Individually Conditional Individual Mutual Information Bound on Generalization Error

Ruida Zhou, Chao Tian, and Tie Liu
Department of Electrical and Computer Engineering
Texas A&M University
Email: {ruida, chao.tian, tieliu}@tamu.edu

2021 IEEE International Symposium on Information Theory (ISIT) | 978-1-5386-8209-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISIT45174.2021.9518016

***Abstract*—We propose a new information-theoretic bound on generalization error based on a combination of the error decomposition technique of Bu et al. and the conditional mutual information (CMI) construction of Steinke and Zakynthinou. In a previous work, Haghifam et al. proposed a different bound combining the two aforementioned techniques, which we refer to as the conditional individual mutual information (CIMI) bound. However, in a simple Gaussian setting, both the CMI and the CIMI bounds are order-wise worse than that by Bu et al.. This observation motivated us to propose the new bound, which overcomes this issue by reducing the conditioning terms in the conditional mutual information. In the process of establishing this bound, a conditional decoupling lemma is established, which also leads to a meaningful dichotomy and comparison among these information-theoretic bounds.**

I. Introduction

Bounding the generalization error of learning algorithms is of fundamental importance in statistical machine learning. The conventional approach is to bound it using a quantity related to the hypothesis class, such as the VC-dimension [1], and such bounds are therefore oblivious to the learning algorithm and data distribution. The obtained results are usually rather conservative, and cannot fully explain the recent success of deep learning. Recently, information theoretic approaches that jointly take into consideration the hypothesis class, the learning algorithm, and the data distribution, has drawn considerable attention [2]–[13].

The effort of deriving generalization error bounds using information theoretic approaches was perhaps first initiated in [2] and [8]. The bound was further tightened in [9], by decomposing the error, and bounding each term individually. Steinke and Zakynthinou [10] proposed a conditional mutual information (CMI) based bound, by introducing a dependence structure which resembles that in the analysis of the Rademacher complexity [1]. Combining the idea of error decomposition [9] and the CMI bound in [10], Haghifam et al. [11] subsequently provided a sharpened bound based on conditional individual mutual information (CIMI).

In this work, we propose a new generalization error bound, which is also based on a combination of the error decomposition technique and the CMI construction. This new bound is motivated by the observation that in a simple Gaussian setting, the CIMI bound in [11] (as well as the CMI bound in [10]) is of constant order, while the bound in [9] is of order $\Theta(\frac{1}{\sqrt{n}})$, where n is the number of training samples. We further observe that the conditioning term in CIMI is the same as CMI, and it tends to reveal too much information which makes the bounds loose. The proposed new bound is thus obtained by making the mutual information conditioned on an individual sample (pair), which we refer to as the individually conditional individual mutual information (ICIMI) bound. In order to establish the new bound, we introduce a new conditional decoupling lemma. This lemma allows us to view the bounds in [8]–[11] and the new bound in a unified manner, which not only yields a dichotomy of these bounds, but also makes possible a meaningful comparison among them. Finally, we show that in the Gaussian setting mentioned earlier, the proposed new bound is also able to provide a bound of the same order as, but with an improved leading constant than, that in [9].

After our initial preprint was posted on Arxiv, we were made aware of an independent work by Rodríguez-Gálvez et al. [14], where a similar ICIMI-based generalization bound was proposed under the restricted assumption of bounded loss. In contrast, our result applies under more general conditions. Our work was mainly motivated by the looseness of the CIMI bound in the Gaussian setting, for which the restricted assumption in [14] makes their result not applicable. Furthermore, the proposed conditional decoupling lemma, which we believe is of fundamental importance, was not present in [14].

II. Preliminary

We study the classic supervised learning setting. Denote the data domain as $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the feature domain and $\mathcal{Y}$ is the label set. The parametric hypothesis class is denoted as $\mathcal{H}_{\mathcal{W}} = \{h_W : W \in \mathcal{W}\} \subseteq \mathcal{Y}^{\mathcal{X}}$, where $\mathcal{W}$ is the parameter space. During the training, the learning algorithm (learner) has access to a sequence of training samples $Z_{[n]} = (Z_1, Z_2, \ldots, Z_n)$, where each Z_i is drawn independently from $\mathcal{Z}$ following some unknown probability distribution ξ. The learner can be represented by $P_{W|Z_{[n]}}$, which is a kernel (channel) that (randomly) maps $\mathcal{Z}^n$ to $\mathcal{W}$.

