization Error via $I(W; Z_i)$

In this section, we first generalize the decoupling lemma in [18, Lemma 1] to a different setting, and then tighten the bound on generalization error via the individual sample mutual information $I(W; Z_i)$.

A. General Decoupling Estimate

Consider a pair of random variables W and Z with joint distribution $P_{W,Z}$. Let $\widetilde{W}$ be an independent copy of W, and $\widetilde{Z}$ be an independent copy of Z, such that $P_{\widetilde{W}\widetilde{Z}} = P_W \otimes P_Z$. Suppose $f : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$ is a real-valued function. If the CGF $\Lambda_{f(\widetilde{W},\widetilde{Z})}(\lambda)$ of $f(\widetilde{W}, \widetilde{Z})$ can be upper bounded by some function ψ for $\lambda \in (b_-, b_+)$, we have the following theorem.

Theorem 1. *Assume that $\Lambda_{f(\widetilde{W},\widetilde{Z})}(\lambda) \leq \psi_+(\lambda)$ for $\lambda \in [0, b_+)$, and $\Lambda_{f(\widetilde{W},\widetilde{Z})}(\lambda) \leq \psi_-(-\lambda)$ for $\lambda \in (b_-, 0]$ under distribution $P_{\widetilde{W}\widetilde{Z}} = P_W \otimes P_Z$, where $0 < b_+ \leq \infty$ and $-\infty \leq b_- < 0$. Suppose that $\psi_+(\lambda)$ and $\psi_-(\lambda)$ are convex, and $\psi_+(0) = \psi'_+(0) = \psi_-(0) = \psi'_-(0) = 0$. Then,*

$$\mathbb{E}[f(W, Z)] - \mathbb{E}[f(\widetilde{W}, \widetilde{Z})] \leq \psi_+^{*-1}\big(I(W; Z)\big), \tag{8}$$

$$\mathbb{E}[f(\widetilde{W}, \widetilde{Z})] - \mathbb{E}[f(W, Z)] \leq \psi_-^{*-1}\big(I(W; Z)\big). \tag{9}$$

Proof. Consider the variational representation of the relative entropy between two probability measures P and Q defined on $\mathcal{X}$:

$$D(P\|Q) = \sup_{g\in\mathcal{G}} \Big\{\mathbb{E}_P[g(X)] - \log \mathbb{E}_Q[e^{g(X)}]\Big\}, \tag{10}$$

where the supremum is over all measurable functions $\mathcal{G} = \{g : \mathcal{X} \to \mathbb{R}, \text{ s.t. } \mathbb{E}_Q[e^{g(X)}] < \infty\}$, and equality is achieved when $g = \log \frac{dP}{dQ}$, where $\frac{dP}{dQ}$ is the Radon–Nikodym derivative. It then follows that $\forall \lambda \in [0, b_+)$,

$$\begin{aligned} I(W; Z) &= D(P_{W,Z}\|P_W \otimes P_Z) \\ &\geq \mathbb{E}[\lambda f(W, Z)] - \log \mathbb{E}[e^{\lambda f(\widetilde{W},\widetilde{Z})}] \\ &\geq \lambda(\mathbb{E}[f(W, Z)] - \mathbb{E}[f(\widetilde{W}, \widetilde{Z})]) - \psi_+(\lambda), \end{aligned} \tag{11}$$

where the last inequality follows from the assumption that

$$\Lambda_f(\widetilde{W}; \widetilde{Z}) = \log \mathbb{E}[e^{\lambda(f(\widetilde{W},\widetilde{Z}) - \mathbb{E}f(\widetilde{W},\widetilde{Z}))}] \leq \psi_+(\lambda), \tag{12}$$

for $\lambda \in [0, b_+)$. Similarly, for $\lambda \in (b_-, 0]$, it follows that

$$\begin{aligned} &D(P_{W,Z}\|P_W \otimes P_Z) \\ &\quad\geq \lambda(\mathbb{E}[f(W, Z)] - \mathbb{E}[f(\widetilde{W}, \widetilde{Z})]) - \psi_-(-\lambda). \end{aligned} \tag{13}$$

From (11) it follows that

$$\begin{aligned} \mathbb{E}[f(W, Z)] - \mathbb{E}[f(\widetilde{W}, \widetilde{Z})] &\leq \inf_{\lambda\in[0,b_+)} \frac{I(W; Z) + \psi_+(\lambda)}{\lambda} \\ &= \psi_+^{*-1}\big(I(W; Z)\big), \end{aligned} \tag{14}$$

and from (13) it follows that

$$\begin{aligned} \mathbb{E}[f(\widetilde{W}, \widetilde{Z})] - \mathbb{E}[f(W, Z)] &\leq \inf_{\lambda\in[0,-b_-)} \frac{I(W; Z) + \psi_-(\lambda)}{\lambda} \\ &= \psi_-^{*-1}\big(I(W; Z)\big), \end{aligned} \tag{15}$$

where the equalities in (14) and (15) follow from Lemma 2. □

Theorem 1 provides a different characterization of the decoupling estimate than existing results. Specifically, it is assumed that the CGF of $f(w, Z)$ is bounded for all $w \in \mathcal{W}$ and $Z \sim \mu$ in [18, Lemma 1] and [31, Theorem 2], whereas in Theorem 1, it is assumed that the CGF of $f(\widetilde{W}, \widetilde{Z})$ is bounded in expectation under $P_{\widetilde{W}\widetilde{Z}} = P_W \otimes P_Z$.

B. Individual Sample Mutual Information Bound

Motivated by the idea of algorithmic stability, which measures how much an output hypothesis changes with the replacement of an *individual* training sample, we construct the following upper bound on the generalization error via $I(W; Z_i)$.

Theorem 2. *Suppose $\ell(\widetilde{W}, \widetilde{Z})$ satisfies $\Lambda_{\ell(\widetilde{W},\widetilde{Z})}(\lambda) \leq \psi_+(\lambda)$ for $\lambda \in [0, b_+)$, and $\Lambda_{\ell(\widetilde{W},\widetilde{Z})}(\lambda) \leq \psi_-(-\lambda)$ for $\lambda \in (b_-, 0]$ under $P_{\widetilde{Z},\widetilde{W}} = \mu \otimes P_W$, where $0 < b_+ \leq \infty$ and $-\infty \leq b_- < 0$. Then,*

$$\mathrm{gen}(\mu, P_{W|S}) \leq \frac{1}{n}\sum_{i=1}^{n} \psi_-^{*-1}\big(I(W; Z_i)\big), \tag{16}$$

$$-\mathrm{gen}(\mu, P_{W|S}) \leq \frac{1}{n}\sum_{i=1}^{n} \psi_+^{*-1}\big(I(W; Z_i)\big). \tag{17}$$

Proof. The generalization error can be written as follows:

$$\mathrm{gen}(\mu, P_{W|S}) = \frac{1}{n}\sum_{i=1}^{n} \Big(\mathbb{E}_{W,\widetilde{Z}}[\ell(W, \widetilde{Z})] - \mathbb{E}_{W,Z_i}[\ell(W, Z_i)]\Big), \tag{18}$$

where W and Z_i in the second term are dependent with $P_{W,Z_i} = \mu \otimes P_{W|Z_i}$, and W and $\widetilde{Z}$ in the first term are independent with the same marginal distributions. Applying Theorem 1 completes the proof. □

In the following proposition, we derive the ISMI bounds under two different sub-Gaussian assumptions.

