Tightening Mutual Information Based Bounds on Generalization Error

Yuheng Bu* Shaofeng Zou† Venugopal V. Veeravalli*
*University of Illinois at Urbana-Champaign †University at Buffalo, the State University of New York
Email: bu3@illinois.edu, szou3@buffalo.edu, vvv@illinois.edu

***Abstract*—A mutual information based upper bound on the generalization error of a supervised learning algorithm is derived in this paper. The bound is constructed in terms of the mutual information between each individual training sample and the output of the learning algorithm, which requires weaker conditions on the loss function, but provides a tighter characterization of the generalization error than existing studies. Examples are further provided to demonstrate that the bound derived in this paper is tighter, and has a broader 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.**

I. Introduction

Consider an instance space $\mathcal{Z}$, a continuous hypothesis space $\mathcal{W}$, and a nonnegative loss function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}^+$. A training dataset $S = \{Z_1, \cdots, Z_n\}$ consists of n i.i.d samples $Z_i \in \mathcal{Z}$ drawn from an unknown distribution μ. 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 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 data set S to a hypothesis W according to a conditional distribution $P_{W|S}$. The *generalization error* 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:

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

where the expectation is taken over the joint distribution $P_{S,W} = P_S \otimes P_{W|S}$. The generalization error is used to measure the extent to which the learning algorithm overfits the training data.

Traditional ways of bounding the generalization error can be categorized into two groups: (1) by measuring the complexity of the hypothesis space $\mathcal{W}$, e.g., VC dimension and Rademacher complexity [1]; and (2) by exploring properties of the learning algorithm, e.g., uniform stability [2]. Recently, it was proposed in [3] and further studied in [4] and [5] that the metric of mutual information can be used to develop upper bounds on the generalization error of a learning algorithm. Such an information-theoretic framework can handle a broader range of problems, e.g., problems with unbounded loss function. More importantly, it offers an information-theoretic point of view on how to improve the generalization capability of a learning algorithm.

In this paper, we follow the information-theoretic framework in [3]–[5]. Our main contribution is a tighter upper bound on the generalization error using the mutual information $I(Z_i; W)$ between an individual training sample Z_i and the output hypothesis W of the learning algorithm. We show that compared to existing studies, our bound has a broader applicability, and can be considerably tighter.

A. Main Contributions and Comparison to Related Works

The following lemma from [4] provides an upper bound on the generalization error using the mutual information $I(S; W)$ between the training data set S and the output hypothesis W.

Lemma 1. *[4, Theorem 1] Suppose $\ell(w, Z)$ is R-sub-Gaussian* [1] *under $Z \sim \mu$ for all $w \in \mathcal{W}$, then*

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

This mutual information based bound in (4) is related to the on-average stability [6], and quantifies the overall dependence between the output of the learning algorithm and its input dataset using $I(S;W)$. By further exploiting the structure of the hypothesis space and the dependency between the algorithm input and output, the authors of [5] 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 [5] both suffer from the following two shortcomings. First, for empirical risk minimization (ERM), if W is the unique minimizer of $L_S(w)$ in $\mathcal{W}$, 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)

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

978-1-5386-9291-2/19/$31.00 ©2019 IEEE

under $Z \sim \mu$ *for all* $w \in \mathcal{W}$, which may not hold for many problems.

In this paper, we get around these shortcomings by combining the idea of algorithmic stability [6], [7] and the information theoretic framework. Specifically, an algorithm is stable if the output hypothesis does not change too much with the replacement of any *individual* training sample, and if an algorithm is stable, then it generalizes well [6], [7]. Motivated by these facts, we tighten the mutual information based generalization error bound by considering the individual sample mutual information (ISMI) $I(W;Z_i)$. Compared with the bound in Lemma 1, and the CMI bound in [5], the ISMI bound requires a weaker condition on the CGF of the loss function, is applicable to a broader range of problems, and provides a tighter characterization of the generalization error. We also comprehensively study three examples, and compare the ISMI bound with existing results to demonstrate its superiority.

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 logarithms are natural ones.