To complete the classification or regression task, the learner in principle would choose a hypothesis $w \in \mathcal{W}$ to minimize the following population loss, under a given loss function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$,

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

The work of T. Liu was supported in part by the National Science Foundation under Grant CCF-1719017.

However, since only a training data vector $Z_{[n]}$ is available, the empirical loss of w is usually computed (and minimized during training), which is given as

$$L_{Z_{[n]}}(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, Z_i). \tag{2}$$

The expected generalization error of the learner $P_{W|Z_{[n]}}$ is

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) := \mathbb{E}\left[L_\xi(W) - L_{Z_{[n]}}(W)\right], \tag{3}$$

where the expectation is taken over the joint distribution $P(W, Z_{[n]}) = \xi^n \otimes P_{W|Z_{[n]}}$. This quantity captures the effect of the learner's expected overfitting error due to limited training data, which we shall study in this work.

III. Review of Related Results

In this section, we briefly review a few information theoretic bounds on the generalization error relevant to this work. A more thorough discussion of their relation is deferred to Section IV-D and IV-E, after a unified framework is given.

A. Mutual information based bounds

Xu and Raginsky, motivated by a previous work by Russo and Zou [2], provided a mutual information (MI) based bound on the expected generalization error [8].

Theorem 1 (MI Bound [8])**.** *Suppose $\ell(w, Z)$ is σ^2-sub-Gaussian under ξ for all $w \in \mathcal{W}$, then*

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) \le \sqrt{\frac{2\sigma^2}{n} I\left(W; Z_{[n]}\right)}. \tag{4}$$

The generalization can be written in two ways

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) = \mathbb{E}\left[L_{\tilde{Z}_{[n]}}(\tilde{W})\right] - \mathbb{E}\left[L_{Z_{[n]}}(W)\right] \tag{5}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\left(\ell(\tilde{W}, \tilde{Z}_i) - \ell(W, Z_i)\right)\right], \tag{6}$$

where $\tilde{W}$ and $\tilde{Z}_i$ are independent random variables that have the same marginal distributions as W and Z_i, respectively. Instead of bounding the difference (5) as in [8], Bu et al. [9] bounded each individual difference in (6) and derived an individual mutual information (IMI) based bound. Furthermore, the following inverse Fenchel conjugate function was utilized to obtain a tightened bound. For any random variables F, its cumulant generating function is

$$\psi_F(\lambda) := \ln \mathbb{E}\left[e^{\lambda(F - \mathbb{E}[F])}\right], \tag{7}$$

and the inverse of its Fenchel conjugate is given as

$$\psi_F^{*-1}(\eta) := \inf_{\lambda > 0} \frac{\eta + \psi_F(\lambda)}{\lambda}, \quad \eta \in [0, \infty). \tag{8}$$

The tightened bound is summarized in the following theorem.

Theorem 2 (IMI Bound [9])**.** *Suppose ψ_- is an upper bound of $\psi_{-\ell(\tilde{W}, \tilde{Z}_i)}$, then*

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) \le \frac{1}{n}\sum_{i=1}^{n} \psi_-^{*-1}\left(I\left(W; Z_i\right)\right), \tag{9}$$

where $\tilde{W}$ and $\tilde{Z}_i$ are independent random variables that have the same marginal distributions as W and Z_i, respectively.