Proposition 1. *1) Suppose that $\ell(w, Z)$ is R-sub-Gaussian under $Z \sim \mu$ for all $w \in \mathcal{W}$, then*

$$\left|\mathrm{gen}(\mu, P_{W|S})\right| \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2R^2 I(W; Z_i)}. \tag{19}$$

2) Suppose that $\ell(\widetilde{W}, \widetilde{Z})$ is R-sub-Gaussian under distribution $P_{\widetilde{W}\widetilde{Z}} = P_W \otimes P_Z$, then

$$\left|\mathrm{gen}(\mu, P_{W|S})\right| \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2R^2 I(W; Z_i)}. \tag{20}$$

Proof. 1) The generalization error can be written as in (18), where W and Z_i in the second term are dependent with $P_{W,Z_i} = \mu \otimes P_{W|Z_i}$, and W and $\widetilde{Z}$ in the first term are independent whose marginal distributions are the same as those of W and Z_i. The first inequality then follows from Lemma 1 by letting $S = Z_i$ and $n = 1$, for each $i = 1, \cdots, n$.

2) For an R-sub-Gaussian random variable, $\psi_+^{-1}(y) = \psi_-^{-1}(y) = \sqrt{2R^2 y}$ is an upper bound on its CGF. The second inequality then follows from Theorem 2. □

Remark 1. *The condition that $\ell(w, Z)$ is R-sub-Gaussian under $Z \sim \mu$ for all $w \in \mathcal{W}$ in the first part of Proposition 1 is the same as the one in Lemma 1, which is not stronger than the condition in the second part of Proposition 1. An example was given in [32] for this argument. Specifically, consider $\mathcal{W} = \mathcal{Z} = \mathbb{R}$, with $\ell(w, z) = w + z$, and $(\widetilde{W}, \widetilde{Z}) \sim$ Cauchy$\otimes\mathcal{N}(0, 1)$. Then, $\ell(w, Z)$ is 1-sub-Gaussian for any $w \in \mathcal{W}$, whereas $\ell(\widetilde{W}, \widetilde{Z})$ does not even have bounded absolute first moment.*

The following proposition shows that the proposed ISMI bound is always no worse than the bound using $I(S; W)$ in Lemma 1 and [31, Theorem 2].

Proposition 2. *Suppose that $S = \{Z_1, \cdots, Z_n\}$ consists of n independent samples, and ψ^{*-1} is a concave function, then*

$$\frac{1}{n}\sum_{i=1}^{n}\psi^{*-1}\Big(I(W; Z_i)\Big) \leq \psi^{*-1}\left(\frac{I(S; W)}{n}\right). \tag{21}$$

Proof. By the chain rule of mutual information,

$$I(W; S) = \sum_{i=1}^{n} I(W; Z_i|Z^{i-1}) \tag{22}$$

where $Z^j = \{Z_1, \cdots, Z_j\}$. Note that Z_i and Z^{i-1} are independent, i.e., $I(Z_i; Z^{i-1}) = 0$, it then follows that

$$\begin{aligned} I(W; Z_i|Z^{i-1}) &= I(W; Z_i|Z^{i-1}) + I(Z_i; Z^{i-1}) \\ &= I(W, Z^{i-1}; Z_i) \\ &= I(W; Z_i) + I(Z^{i-1}; Z_i|W) \\ &\geq I(W; Z_i). \end{aligned} \tag{23}$$

Thus,

$$I(W; S) = \sum_{i=1}^{n} I(W; Z_i|Z^{i-1}) \geq \sum_{i=1}^{n} I(W; Z_i), \tag{24}$$

and applying Jensen's inequality completes the proof. □

Remark 2. *Under the sub-Gaussian condition (see sentence following Lemma 1), we can let $\psi^{*-1}(y) = \sqrt{2R^2 y}$. Then by Proposition 2, the ISMI bound in Proposition 1 is always no worse than the bound based on $I(S; W)$ in Lemma 1.*

Remark 3. *Following arguments similar to those used in the proof of $I(W; Z_i) \leq I(W; Z_i|Z^{i-1})$, we can also show that $I(W; Z_i) \leq I(W; Z_i|S^{-i})$, where S^{-i} denotes the set obtained by deleting Z_i from S. Therefore, the ISMI bound is always no worse than the bound based on $I(W; Z_i|S^{-i})$ in [23, Theorem 2].*

In the next section, we will also show via several examples that the ISMI bound provides a more accurate characterization of the generalization error than the bound in Lemma 1 and the chaining bound in [19].

IV. Examples with Infinite $I(W; S)$

In this section, we consider two examples of learning algorithms with infinite $I(W; S)$. We show that for these examples, the upper bound on generalization error in Lemma 1 blows up, whereas the ISMI bound in Theorem 2 still provides an accurate approximation. The details of the derivations of the bounds can be found in the Appendices.

A. Estimating the Mean

We first consider the problem of learning the mean of a Gaussian random vector $Z \sim \mathcal{N}(\mu, \sigma^2 I_d)$, which minimizes the square error $\ell(w, Z) \triangleq \|w - Z\|_2^2$. The empirical risk with n i.i.d. samples is

$$L_S(w) \triangleq \frac{1}{n}\sum_{i=1}^{n}\|w - Z_i\|_2^2, \quad w \in \mathbb{R}^d. \tag{25}$$

The empirical risk minimization (ERM) solution is the sample mean $W = \frac{1}{n}\sum_{i=1}^{n} Z_i$, which is deterministic given S. Its generalization error can be computed exactly as (see Appendix A):

$$\mathrm{gen}(\mu, P_{W|S}) = \frac{2\sigma^2 d}{n}. \tag{26}$$

The bound in Lemma 1 is not applicable here due to the following two reasons: (1) W is a deterministic function of S, and hence $I(S; W) = \infty$; and (2) since Z is a Gaussian random vector, the loss function $\ell(w, Z) = \|w - Z\|_2^2$ is not sub-Gaussian for all $w \in \mathbb{R}^d$. Specifically, the variance of the loss function $\ell(w, Z)$ diverges as $\|w\|_2 \to \infty$, which implies that a uniform upper bound on $\Lambda_{\ell(w,Z)}(\lambda)$, $\forall w \in \mathbb{R}^d$ does not exist.