The cumulant generating function (CGF) of a random variable X is defined as $\Lambda_X(\lambda) \triangleq \log \mathbb{E}[e^{\lambda(X-\mathbb{E}X)}]$. It can be verified that $\Lambda_X(0) = \Lambda_X'(0) = 0$, and that $\Lambda_X(\lambda)$ is convex if it exists.

Definition 1. *For a convex function* ψ *defined on the interval* $[0,b)$*, where* $0 < b \le \infty$*, its Legendre dual* ψ^* *is defined as*

$$\psi^*(x) \triangleq \sup_{\lambda\in[0,b)} \big(\lambda x - \psi(\lambda)\big). \tag{5}$$

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

Lemma 2. *[8, Lemma 2.4] Assume that* $\psi(0) = \psi'(0) = 0$*. Then* $\psi^*(x)$ *defined above is a nonnegative convex and non-decreasing function on* $[0,\infty)$ *with* $\psi^*(0) = 0$*. Moreover, its inverse function* $\psi^{*-1}(y) = \inf\{x \ge 0 : \psi^*(x) \ge y\}$ *is concave, and can be written as*

$$\psi^{*-1}(y) = \inf_{\lambda\in(0,b)} \Big(\frac{y+\psi(\lambda)}{\lambda}\Big). \tag{6}$$

For a R-sub-Gaussian random variable X, let $\psi(\lambda) = \Lambda_X(\lambda) = \frac{R^2\lambda^2}{2}$, then by Lemma 2, $\psi^{*-1}(y) = \sqrt{2R^2y}$.

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

In this section, we first generalize the decoupling lemma in [4, Lemma 1] to a more general setting, and then tighten the bound on generalization error via $I(W;Z_i)$.

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 $f(\widetilde{W},\widetilde{Z})$ is upper bounded for $\lambda \in (b_-, b_+)$, we have the following theorem.

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

$$\mathbb{E}[f(W,Z)] - \mathbb{E}[f(\widetilde{W},\widetilde{Z})] \le \psi_+^{*-1}\big(I(W;Z)\big), \tag{7}$$

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

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

$$D(P\|Q) = \sup_{g\in\mathcal{G}} \Big\{\mathbb{E}_P[g(X)] - \log \mathbb{E}_Q[e^{g(X)}]\Big\}, \tag{9}$$

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 the equality is achieved when $g = \log\frac{P}{Q}$. It then follows that $\forall\lambda \in [0,b_+)$,

$$\begin{aligned} D(P_{W,Z}\|P_W\otimes P_Z) &\ge \mathbb{E}[\lambda f(W,Z)] - \log\mathbb{E}[e^{\lambda f(\widetilde{W},\widetilde{Z})}] \\ &\ge \lambda(\mathbb{E}[f(W,Z)] - \mathbb{E}[f(\widetilde{W},\widetilde{Z})]) - \psi_+(\lambda), \end{aligned} \tag{10}$$

where the last inequality follows from the assumption that

$$\log\mathbb{E}[e^{\lambda(f(\widetilde{W},\widetilde{Z})-\mathbb{E}f(\widetilde{W},\widetilde{Z}))}] \le \psi_+(\lambda), \quad \forall\lambda\in[0,b_+). \tag{11}$$

Similarly, $\forall\lambda \in (b_-,0]$, it follows that

$$\begin{aligned} &D(P_{W,Z}\|P_W\otimes P_Z) \\ &\qquad \ge \lambda(\mathbb{E}[f(W,Z)] - \mathbb{E}[f(\widetilde{W},\widetilde{Z})]) - \psi_-(-\lambda). \end{aligned} \tag{12}$$

If $\lambda \in [0,b_+)$,

$$\begin{aligned} \mathbb{E}[f(W,Z)] - \mathbb{E}[f(\widetilde{W},\widetilde{Z})] &\le \inf_{\lambda\in[0,b_+)} \frac{I(W;Z)+\psi_+(\lambda)}{\lambda} \\ &= \psi_+^{*-1}\big(I(W;Z)\big), \end{aligned} \tag{13}$$

and if $\lambda \in (b_-,0]$,

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

where the equalities in (13) and (14) follow from Lemma 2. □

Theorem 1 provides a more general characterization of the decoupling estimate than existing results. Specifically, it is assumed that the CGF of $f(w,Z)$ is bounded for all $w \in \mathcal{W}$ in [4, Lemma 1] and [9, Theorem 2], whereas in Theorem 1, it is only assumed that the CGF of $f(\widetilde{W},\widetilde{Z})$ is bounded in expectation under $P_W \otimes P_Z$.