B. Conditional mutual information based bounds

Steinke and Zakynthinou [10] recently introduced a novel bounding approach. In their approach, $Z_{[n]}^{\pm} := (Z_1^{\pm 1}, Z_2^{\pm 1}, \ldots, Z_n^{\pm 1})$ is a $2 \times n$ table of samples that each Z_i^s, for $s = -1, 1$ and $i = 1, \ldots, n$ is independently drawn following ξ. The training vector $(Z_1^{R_1}, Z_2^{R_2}, \ldots, Z_n^{R_n})$ is selected from the table $Z_{[n]}^{\pm}$, where R_i's are independent Rademacher random variables, i.e., R_i takes 1 or -1 equally likely. The vector $R_{[n]} = (R_1, \ldots, R_n) \in \{-1, 1\}^n$ essentially selects one sample from each column in the table, which partition $Z_{[n]}^{\pm}$ into a training vector and a testing vector. For simplicity, we shall write Z_i^{-1} and Z_i^{+1} as Z_i^- and Z_i^+, when the meaning is clear from the context.

With the structure given above, the expected generalization error of the algorithm can be written as

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) = \mathbb{E}_{Z_{[n]}^{\pm}}\left[\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} R_i\left(\ell(W, Z_i^-) - \ell(W, Z_i^+)\right)\Big| Z_{[n]}^{\pm}\right]\right]. \tag{10}$$

Steinke and Zakynthinou obtained the following conditional mutual information (CMI) based result.

Theorem 3 (CMI Bound [10])**.** *Suppose $\sup_{w \in \mathcal{W}} |\ell(w, z_1) - \ell(w, z_2)| \le \Delta(z_1, z_2)$ for any $z_1, z_2 \in \mathcal{Z}$, then*

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) \le \sqrt{\frac{2}{n}\mathbb{E}[\Delta(Z_1, Z_2)^2] I\left(W; R_{[n]} | Z_{[n]}^{\pm}\right)}, \tag{11}$$

where Z_1, Z_2 are independent samples distributed as ξ.

Since R_i is binary, the conditional mutual information is always bounded; in contrast, mutual information based bounds (i.e., MI and IMI bounds) can be unbounded, particularly when the random variables W, Z_i are both continuous.

Motivated by the results in [9], Haghifam et al. [11] proposed a sharpened bound by similarly bounding each term in (10). Moreover, they provided a conditional individual mutual information (CIMI) based bound represented by *the sample-conditioned mutual information*, which is defined as

$$I_u(X; Y) := I(X; Y | U = u). \tag{12}$$

Clearly $I_U(X; Y)$ is a function of the random variable U, thus also a random variable, and $\mathbb{E}[I_U(X; Y)] = I(X; Y|U)$. These sharpened bounds are summarized in the following theorem.

Theorem 4 (CIMI Bound [11])**.** *Suppose $\ell \in [0, 1]$, then*

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) \le \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\sqrt{2 I_{Z_{[n]}^{\pm}}(W; R_i)}\right] \tag{13}$$

$$\le \frac{1}{n}\sum_{i=1}^{n} \sqrt{2 I\left(W; R_i | Z_{[n]}^{\pm}\right)}. \tag{14}$$

IV. New Result

A. A motivating example

Let us consider the simple setting of estimating the mean from samples generated from a Gaussian distribution $N(\mu, \sigma^2)$, by averaging the $i.i.d.$ training samples under the squared loss.

Example 1 (Estimating the Gaussian mean)**.** *The training samples $Z_{[n]}$ are drawn i.i.d. following $N(\mu, \sigma^2)$ for some unknown μ. The learner deterministically estimates μ by averaging the training samples, i.e., $W = \frac{1}{n}\sum_{i=1}^{n} Z_i$, whose empirical error is*

$$L_{Z_{[n]}}(W) = \frac{1}{n}\sum_{i=1}^{n}(W - Z_i)^2. \tag{15}$$

Bu et al. [9] showed that the mutual information term in the IMI bound is

$$I(W; Z_i) = \frac{1}{2}\log\frac{n}{n-1} = \frac{1}{2(n-1)} + o\left(\frac{1}{n}\right), \tag{16}$$

and obtained the following IMI based bound

$$\sigma^2\sqrt{\frac{2(n+1)^2}{n^2}\log\frac{n}{n-1}} = \sigma^2\sqrt{\frac{2}{n-1}} + o\left(\frac{1}{\sqrt{n}}\right). \tag{17}$$

For this simple setting, the generalization error can in fact be calculated exactly to be $\frac{2\sigma^2}{n}$. Though the error bound above does not have the same order as the true generalization error, it is consistent with the VC dimension-based bound and is the best known for this case. Note that the MI bound will be unbounded, since $I(W; Z_{[n]})$ is unbounded.