We can get around both of these issues by applying the ISMI bound in Theorem 2. Since $W \sim \mathcal{N}(\mu, \frac{\sigma^2 I_d}{n})$, the mutual information between each individual sample and the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAIT.2020.2991139, IEEE Journal on Selected Areas in Information Theory

output hypothesis $I(W;Z_i)$ can be computed exactly as (see Appendix A):

$$I(W;Z_i) = \frac{d}{2}\log\frac{n}{n-1}, \qquad i = 1,\cdots,n, \quad n \geq 2. \tag{27}$$

In addition, since $W \sim \mathcal{N}(\mu, \frac{\sigma^2 I_d}{n})$, it can be shown that $\ell(W,\widetilde{Z}) \sim \sigma_\ell^2 \chi_d^2$, where $\sigma_\ell^2 \triangleq \frac{(n+1)\sigma^2}{n}$, and χ_d^2 denotes the chi-squared distribution with d degrees of freedom. Note that the expectation of χ_d^2 distribution is d and its moment generating function is $(1-2\lambda)^{d/2}$. Therefore, the CGF of $\ell(\widetilde{W},\widetilde{Z})$ is given by

$$\Lambda_{\ell(\widetilde{W},\widetilde{Z})}(\lambda) = -d\sigma_\ell^2\lambda - \frac{d}{2}\log(1-2\sigma_\ell^2\lambda), \tag{28}$$

for $\lambda \in (-\infty, \frac{1}{2\sigma_\ell^2})$. Since W is the ERM solution, it follows that $\mathrm{gen}(\mu, P_{W|S}) \geq 0$, and we only need to consider the case $\lambda < 0$. It can be shown that (see Appendix A):

$$\Lambda_{\ell(\widetilde{W},\widetilde{Z})}(\lambda) \leq d\sigma_\ell^4\lambda^2 \triangleq \psi_-(-\lambda), \quad \lambda < 0. \tag{29}$$

Then, $\psi_-^{*-1}(y) = 2\sqrt{d\sigma_\ell^4 y}$. Combining the results in (27), we have

$$\mathrm{gen}(\mu, P_{W|S}) \leq \sigma^2 d\sqrt{\frac{2(n+1)^2}{n^2}\log\frac{n}{n-1}}. \tag{30}$$

As $n \to \infty$, the above bound is $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$, which is sub-optimal compared to the true generalization error computed in (26). We should note that techniques based on VC dimension [7] and algorithmic stability [13] also yield bounds of $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$.

B. Gaussian Process

In this subsection, we revisit the Gaussian process example studied in [19]. Let $\mathcal{W} = \{w \in \mathbb{R}^2 : \|w\|_2 = 1\}$, and $Z \sim \mathcal{N}(0, I_2)$ be a standard normal random vector in $\mathbb{R}^2$. The loss function is defined to be the following Gaussian process indexed by w:

$$\ell(w,Z) \triangleq -\langle w, Z\rangle, \quad \forall w \in \mathcal{W}. \tag{31}$$

Note that the loss function[3] $\ell(w,Z)$ is sub-Gaussian with parameter $R = 1$ for all $w \in \mathcal{W}$. In addition, the output hypothesis $w \in \mathcal{W}$ can also be represented equivalently using the phase of w. In other words, we can let ϕ be the unique number in $[0, 2\pi)$ such that $w = (\sin\phi, \cos\phi)$. For this problem, the empirical risk of a hypothesis $w \in \mathcal{W}$ is given by

$$L_S(w) = -\frac{1}{n}\sum_{i=1}^n \langle w, Z_i\rangle. \tag{32}$$

We consider two learning algorithms which are the same as the ones in [19]. The first is the ERM algorithm:

$$W = \arg\min_{\phi\in[0,2\pi)} L_S(w) = \arg\max_{\phi\in[0,2\pi)} \big\langle w, \frac{1}{n}\sum_{i=1}^n Z_i\big\rangle. \tag{33}$$

[3]The loss function can be negative here. We ignore the non-negativity assumption of the loss function; this does not affect our analysis.

The second is the ERM algorithm with additive noise:

$$W' = \Big(\arg\max_{\phi\in[0,2\pi)} \big\langle w, \frac{1}{n}\sum_{i=1}^n Z_i\big\rangle\Big) \oplus \xi \ (\mathrm{mod}\ 2\pi), \tag{34}$$

where the noise ξ is independent of S, and has an atom with probability mass ϵ at 0, and probability $1-\epsilon$ uniformly distributed on $(-\pi, \pi)$. Due to the symmetry of the problem, W and W' are uniformly distributed over $[0, 2\pi)$.

The generalization error of the ERM algorithm W can be computed exactly as (see Appendix B):

$$\mathrm{gen}(\mu, P_{W|S}) = \sqrt{\frac{\pi}{2n}}. \tag{35}$$

For the second algorithm W', since the noise ξ is independent from S, it follows that

$$\mathrm{gen}(\mu, P_{W'|S}) = \epsilon\sqrt{\frac{\pi}{2n}}. \tag{36}$$

The bound via $I(W;S)$ in Lemma 1 is not applicable, since W is deterministic given S and $I(W;S) = \infty$. Moreover, for the second algorithm W',

$$\begin{aligned} I(W';S) &= h(W') - h(W'|S) \\ &= \log 2\pi - h(\xi) = \infty, \end{aligned} \tag{37}$$

since ξ has a singular component at 0, and $h(\xi) = -\infty$.

Applying the ISMI bound in Theorem 2 to the ERM algorithm W, we have that

$$\begin{aligned} I(W;Z_i) &= h(W) - h(W|Z_i) \\ &= \log 2\pi - h(W|Z_i) \\ &= \log 2\pi - \mathbb{E}_{Z_i}[h(W|Z_i = z_i)], \end{aligned} \tag{38}$$

which we need to compute the conditional distribution $P_{W|Z_i=z_i}$. Note that given $Z_i = z_i$, the ERM solution

$$W = \arg\max_{\phi\in[0,2\pi)} \langle w, \frac{z_i}{n} + \frac{1}{n}\sum_{j\neq i} Z_i\rangle \tag{39}$$

depends on the other samples Z_j, $j \neq i$. Moreover, it can be shown that $P_{W|Z_i=z_i}$ is equivalent to the phase distribution of a Gaussian random vector $\mathcal{N}(\frac{z_i}{n}, \frac{n-1}{n^2}I_2)$ in polar coordinates.

Due to symmetry, we can always rotate the polar coordinates, such that $z_i = (r, 0)$, where $r \in \mathbb{R}^+$ is the ℓ_2 norm of z_i. Then, $P_{W|Z_i=z_i}$ is a function of r, and can be equivalently characterized and computed by the distribution $f\big(\phi\big|\|Z_i\| = r\big)$ provided in Appendix B. Since the norm of Z_i has a Rayleigh distribution with unit variance, it then follows that

$$I(W;Z_i) = \log 2\pi - \mathbb{E}_{\|Z_i\|}\Big[h\big(f(\phi\big|\|Z_i\| = r)\big)\Big]. \tag{40}$$

Applying Theorem 2, we obtain

$$|\mathrm{gen}(\mu, P_{W|S})| \leq \frac{1}{n}\sum_{i=1}^n \sqrt{2I(W;Z_i)} = \sqrt{2I(W;Z_i)}. \tag{41}$$

Similarly, we can compute the ISMI bound for W'.

Numerical comparisons are presented in Fig. 1 and Fig. 2. In both figures, we plot the ISMI bound, the CMI bound in [19],

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAIT.2020.2991139, IEEE Journal on Selected Areas in Information Theory

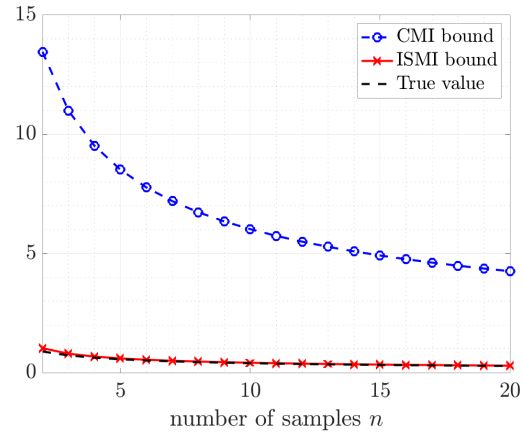


Fig. 1. Comparison of generalization bounds for the ERM algorithm.