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 an upper bound on the generalization error via $I(W;Z_i)$.

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

$$\mathrm{gen}(\mu, P_{W|S}) \le \frac{1}{n}\sum_{i=1}^{n} \psi_-^{*-1}\big(I(W;Z_i)\big), \tag{15}$$

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

Proof. The generalization error can be written as follows:

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

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. □

The following Proposition shows that the ISMI bound is always tighter than the bound in Lemma 1.

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

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

Proof. It is clear that if $\ell(w,Z)$ is R-sub-Gaussian under $Z \sim \mu$ for all $w \in \mathcal{W}$, then $\ell(\widetilde{W},\widetilde{Z})$ is also R-sub-Gaussian. For R-sub-Gaussian random variables, it is easy to show that $\psi_+^{-1}(y) = \psi_-^{-1}(y) = \sqrt{2R^2 y}$. The first inequality then follows from Theorem 2.

For the second part, by the chain rule of mutual information,

$$I(W;S) = \sum_{i=1}^{n} I(W;Z_i|Z^{i-1}) \ge \sum_{i=1}^{n} I(W;Z_i), \tag{17}$$

where $Z^j = \{Z_1,\cdots,Z_j\}$, and the last step follows by the fact that Z_i and Z^{i-1} are independent. Applying Jensen's inequality completes the proof. □

Remark 1. *If $\psi_+^{*-1}(y)$ and $\psi_-^{*-1}(y)$ are concave, it can be shown that the ISMI bound in Theorem 2 is also tighter than the bound using $I(S;W)$ in [9].*

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

In this section, we consider two examples with infinite $I(W;S)$. We show that for these two examples, the upper bound on generalization error in Lemma 1 blows up, whereas the ISMI bound in Theorem 2 still provides an accurate approximation.

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 mean square error $\ell(w,Z) \triangleq \mathbb{E}\|w-Z\|_2^2$. The empirical risk with n i.i.d. samples is $L_S(w) \triangleq \frac{1}{n}\sum_{i=1}^{n}\|w-Z_i\|_2^2$. 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 follows:

$$\mathrm{gen}(\mu, P_{W|S}) = \frac{2\sigma^2 d}{n}. \tag{18}$$

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. 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.

Both of these issues can be solved 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 output hypothesis $I(W;Z_i)$ can be computed exactly as follows:

$$I(W;Z_i) = \frac{d}{2}\log\frac{n}{n-1}, \qquad i = 1,\cdots,n. \tag{19}$$

In addition, since $W \sim \mathcal{N}(\mu, \frac{\sigma^2 I_d}{n})$, it can be shown that $\ell(\widetilde{W},\widetilde{Z}) \sim \sigma_\ell^2 \chi_d^2$, 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. Then, the CGF of $\ell(\widetilde{W},\widetilde{Z})$ is

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

Since W is the ERM solution, it follows that $\mathrm{gen}(\mu, P_{W|S}) \ge 0$. We only need to consider the case $\lambda < 0$. It can be shown that

$$\Lambda_{\ell(\widetilde{W},\widetilde{Z})}(\lambda) \le d\sigma_\ell^4\lambda^2 \triangleq \psi_-(-\lambda), \quad \lambda < 0. \tag{20}$$

Then, $\psi_-^{*-1}(y) = 2\sqrt{d\sigma_\ell^4 y}$. Combining the results in (19), we have

$$\mathrm{gen}(\mu, P_{W|S}) \le \sigma^2 d\sqrt{\frac{2(n+1)^2}{n^2}\log\frac{n}{n-1}}. \tag{21}$$

As $n \to \infty$, the above bound is $\mathcal{O}(\frac{1}{\sqrt{n}})$, which is usually the case when one applies bounding techniques based on the VC dimension [1], and algorithmic stability [2].