Next consider the CMI and CIMI bounds, and let us focus on the mutual information terms in these bounds, which give

$$I(W; R_{[n]}|Z^{\pm}_{[n]}) = n/\log_2 e, \tag{18}$$

$$I_{Z^{\pm}_{[n]}}(W; R_i) = 1/\log_2 e, \quad a.s.. \tag{19}$$

It is seen that they are order-wise worse than (16), which suggests that the bounds obtained from the CMI and CIMI bounds would be order-wise worse than (17).

Theorem 3 and Theorem 4 in fact do not apply directly in this setting, since their required conditions do not hold. In Theorem 3, the function $\Delta(z_1, z_2)$ does not exist (i.e., unbounded); even if it existed, the term $\mathbb{E}[\Delta(Z_1, Z_2)^2]$ would be a constant, thus the CMI bound would be of constant order. Similarly, if the condition $\ell \in [0, 1]$ held, the CIMI bound would also be of constant order. As we shall show shortly, the CMI and CIMI bounds can be generalized and strengthened, yet the resultant strengthened bounds in this setting still do not diminish as $n \to \infty$, and thus would be order-wise worse than the IMI bound.

A question arises naturally: Is the looseness of the CMI and CIMI bounds here due to the introduction of the conditioning terms? As we shall show next, it is in fact caused by too much information being revealed in the conditioning terms, and there is indeed a natural way to resolve this issue.

B. A conditional decoupling lemma

Our main result relies on a key lemma. A few more definitions are first introduced in order to present this lemma and the main result.

For any random variables F and U, define the *sample-conditioned cumulant generating function (CGF)* for any realization $U = u$,

$$\Lambda_{F|U}(\lambda, u) := \ln \mathbb{E}\left[e^{\lambda F}\middle| U = u\right], \quad \lambda \in \mathbb{R}. \tag{20}$$

Similar to the regular CGF, $\Lambda_{F|U}(\lambda, u)$ may not exist for some $\lambda \in \mathbb{R}$. Define the *extended-value centered sample-conditioned CGF* as $\psi_{F|U}(\lambda, u) := \infty$ for such λ that $\Lambda_{F|U}(\lambda, u)$ does not exist, and $\psi_{F|U}(\lambda, u) := \Lambda_{F|U}(\lambda, u) - \lambda\mathbb{E}[F|U = u]$ otherwise. It is straightforward to verify that for any realization $U = u$, $\psi_{F|U}(0, u) = \psi'_{F|U}(0, u) = 0$ and $\psi''_{F|U}(0, u) > 0$. Hence the inverse of its Fenchel conjugate

$$\psi^{*-1}_{F|U}(\eta, u) := \inf_{\lambda > 0}\frac{\eta + \psi_{F|U}(\lambda, u)}{\lambda}, \quad \eta \in [0, \infty) \tag{21}$$

is concave and non-decreasing; see e.g., [9] and [15]. The unconditioned version of this function was introduced earlier by Asadi et al. [3] and Bu et al. [9]. When it is clear from context, we will write

$$\Psi_{F|U}(\lambda) := \psi_{F|U}(\lambda, U), \quad \Psi^{*-1}_{F|U}(\eta) := \psi^{*-1}_{F|U}(\eta, U), \tag{22}$$

which are functions of U, thus random. Next define *the conditional cumulant generating function*

$$\bar{\psi}_{F|U} = \mathbb{E}\left[\Psi_{F|U}\right], \tag{23}$$

and similarly its inverse Fenchel conjugate as $\bar{\psi}^{*-1}_{F|U}$.