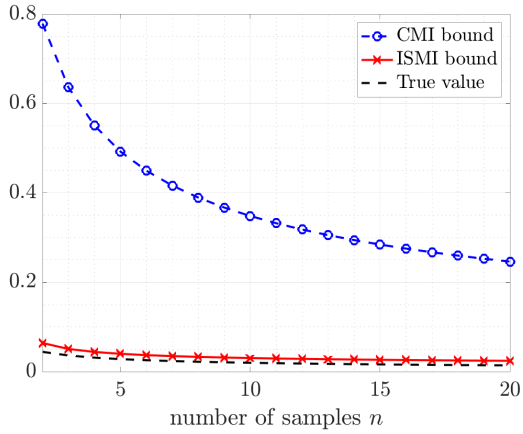


Fig. 2. Comparison of different generalization bounds for the ERM algorithm with additive noise $\epsilon = 0.05$.

and the true values of the generalization error, as functions of the number of samples n. In Fig. 1, we compare these bounds for the ERM solution W. Note that the CMI bound reduces to the classical chaining bound in this case. In Fig. 2, we evaluate these bounds for the noisy algorithm W' with $\epsilon = 0.05$. Both figures demonstrate that the ISMI bound is closer to the true values of the generalization error, and outperforms the CMI bound significantly. More details about the computations of both bounds can be found in Appendix B.

V. Noisy, Iterative Algorithms

In this section, we apply the ISMI bound in Theorem 2 to a class of noisy, iterative algorithms, specifically, stochastic gradient Langevin dynamics (SGLD).

A. SGLD Algorithm

We begin by introducing some notation to be used in this section. Denote the parameter vector at iteration t by $W_{(t)} \in \mathbb{R}^d$, and let $W_{(0)} \in \mathcal{W}$ denote an arbitrary initialization. At each iteration $t \geq 1$, we sample a training data point $Z_{U_{(t)}} \in S$, where $U_{(t)} \in \{1, ..., n\}$ denotes the random index of the sample selected at iteration t, and compute the gradient $\nabla \ell(W_{(t-1)}, Z_{U_{(t)}})$. We then scale the gradient by a step size $\eta_{(t)}$ and perturb it by isotropic Gaussian noise $\xi \sim \mathcal{N}(0, I_d)$. The overall update rule is as follows [33]:

$$W_{(t)} = W_{(t-1)} - \eta_{(t)} \nabla \ell(W_{(t-1)}, Z_{U_{(t)}}) + \sigma_{(t)} \xi, \qquad (42)$$

where $\sigma_{(t)}$ controls the variance of the Gaussian noise.

For $t \geq 0$, let $W^{(t)} \triangleq \{W_{(1)}, \cdots, W_{(t)}\}$ and $U^{(t)} \triangleq \{U_{(1)}, \cdots, U_{(t)}\}$. We assume that the training process takes K epochs, and the total number of iterations is $T = nK$. The output of the algorithm is $W = W_{(T)}$.

In the following, we use the same assumptions as in [29].

Assumption 1. *$\ell(w, Z)$ is R-sub-Gaussian with respect to $Z \sim \mu$, for every $w \in \mathcal{W}$.*

Assumption 2. *The gradients are bounded, i.e., $\sup_{w \in \mathcal{W}, z \in \mathcal{Z}} \|\nabla \ell(W, z)\|_2 \leq L$, for some $L > 0$.*

In [29], the following bound was obtained by upper bounding $I(W; S)$ in Lemma 1.

Lemma 3 ([29, Corollary 1])**.** *The generalization error of the SGLD algorithm is bounded by*

$$|\mathrm{gen}(\mu, P_{W|S})| \leq \sqrt{\frac{R^2}{n} \sum_{t=1}^{T} \frac{\eta_t^2 L^2}{\sigma_t^2}}. \qquad (43)$$

B. ISMI Bound for SGLD

We have the following proposition which characterizes the ISMI bound for the SGLD algorithm.

Proposition 3. *Suppose Assumption 1 and 2 hold, then we have the following ISMI bound on the generalization error for SGLD algorithm,*

$$|\mathrm{gen}(\mu, P_{W|S})| \leq \mathbb{E}_{U^{(T)}} \left[\frac{R}{n} \sum_{i=1}^{n} \sqrt{\sum_{\tau \in \mathcal{T}_i(U^{(T)})} \frac{\eta_{(\tau)}^2 L^2}{\sigma_{(\tau)}^2}} \right], \quad (44)$$

where $U^{(T)}$ denotes the random sample path, and $\mathcal{T}_i(U^{(T)})$ denote the set of iterations for which samples Z_i is selected for a given sample path $U^{(T)}$.

Proof. To apply the ISMI bound for SGLD, we modify the result in Theorem 2 by conditioning on the random sample path $U^{(T)}$,

$$\begin{aligned} &|\mathrm{gen}(\mu, P_{W|S})| \\ &= \left| \mathbb{E}_{U^{(T)}} \left[\frac{1}{n} \sum_{i=1}^{n} \Big(\mathbb{E}_{W, \widetilde{Z}} [\ell(W, \widetilde{Z}) | U^{(T)}] \right. \right. \\ &\qquad \left. \left. - \mathbb{E}_{W, Z_i} [\ell(W, Z_i) | U^{(T)}] \Big) \right] \right| \\ &\leq \frac{1}{|\mathcal{U}|} \sum_{u^{(T)} \in \mathcal{U}} \Big(\frac{1}{n} \sum_{i=1}^{n} \sqrt{2R^2 I(W; Z_i | U^{(T)} = u^{(T)})} \Big), \quad (45) \end{aligned}$$

where $\mathcal{U}$ denotes the set of all possible sample paths, and $I(W; Z_i | U^{(T)} = u^{(T)})$ is the mutual information [4] with conditional distribution $P(W, Z_i | U^{(T)} = u^{(T)})$.