B. Gaussian Process

In this subsection, we revisit the example studied in [5]. 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:

$$\ell(w,Z) \triangleq -\langle w, Z\rangle, \quad \forall w \in \mathcal{W}. \tag{22}$$

Note that the loss function $\ell(w,Z)$ 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

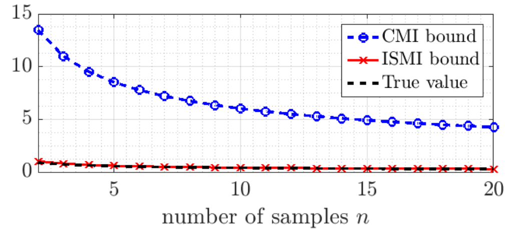

Fig. 1. Comparison of generalization bounds for the ERM algorithm.

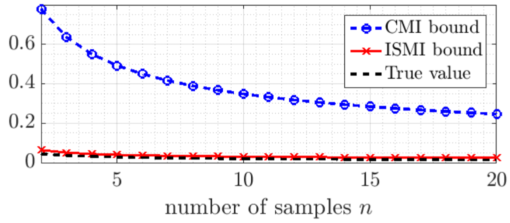

Fig. 2. Comparison of different generalization bounds for the ERM algorithm with an additive noise.

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

We consider two learning algorithms which are the same as the ones in [5]. The first is the ERM algorithm:

$$W = \arg\min_{\phi\in[0,2\pi)} L_S(w) = \arg\max_{\phi\in[0,2\pi)} \langle w, \frac{1}{n}\sum_{i=1}^{n} Z_i\rangle. \tag{23}$$

The second is the ERM algorithm with additive noise:

$$W' = \Big(\arg\max_{\phi\in[0,2\pi)} \langle w, \frac{1}{n}\sum_{i=1}^{n} Z_i\rangle\Big) \oplus \xi \pmod{2\pi}, \tag{24}$$

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 $[0, 2\pi)$.

For this example, the generalization error of W can be computed exactly as follows:

$$\mathrm{gen}(\mu, P_{W|S}) = \mathbb{E}_{W,S}\Big\|\frac{1}{n}\sum_{i=1}^{n} Z_i\Big\|_2 = \sqrt{\frac{\pi}{2n}}, \tag{25}$$

where the last step is due to the fact that the distribution of $\|\frac{1}{n}\sum_{i=1}^{n} Z_i\|_2$ is Rayleigh$(\frac{1}{n})$. For the second algorithm W', since the noise ξ is independent from S, it follows that

$$\mathrm{gen}(\mu, P_{W'|S}) = \epsilon\sqrt{\frac{\pi}{2n}}. \tag{26}$$

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',

$$I(W';S) = h(W') - h(W'|S) = \log 2\pi - h(\xi) = \infty, \tag{27}$$

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

$$\begin{aligned} I(W;Z_i) &= h(W) - h(W|Z_i) = \log 2\pi - h(W|Z_i) \\ &= \log 2\pi - \mathbb{E}_{Z_i}[h(W|Z_i = z_i)]. \end{aligned} \tag{28}$$

Note that given $Z_i = z_i$, the ERM solution is

$$W = \arg\max_{\phi\in[0,2\pi)} \langle w, \frac{z_i}{n} + \frac{1}{n}\sum_{j\neq i} Z_i\rangle, \tag{29}$$

which 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 variable $\mathcal{N}(\frac{z_i}{n}, \frac{n-1}{n^2}I_2)$ 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 Euclidian norm of z_i. Then, $P_{W|Z_i=z_i}$ is a function of r, and can be equivalently characterized by

$$\begin{aligned} f\big(\phi\big|\|Z_i\| = r\big) =& \frac{1}{2\pi}e^{-\frac{r^2}{2(n-1)}} \\ &+ \frac{r\cos\phi}{\sqrt{2\pi(n-1)}}e^{-\frac{r^2\sin^2\phi}{2(n-1)}}Q(-\frac{r\cos\phi}{n-1}), \end{aligned} \tag{30}$$

where $Q(x)$ is the tail distribution function of the standard normal distribution. Since the norm of Z_i has a Rayleigh distribution with unit variance, it then follows that

$$I(W;Z_i) = \log 2\pi - \mathbb{E}_{\|Z_i\|}\Big[h\big(f(\phi\big|\|Z_i\| = r)\big)\Big]. \tag{31}$$

Applying Theorem 2, we obtain

$$|\mathrm{gen}(\mu, P_{W|S})| \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2I(W;Z_i)} = \sqrt{2I(W;Z_i)}. \tag{32}$$

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 [5], 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.