For a pair of random variables (X, Y), its *decoupled pair conditioned on a third random variable U* is a pair of random variables $(\tilde{X}, \tilde{Y})$, such that

$$(\tilde{X}, U) \overset{D}{=} (X, U), \quad (\tilde{Y}, U) \overset{D}{=} (Y, U), \tag{24}$$

i.e., $(\tilde{X}, U)$ and (X, U) are identically distributed, and $(\tilde{Y}, U)$ and (Y, U) are identically distributed, and moreover

$$\tilde{X} \leftrightarrow U \leftrightarrow \tilde{Y} \tag{25}$$

forms a Markov string. It follows from this definition that

$$I_U(X; Y) = D(P_{X,Y|U}||P_{\tilde{X},\tilde{Y}|U}). \tag{26}$$

We next introduce a conditional decoupling (CD) lemma, which serves an instrumental role in our work. The unconditioned version was presented in [9].

Lemma 1 (The CD lemma)**.** *For any three random variables X, Y, U, let $\tilde{X}, \tilde{Y}$ be the decoupled pair of X, Y conditioned on U. Let $F := f(X, Y)$ and $\tilde{F} := f(\tilde{X}, \tilde{Y})$, for some real-valued measurable function f. The following inequalities hold*

$$\begin{aligned}\mathbb{E}[F] - \mathbb{E}[\tilde{F}] &\le \mathbb{E}\left[\Psi^{*-1}_{\tilde{F}|U}\left(I_U(X; Y)\right)\right] \\ &\le \bar{\psi}^{*-1}_{\tilde{F}|U}\left(I(X; Y|U)\right).\end{aligned} \tag{27}$$

This lemma is proved by utilizing the Donsker–Varadhan variational representation of KL divergence and the concavity of the inverse Fenchel conjugate function. The proof details can be found in [16].

C. The ICIMI bound

Let $(W, Z^{\pm}_{[n]}, R_{[n]})$ be as given previously in Section III-B. For each $i = 1, \ldots, n$, let $(\tilde{W}_i, \tilde{R}_i)$ be a decoupled pair of (W, R_i) conditioned on $Z^{\pm}_i$. The new bound we propose is presented in Theorem 5.

Theorem 5. *(ICIMI Bound) Given an algorithm $P_{W|Z_{[n]}}$, the following bounds on the generalization hold*

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\Psi^{*-1}_{\tilde{G}_i|Z^{\pm}_i}(I_{Z^{\pm}_i}(W; R_i))\right] \tag{28}$$

$$\leq \frac{1}{n}\sum_{i=1}^{n} \bar{\psi}^{*-1}_{\tilde{G}_i|Z_i^{\pm}}(I(W;R_i|Z_i^{\pm})), \tag{29}$$

where $\tilde{G}_i = \tilde{R}_i\left(\ell(\tilde{W}_i, Z_i^-) - \ell(\tilde{W}_i, Z_i^+)\right)$.

There are two bounds in this theorem. The stronger bound is in terms of the sample-conditioned mutual information, which is different from the conventional notion of conditional mutual information and may be more difficult to evaluate. The weaker bound is in terms of the conventional mutual information.

In the proposed bounds, the mutual information is conditioned on the individual data pair $Z_i^{\pm}$, instead of the full data pair set $Z_{[n]}^{\pm}$. Intuitively, revealing only $Z_i^{\pm}$ makes it more difficult, than revealing all data pairs $Z_{[n]}^{\pm}$, to deduce information regarding R_i from W. As a consequence, the mutual information $I(W;R_i|Z_i^{\pm})$ is less than $I(W;R_i|Z_{[n]}^{\pm})$, yielding a potentially tighter bound.