[4]Note that this mutual information is different from the conditional mutual information $I(W; Z_i | U^{(T)}) = \mathbb{E}_{U^{(T)}} [I(W; Z_i | U^{(T)} = u^{(T)})]$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAIT.2020.2991139, IEEE Journal on Selected Areas in Information Theory

Let $\mathcal{T}_i(u^{(T)})$ denote the set of iterations for which sample Z_i is selected for a given sample path $u^{(T)}$. Using the chain rule of mutual information, we have

$$\begin{aligned}
&I(W; Z_i|U^{(T)} = u^{(T)}) \\
&\leq I(Z_i; W^{(T)}|U^{(T)} = u^{(T)}) \\
&= \sum_{\tau=1}^{T} I(Z_i; W_{(\tau)}|W_{(\tau-1)}, U^{(T)} = u^{(T)}) \\
&= \sum_{\tau\in\mathcal{T}_i(u^{(T)})} I(Z_i; W_{(\tau)}|W_{(\tau-1)}, U^{(T)} = u^{(T)}), \qquad (46)
\end{aligned}$$

where the last equality is due to the fact that given $u^{(T)}$ and $W_{(\tau-1)}$, Z_i is independent of $W_{(\tau)}$, if $\tau \notin \mathcal{T}_i(u^{(T)})$. For $\tau \in \mathcal{T}_i(u^{(T)})$, i.e., if Z_i is selected at iteration τ, we have

$$\begin{aligned}
&I(Z_i; W_{(\tau)}|W_{(\tau-1)}, U^{(T)} = u^{(T)}) \\
&= h\big(\eta_{(\tau)}\nabla\ell(W_{(\tau-1)}, Z_i) + \sigma_{(\tau)}\xi|W_{(\tau-1)}\big) - h(\sigma_{(\tau)}\xi).
\end{aligned}$$

Since we assume that $\sup_{w\in\mathcal{W}, Z\in\mathcal{Z}} \|\nabla\ell(W, Z)\|_2 \leq L$, we have

$$\begin{aligned}
&h\big(\eta_{(\tau)}\nabla\ell(W_{(\tau-1)}, Z_i) + \sigma_{(\tau)}\xi|W_{(\tau-1)}\big) \\
&\leq h\big(\eta_{(\tau)}\nabla\ell(W_{(\tau-1)}, Z_i) + \sigma_{(\tau)}\xi\big) \\
&\leq \frac{d}{2}\log\big(2\pi e\frac{\eta_{(\tau)}^2 L^2 + d\sigma_{(\tau)}^2}{d}\big). \qquad (47)
\end{aligned}$$

Due to the fact that ξ is an independent Gaussian noise, $h(\sigma_{(\tau)}\xi|W_{(\tau-1)}) = \frac{d}{2}\log\big(2\pi e\sigma_\tau^2\big)$, we have

$$I(Z_i; W_{(\tau)}|W_{(\tau-1)}, U^{(T)} = u^{(T)}) \leq \frac{d}{2}\log\big(1 + \frac{\eta_{(\tau)}^2 L^2}{d\sigma_{(\tau)}^2}\big).$$

Combining with (45), it follows that

$$|\mathrm{gen}(\mu, P_{W|S})| \leq \mathbb{E}_{U^{(T)}}\left[\frac{R}{n}\sum_{i=1}^{n}\sqrt{\sum_{\tau\in\mathcal{T}_i(U^{(T)})}\frac{\eta_{(\tau)}^2 L^2}{\sigma_{(\tau)}^2}}\right], \quad (48)$$

where we remove the log term by using $\log(1 + x) \leq x$. □

To compare the result of the ISMI bound in Proposition 3 and the bound in Lemma 3, we specify the parameters in the SGLD algorithms. As in [29], we set $\eta_{(t)} = \frac{c}{t}$, and $\sigma_{(t)} = \sqrt{\eta_t}$. We use the following "without replacement" sampling scheme for SGLD to further simplify the computation. Specifically, for the k-th training epoch, i.e., from the $((k-1)n+1)$-th to kn-th iterations, all training samples in S are used exactly once.

Then, the ISMI bound can be further bounded as follows:

$$\begin{aligned}
&|\mathrm{gen}(\mu, P_{W|S})| \\
&\leq \frac{RL}{n}\mathbb{E}_{U^{(T)}}\Big[\sum_{i=1}^{n}\sqrt{\sum_{\tau\in\mathcal{T}_i(U^{(T)})}\frac{c}{\tau}}\Big] \\
&\overset{(a)}{\leq} \frac{RL\sqrt{c}}{n}\sum_{i=1}^{n}\sqrt{\frac{1}{i} + \sum_{k=1}^{K-1}\frac{1}{nk}} \\
&\overset{(b)}{\leq} \frac{RL\sqrt{c}}{n}\sum_{i=1}^{n}\sqrt{\frac{1}{i} + \frac{\log(K-1)+1}{n}} \\
&\overset{(c)}{\leq} \frac{RL}{\sqrt{n}}\Big(\sqrt{c\log(K-1)+c} + o(\log\log K)\Big), \qquad (49)
\end{aligned}$$

where (a) follows from the sampling scheme that all samples are used exactly once in each epoch; (b) is due to the fact that $\sum_{k=1}^{K}\frac{1}{k} \leq \log(K)+1$; and (c) follows by computing the integral $\int_0^1\sqrt{\frac{1}{x}+1+\log(K-1)}dx$.

Comparing with the bound in [29],

$$|\mathrm{gen}(\mu, P_{W|S})| \leq \frac{RL}{\sqrt{n}}\sqrt{c\log(nK)+c}, \qquad (50)$$

it can be seen that our bound is tighter by a factor of $\sqrt{\log n}$ with the "without replacement" sampling scheme.

Remark 4. *We note that for the typical use of SGLD, the standard deviation of the noise is $\sigma_t = \sqrt{2\eta_t/\beta_t}$, where β_t denotes the inverse temperature at iteration t, and it is often set to be $\Theta(n)$. It is clear that $\beta = \Theta(n)$ will lead to a generalization bound that does not decay in n. Here, we choose $\beta_t = 2$ for comparison with the bound in [29], while in practice β may be a function of n and grow with t. An analysis of the generalization error bound of SGLD for arbitrary choices of β_t can be found in [32].*

VI. Empirical Evaluation of ISMI Bound for Logistic Regression

For some learning algorithms applied in practice, it is difficult to analytically characterize $P_{W|S}$, which makes the analytical evaluation of the ISMI bound challenging. In this section, we provide such an example, that of logistic regression, for which it is difficult to analytically characterize the learning algorithm via the conditional distribution $P_{W|S}$. We therefore empirically evaluate the ISMI bound via a mutual information estimator, and compare it to an empirical evaluation of the generalization error. We further note that the ISMI bound is much easier to estimate than the bound in Lemma 1 and the chaining bound in [19] due to the significant reduction in dimension that comes from estimating $I(Z_i; W)$ instead of $I(S; W)$.

Consider the binary classification problem, where the samples $Z = (X, Y)$, consisting of features $X \in \mathbb{R}^d$ and labels $Y \in \{\pm 1\}$. We assume that training samples are generated from the following distribution,

$$X \sim \mathcal{N}(\mu_Y, \Sigma), \quad Y \in \{\pm 1\}, \quad \mu_Y \in \mathbb{R}^d. \qquad (51)$$

The marginal distribution of X is the mixture of two Gaussian distributions, and we assume that $P(Y = -1) = P(Y = 1) = 1/2$.

A binary classifier is constructed as follows:

$$\hat{Y} = \begin{cases} 1, & w^T X \geq 0; \\ -1, & \text{else.} \end{cases} \qquad (52)$$

We adopt classification error $\ell(w, Z) = \mathbb{1}_{\{Y\neq\hat{Y}\}}$ to compute the generalization error, then the empirical risk with n i.i.d. samples is

$$L_S(w) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{Y_i\neq\hat{Y}_i\}}. \qquad (53)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAIT.2020.2991139, IEEE Journal on Selected Areas in Information Theory

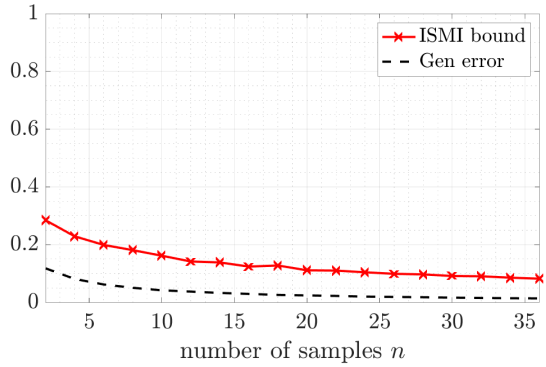


Fig. 3. Empirical evaluation of ISMI bound and generalization error in Logistic regression.