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

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 updating rule is as follows [10]:

$$W_{(t)} = W_{(t-1)} - \eta_{(t)} \nabla \ell(W_{(t-1)}, Z_{U_{(t)}}) + \sigma_{(t)} \xi, \qquad (33)$$

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. For the k-th training epoch, i.e., from $((k-1)n+1)$-th to kn-th iterations, all training samples in S are used exactly once. 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 [11].

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

Lemma 3. *[11, Corollary 1] The generalization error of the SGLD algorithm is bounded by*

$$|\mathrm{gen}(\mu, P_{W|S})| \leq \sqrt{\frac{R^2}{n} \sum_{t=1}^{T} \frac{\eta_t^2 L^2}{\sigma_t^2}}. \qquad (34)$$

B. ISMI Bound for SGLD

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

$$\begin{aligned} &|\mathrm{gen}(\mu, P_{W|S})| \\ &= \Big| \mathbb{E}_{U^{(T)}} \Big[\frac{1}{n} \sum_{i=1}^{n} \Big(\mathbb{E}_{W, \widetilde{Z}}[\ell(W, \widetilde{Z})] - \mathbb{E}_{W, Z_i}[\ell(W, Z_i) | U^{(T)}] \Big) \Big] \Big| \\ &\leq \frac{1}{|\mathcal{U}|} \sum_{u^{(T)} \in \mathcal{U}} \Big(\frac{1}{n} \sum_{i=1}^{n} \sqrt{2R^2 I(W; Z_i | U^{(T)} = u^{(T)})} \Big), \qquad (35) \end{aligned}$$

where $\mathcal{U}$ denotes the set of all possible sample paths.

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

$$\begin{aligned} &I(W; Z_i | U^{(T)} = u^{(T)}) \\ &\leq I(Z_i; W^{(T)} | U^{(T)} = u^{(T)}) \\ &= \sum_{\tau=1}^{T} I(Z_i; W_{(\tau)} | W_{(\tau-1)}, U^{(T)} = u^{(T)}) \\ &= \sum_{\tau \in \mathcal{T}_i(u^{(T)})} I(Z_i; W_{(\tau)} | W_{(\tau-1)}, U^{(T)} = u^{(T)}), \qquad (36) \end{aligned}$$

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 $\tau \in \mathcal{T}_i(B^{(T)})$, i.e., if Z_i is selected at iteration τ, we have

$$\begin{aligned} &I(Z_i; W_{(\tau)} | W_{(\tau-1)}, U^{(T)} = u^{(T)}) \\ &= h\big(\eta_{(\tau)} \nabla \ell(W_{(\tau-1)}, Z_i) + \sigma_{(\tau)} \xi | W_{(\tau-1)}\big) - h(\sigma_{(\tau)} \xi) \\ &\leq \frac{d}{2} \log\big(1 + \frac{\eta_{(\tau)}^2 L^2}{d \sigma_{(\tau)}^2}\big), \qquad (37) \end{aligned}$$

where the last step follows from Assumption 2 and the fact that ξ is an independent Gaussian noise as in [11].

Combining with (35), it follows that

$$|\mathrm{gen}(\mu, P_{W|S})| \leq \mathbb{E}_{U^{(T)}} \left[\frac{R}{n} \sum_{i=1}^{n} \sqrt{\sum_{\tau \in \mathcal{T}_i(U^{(T)})} \frac{\eta_{(\tau)}^2 L^2}{\sigma_{(\tau)}^2}} \right], \qquad (38)$$

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

C. Discussion

As in [11], we set $\eta_{(t)} = \frac{c}{t}$, and $\sigma_{(t)} = \sqrt{\eta_t}$. Then,

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

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 [11],

$$|\mathrm{gen}(\mu, P_{W|S})| \leq \frac{RL}{\sqrt{n}} \sqrt{c \log(nK) + c}, \qquad (39)$$

it can be seen that our bound is tighter by a factor of $\sqrt{\log n}$.

References

[1] 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.

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

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

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

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

[6] 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.

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

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

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

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

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