Since the empirical risk function is not differentiable, we learn W by minimizing the following loss function of logistic regression:

$$W = \arg\min_{w\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^{n} \log(1+e^{-Y_i w^T X_i}). \tag{54}$$

In general, it is difficult to obtain a closed form solution for this optimization problem, and therefore, (54) is usually solved numerically. This makes it difficult to analytically characterize the conditional distribution $P_{W|Z_i}$, which in turn makes it challenging to compute the generalization error and the ISMI bound analytically.

Alternatively, we can empirically estimate the generalization error and the ISMI bound. Specifically, we train W for N times using N sets of independent samples, and we use the K-nearest neighbor based mutual information estimator [34], [35] to estimate $I(W;Z_i)$ with N i.i.d. samples of W and Z_i. Note that the K-nearest neighbor based mutual information estimator is consistent, and its mean squared estimation error can be upper bounded by $\mathcal{O}(N^{-\frac{2}{d_W+d_Z}})$, where $d_W = d$ is the dimension of the weights W, and $d_Z = d+1$ is the dimension of Z [35]. Moreover, since we use classification error to compute generalization error, $\ell(W,Z)$ is bounded by 1. Then, by Hoeffding's lemma, $\ell(W,Z)$ is $\frac{1}{2}$-sub-Gaussian. Thus, the ISMI bound can be estimated by

$$\frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{\hat{I}(W;Z_i)}{2}}, \tag{55}$$

where $\hat{I}(W;Z_i)$ is the estimate of $I(W;Z_i)$. If we apply an optimization algorithm that does not depend on the order of the samples, e.g., gradient descent and stochastic gradient descent with random shuffling, we then only need to estimate one $\sqrt{\frac{\hat{I}(W;Z_1)}{2}}$ instead of estimating $\hat{I}(W;Z_i)$ for all $1\le i\le n$.

We note that the bound in Lemma 1 is difficult to estimate due to the high dimension of S, which scales linearly with n. Specifically, the training dataset S consists of n samples, and therefore $I(W;S)$ is the mutual information between two random vectors with dimensions d and $n(d+1)$. As shown in [35], due to the curse of dimensionality, it is impossible to construct a consistent mutual information estimator for large n. We also note that the exact computation of Wasserstein distances is costly in general, as it requires the solution of an optimal transport problem [36]. Moreover, similar high dimensional issue makes it even more difficult to directly estimate $\mathbb{W}(P_W, P_{W|S})$ in the Wasserstein distance based generalization bound in [24], [25].

In Fig. 3, we plot an empirical estimate of the ISMI bound using (55), and compare it to the generalization error. In the simulation, we chose the following model parameters: $d=2$ and $\mu_1=(1,1)$, $\mu_{-1}=(-1,-1)$ with $\Sigma = 4I$. We used the K-nearest neighbor based mutual information estimator (revised KSG estimator) in [35] with $N=5000$ i.i.d. samples. It can be seen that the ISMI bound has a similar convergence behavior as the true generalization error as number of training samples n increases.

VII. Conclusions

In this paper, we proposed a tighter information-theoretic upper bound on the generalization error using the mutual information $I(Z_i;W)$ between each individual training sample Z_i and the output hypothesis W of the learning algorithm. We showed that compared to existing studies, our bound is more broadly applicable, and is considerably tighter. More importantly, the individual sample mutual information is between two vectors whose dimensions do not scale with the sample size n. Therefore, unlike the existing bounds in [18], [19], [25], the ISMI bound can easily be evaluated empirically in practice. As suggested by recent works, the proposed ISMI bound could be further improved by combining with the chaining method [22], or data-dependent estimates [32]. The proposed information-theoretic framework can also be used to guide model compression in deep learning [37].

Appendix A
Section IV-A Details

A. Generalization Error

For this example, the generalization can be computed as

$$\begin{aligned}
&\mathrm{gen}(\mu, P_{W|S})\\
&= \mathbb{E}_{W,S}[L_\mu(W) - L_S(W)]\\
&= \mathbb{E}_S\Big[\mathbb{E}_{\widetilde{Z}}[\|\bar{Z}-\widetilde{Z}\|_2^2] - \frac{1}{n}\sum_{i=1}^{n}\|\bar{Z}-Z_i\|_2^2\Big]\\
&= \mathbb{E}_{S,\widetilde{Z}}[\mathrm{Tr}((\bar{Z}-\widetilde{Z})(\bar{Z}-\widetilde{Z})^\top)]\\
&\quad - \mathbb{E}_S\Big[\frac{1}{n}\sum_{i=1}^{n}\mathrm{Tr}((\bar{Z}-Z_i)(\bar{Z}-Z_i)^\top)\Big]\\
&= \mathrm{Tr}(\mathrm{Cov}[\bar{Z}]) + \mathrm{Tr}(\mathrm{Cov}[\widetilde{Z}]) - \frac{n-1}{n}\mathrm{Tr}(\mathrm{Cov}[Z])\\
&= \frac{2}{n}\mathrm{Tr}(\mathrm{Cov}[Z])
\end{aligned} \tag{56}$$

Since $\mathrm{Cov}[Z] = \sigma^2 I_d$, we have $\mathrm{gen}(\mu, P_{W|S}) = \frac{2\sigma^2 d}{n}$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAIT.2020.2991139, IEEE Journal on Selected Areas in Information Theory

B. Individual Mutual Information

We note that both W and n i.i.d. samples Z_i are Gaussian, then the individual mutual information $I(W;Z_i)$ is a function captured by the following covariance matrix,

$$\mathrm{Cov}[W, X_i] = \begin{pmatrix} \Sigma/n & \Sigma/n \\ \Sigma/n & \Sigma \end{pmatrix}. \tag{57}$$

Then, we have

$$\begin{aligned} I(W;X_i) &= \frac{1}{2}\log\frac{|\mathrm{Cov}[W]||\mathrm{Cov}[X_i]|}{|\mathrm{Cov}[W,X_i]|} \\ &= \frac{1}{2}\log\frac{|\Sigma/n||\Sigma|}{|\frac{n-1}{n^2}\Sigma||\Sigma|} \\ &= \frac{d}{2}\log\frac{n}{n-1}, \end{aligned} \tag{58}$$

for all $i = 1, \ldots, n$.

C. Upper bound for CGF

Note that the CGF of $\ell(\widetilde{W}, \widetilde{Z})$ is given by

$$\begin{aligned} \Lambda_{\ell(\widetilde{W},\widetilde{Z})}(\lambda) &= -d\sigma_\ell^2\lambda - \frac{d}{2}\log(1-2\sigma_\ell^2\lambda) \\ &= \frac{d}{2}(-u - \log(1-u)),\ \lambda \in (-\infty, \frac{1}{2\sigma_\ell^2}), \end{aligned} \tag{59}$$

where $u \triangleq 2\sigma_\ell^2\lambda$. Further note that

$$-u - \log(1-u) \le \frac{u^2}{2},\ u < 0. \tag{60}$$

We therefore have the following upper bound on the CGF of $\ell(\widetilde{W}, \widetilde{Z})$:

$$\Lambda_{\ell(\widetilde{W},\widetilde{Z})}(\lambda) \le d\sigma_\ell^4\lambda^2, \quad \lambda < 0. \tag{61}$$

Appendix B
Section IV-B Details

A. Generalization Error

Note that the expectation of the population risk is

$$\mathbb{E}_{W,S}[L_\mu(W)] = \mathbb{E}_{W,Z}[-\langle W, Z\rangle] = 0, \tag{62}$$

since W and Z are independent. Then, the generalization error can be computed as

$$\begin{aligned} &\mathrm{gen}(\mu, P_{W|S}) = \mathbb{E}[-L_S(W)] \\ &= \mathbb{E}_{W,S}[\langle W, \frac{1}{n}\sum_{i=1}^n Z_i\rangle] \\ &= \mathbb{E}_{W,S}\Big\|\frac{1}{n}\sum_{i=1}^n Z_i\Big\|_2 = \sqrt{\frac{\pi}{2n}}, \end{aligned} \tag{63}$$

where the last step is due to the fact that the distribution of $\|\frac{1}{n}\sum_{i=1}^n Z_i\|_2$ is $\mathrm{Rayleigh}(\frac{1}{n})$.

B. Individual Sample Mutual Information Bound

To compute the ISMI bound, we need the conditional distribution $P_{W|Z_i=z_i}$. Note that given $Z_i = z_i$, the ERM solution is

$$W = \arg\max_{\phi\in[0,2\pi)}\langle w, \frac{z_i}{n} + \frac{1}{n}\sum_{j\neq i} Z_i\rangle.$$

Also note that since $\phi \in [0, 2\pi)$, $P_{W|Z_i=z_i}$ is equivalent to the phase distribution of a Gaussian random vector $\mathcal{N}(\frac{z_i}{n}, \frac{n-1}{n^2}I_2)$ in polar coordinates. Since entropy is shift-invariant, we can always rotate the polar coordinates, such that $z_i = (r, 0)$, where $r \in \mathbb{R}^+$ is the ℓ_2 norm of z_i.

The joint distribution of radius and phase (ρ, ϕ) in polar coordinates can be obtained by applying the Jacobian method to the Gaussian distribution $\mathcal{N}(\frac{(r,0)}{n}, \frac{n-1}{n^2}I_2)$, and we have

$$f\big(\rho, \phi\ \big|\|Z_i\| = r\big) = \frac{n^2\rho}{2\pi(n-1)}e^{-\frac{n^2\rho^2+r^2}{2(n-1)}}e^{-\frac{nr\rho\cos\phi}{(n-1)}}, \tag{64}$$

for $\rho \in [0, \infty),\ \phi \in [0, 2\pi)$.

Then, the marginal distribution of ϕ can be computed by integrating out ρ from the joint distribution $f\big(\rho, \phi\big|\|Z_i\| = r\big)$:

$$\begin{aligned} &f\big(\phi\big|\|Z_i\| = r\big) \\ &= \int_0^\infty f\big(\rho, \phi\ \big|\|Z_i\| = r\big)d\rho \\ &= \frac{1}{2\pi}e^{-\frac{r^2}{2(n-1)}} + \frac{r\cos\phi}{\sqrt{2\pi(n-1)}}e^{-\frac{r^2\sin^2\phi}{2(n-1)}}Q(-\frac{r\cos\phi}{n-1}), \end{aligned}$$

where $Q(x)$ is the complementary cumulative distribution function of the standard normal distribution.

Thus, the ISMI bound for the ERM algorithm W can be evaluated using the following expression via numerical integration:

$$\begin{aligned} &|\mathrm{gen}(\mu, P_{W|S})| \le \sqrt{2I(W;Z_i)} \\ &= \sqrt{2\log 2\pi - 2\mathbb{E}_{\|Z_i\|}\Big[h\big(f(\phi\big|\|Z_i\| = r)\big)\Big]}. \end{aligned} \tag{65}$$

Similarly, the ISMI bound for algorithm W' can be computed via numerical integration using

$$\begin{aligned} &|\mathrm{gen}(\mu, P_{W'|S})| \le \sqrt{2I(W';Z_i)} \\ &= \sqrt{2\log 2\pi - 2\mathbb{E}_{\|Z_i\|}\Big[h\big(f(W'\big|Z_i = z_i)\big)\Big]}, \end{aligned} \tag{66}$$

where the conditional distribution $P_{W'|Z_i=z_i}$ can be characterized by the phase distribution

$$\begin{aligned} &f(\phi'\big|\|Z_i\| = r) \\ &= \frac{1-\epsilon}{2\pi} + \frac{\epsilon}{2\pi}e^{-\frac{r^2}{2(n-1)}} \\ &\quad + \frac{\epsilon r\cos\phi'}{\sqrt{2\pi(n-1)}}e^{-\frac{r^2\sin^2\phi'}{2(n-1)}}Q(-\frac{r\cos\phi'}{n-1}). \end{aligned} \tag{67}$$

C. Chaining Mutual Information Bound

The CMI bound is computed based on the values provided in Table 1 in [19]. We note that the CMI bound in [19] is evaluated for the case $n = 1$, i.e., there is only one training sample. To plot the CMI bound in [19] as a function of the

number of samples n, we normalize the CMI bound by a $\sqrt{n}$ factor in Figures 1 and 2, since $\mathcal{O}(1/\sqrt{n})$ is the true convergence rate for the generalization error as shown in (63). For instance, Table 1 in [19] shows that the CMI bound for the ERM solution W is 19.0352 when $n = 1$, which is equivalent to the classical chaining bound. We therefore plot the curve $\frac{19.0352}{\sqrt{n}}$ as the CMI bound for comparison with the proposed ISMI bound in Figure 1.

REFERENCES

[1] Y. Bu, S. Zou, and V. V. Veeravalli, "Tightening mutual information based bounds on generalization error," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, pp. 587–591, 2019.

[2] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*. MIT press Cambridge, 2016.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 1097–1105, 2012.

[4] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55–75, 2018.

[5] B. Huval, T. Wang, S. Tandon, J. Kiske, W. Song, J. Pazhayampallil, M. Andriluka, P. Rajpurkar, T. Migimatsu, and R. Cheng-Yue, "An empirical evaluation of deep learning on highway driving," *arXiv preprint arXiv:1504.01716*, 2015.

[6] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in Bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

[7] S. Boucheron, O. Bousquet, and G. Lugosi, "Theory of classification: A survey of some recent advances," *ESAIM: Probability and Statistics*, vol. 9, pp. 323–375, 2005.

[8] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

[9] M. Anthony and P. L. Bartlett, *Neural network learning: Theoretical foundations*. Cambridge University Press, 2009.

[10] B. Neyshabur, R. Tomioka, and N. Srebro, "In search of the real inductive bias: On the role of implicit regularization in deep learning," *arXiv preprint arXiv:1412.6614*, 2014.

[11] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *Proc. International Conference on Learning Representations (ICLR)*, 2017.

[12] D. A. McAllester, "Some PAC-Bayesian theorems," *Machine Learning*, vol. 37, no. 3, pp. 355–363, 1999.

[13] O. Bousquet and A. Elisseeff, "Stability and generalization," *J. Mach. Learn. Res.*, vol. 2, pp. 499–526, Mar 2002.

[14] A. Elisseeff, T. Evgeniou, and M. Pontil, "Stability of randomized learning algorithms," *J. Mach. Learn. Res.*, vol. 6, pp. 55–79, Jan. 2005.

[15] N. Littlestone and M. Warmuth, "Relating data compression and learnability," *Technical report, University of California, Santa Cruz*, 1986.

[16] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky, "Spectrally-normalized margin bounds for neural networks," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 6240–6249, 2017.

[17] D. Russo and J. Zou, "Controlling bias in adaptive data analysis using information theory," in *Proc. International Conference on Artifical Intelligence and Statistics (AISTATS)*, pp. 1232–1240, 2016.

[18] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 2524–2533, 2017.

[19] A. Asadi, E. Abbe, and S. Verdu, "Chaining mutual information and tightening generalization bounds," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 7245–7254, 2018.

[20] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, "Learnability, stability and uniform convergence," *J. Mach. Learn. Res.*, vol. 11, pp. 2635–2670, Oct. 2010.

[21] R. Bassily, S. Moran, I. Nachum, J. Shafer, and A. Yehudayoff, "Learners that leak little information," *arXiv preprint arXiv:1710.05233*, 2017.

[22] A. Asadi and E. Abbe, "Chaining meets chain rule: Multilevel entropic regularization and training of neural nets," *arXiv preprint arXiv:1906.11148*, 2019.

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

[24] A. T. Lopez and V. Jog, "Generalization error bounds using Wasserstein distances," in *Proc. IEEE Information Theory Workshop (ITW)*, pp. 1–5, 2018.

[25] H. Wang, M. Diaz, J. S. S. Filho, and F. P. Calmon, "An information-theoretic view of generalization via Wasserstein distance," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, 2019.

[26] I. Issa and M. Gastpar, "Computable bounds on the exploration bias," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, pp. 576–580, IEEE, 2018.

[27] I. Issa, A. R. Esposito, and M. Gastpar, "Strengthened information-theoretic bounds on the generalization error," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, pp. 582–586, IEEE, 2019.

[28] I. M. Alabdulmohsin, "Algorithmic stability and uniform generalization," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 19–27, 2015.

[29] A. Pensia, V. Jog, and P. Loh, "Generalization error bounds for noisy, iterative algorithms," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, pp. 546–550, June 2018.

[30] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[31] J. Jiao, Y. Han, and T. Weissman, "Dependence measures bounding the exploration bias for general measurements," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, pp. 1475–1479, 2017.

[32] J. Negrea, M. Haghifam, G. K. Dziugaite, A. Khisti, and D. M. Roy, "Information-theoretic generalization bounds for SGLD via data-dependent estimates," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 11013–11023, 2019.

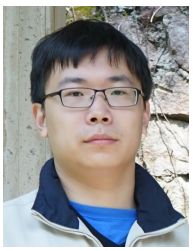
[33] M. Welling and Y. Teh, "Bayesian learning via stochastic gradient Langevin dynamics," in *Proc. International Conference on Machine Learning (ICML)*, pp. 681–688, 2011.

[34] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, p. 066138, 2004.

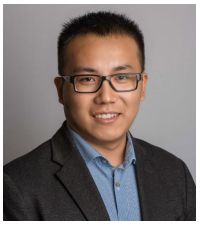
[35] W. Gao, S. Oh, and P. Viswanath, "Demystifying fixed k-nearest neighbor information estimators," *IEEE Trans. Inform. Theory*, vol. 64, no. 8, pp. 5629–5661, 2018.

[36] M. Rowland, J. Hron, Y. Tang, K. Choromanski, T. Sarlos, and A. Weller, "Orthogonal estimation of wasserstein distances," in *Proc. International Conference on Artifical Intelligence and Statistics (AISTATS)*, pp. 186–195, 2019.

[37] Y. Bu, W. Gao, S. Zou, and V. V. Veeravalli, "Information-theoretic understanding of population risk improvement with model compression," in *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2020.



Yuheng Bu (S'16-M'19) received the Ph.D. degree in Electrical and Computer Engineering from University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 2019, and the B.S. degree (with honors) in Electrical Engineering from Tsinghua University, Beijing, China, in 2014. Since September 2019, He has been a postdoctoral research associate at the Institute for Data, Systems, and Society, Massachusetts Institute of Technology. Dr. Bu's research interests include machine learning, information theory and statistical signal processing. He received the Yi-Min Wang and Pi-Yu Chung Research award in 2019.



Shaofeng Zou (S'14-M'16) received the Ph.D. degree in Electrical and Computer Engineering from Syracuse University in 2016. He received the B.E. degree (with honors) from Shanghai Jiao Tong University, Shanghai, China, in 2011. He was a postdoctoral research associate at the Coordinated Science Lab, University of Illinois at Urbana-Champaign during 2016-2018. He joined the Department of Electrical Engineering, University at Buffalo, The State University of New York, in 2018, where he is currently an Assistant Professor. Dr. Zou's research interests include statistical signal processing, machine learning, and information theory.

Venugopal V. Veeravalli (M'92-SM'98-F'06) received the B.Tech. (Silver Medal Honors) degree in electrical engineering from the Indian Institute of Technology, Bombay, India, in 1985, the M.S. degree in electrical engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 1987, and the Ph.D. degree in electrical engineering from the University of Illinois at Urbana-Champaign, Champaign, IL, USA, in 1992.

He joined the University of Illinois at Urbana-Champaign in 2000, where he is currently the Henry Magnuski Professor with the Department of Electrical and Computer Engineering, and where he is also affiliated with the Department of Statistics, the Coordinated Science Laboratory, and the Information Trust Institute. From 2003 to 2005, he was a Program Director for communications research at the U.S. National Science Foundation in Arlington, VA, USA. He has previously held academic positions at Harvard University, Rice University, and Cornell University, and has been on sabbatical at MIT, IISc Bangalore, and Qualcomm, Inc. His research interests include statistical signal processing, machine learning, detection and estimation theory, information theory, and stochastic control, with applications to sensor networks, cyberphysical systems, and wireless communications. A recent emphasis of his research has been on signal processing and machine learning for data science applications.

Prof. Veeravalli was a Distinguished Lecturer for the IEEE Signal Processing Society from 2010 and 2011. He has been on the Board of Governors of the IEEE Information Theory Society. He has been an Associate Editor for Detection and Estimation for the IEEE Transactions on Information Theory and for the IEEE Transactions on Wireless Communications. He is a recipient of the IEEE Browder J. Thompson Best Paper Award, the National Science Foundation CAREER Award, and the Presidential Early Career Award for Scientists and Engineers, and the Wald Prize in Sequential Analysis.