Proof of Theorem 5. We can rewrite the generalization error given in (10) as

$$\mathrm{gen}(\xi, P_{W|Z_{[n]}}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}\left[R_i\left(\ell(W,Z_i^-) - \ell(W,Z_i^+)\right)|Z_i^{\pm}\right]\right]. \tag{30}$$

Now apply the CD lemma on each individual term in (30) by letting $X = W$, $Y_i = R_i$, $U_i = Z_i^{\pm}$, and $F_i = R_i\left(\ell(W,Z_i^-) - \ell(W,Z_i^+)\right)$. Since

$$\mathbb{E}[\tilde{G}_i] = \mathbb{E}[\tilde{F}_i] = \mathbb{E}\left[\tilde{R}_i\left(\ell(\tilde{W}_i,Z_i^-) - \ell(\tilde{W}_i,Z_i^+)\right)\right] = 0,$$

we have

$$\begin{aligned}\mathrm{gen}(\xi, P_{W|Z_{[n]}}) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[F_i] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[F_i] - \mathbb{E}[\tilde{F}_i] \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\Psi^{*-1}_{\tilde{G}_i|Z_i^{\pm}}(I_{Z_i^{\pm}}(W;R_i))\right] \quad (31)\\ &\leq \frac{1}{n}\sum_{i=1}^{n}\bar{\psi}^{*-1}_{\tilde{G}_i|Z_i^{\pm}}(I(W;R_i|Z_i^{\pm})), \quad (32)\end{aligned}$$

which completes the proof. □

We call this bound the individually conditional individual mutual information (ICIMI) bound, since it is derived by applying the CD lemma on the individual conditional terms in (30).

We note that Theorem 5 implies Proposition 3 in [14], which we state below as a corollary.

Corollary 1. *Suppose* $\ell \in [a,b]$ *with* $a < b$*, then*

$$\begin{aligned}\mathrm{gen}(\xi, P_{W|Z_{[n]}}) &\leq \frac{b-a}{n}\sum_{i=1}^{n}\mathbb{E}_{Z_{[n]}^{\pm}}\left[\sqrt{2I_{Z_i^{\pm}}(W;R_i)}\right] \quad (33)\\ &\leq \frac{b-a}{n}\sum_{i=1}^{n}\sqrt{2I(W;R_i|Z_i^{\pm})}. \quad (34)\end{aligned}$$

Proof of Corollary 1. When $\ell \in [a,b]$ and $\tilde{F}_i \in [a-b, b-a]$, it is straightforward to verify that $\tilde{F}_i$ is $\frac{(b-a)^2}{2}$-sub-Gaussian. The definition of the sub-Gaussian distribution in fact gives $\Psi_{\tilde{F}_i|Z_i^{\pm}}(\lambda) \leq \frac{(b-a)^2}{2}\lambda^2$, and thus $\Psi^{*-1}_{\tilde{F}_i|Z_i^{\pm}}(\eta) \leq (b-a)\sqrt{2\eta}$, from which the corollary follows. □

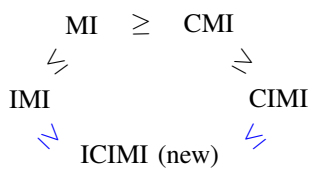

Fig. 1. Relations among generalization bounds, when the inverse Fenchel conjugate functions are assumed to be the same.

D. Dichotomy and generalizations of existing bounds

The CD lemma allows us to view the existing MI, IMI, CMI, and CIMI bounds in a unified framework. By applying the CD lemma in different manners, these bounds can be obtained almost directly. The technical conditions under which the bound hold can also be generalized, and the bounds themselves can be strengthened using the inverse Fenchel conjugate. These results are summarized in Table I. We also provide the bounds for bounded loss function, which eliminate the $\bar{\psi}^{*-1}$ functions.

The CMI and CIMI results can be further strengthened by utilizing the inverse Fenchel conjugate function together with the sample-conditioned mutual information. More precisely, let $(\tilde{R}_{[n]}, \tilde{W})$ be the decoupled pair of $(R_{[n]}, W)$ conditioned on $Z_{[n]}^{\pm}$. Further define

$$\tilde{E}_i = \tilde{R}_i\left(\ell(\tilde{W},Z_i^-) - \ell(\tilde{W},Z_i^+)\right), \quad \tilde{E} = \frac{1}{n}\sum_{i=1}^{n}\tilde{E}_i, \tag{35}$$

then we have the strengthened CMI and CIMI bounds:

$$\mathrm{gen}\left(\xi, P_{W|Z_{[n]}}\right) \leq \mathbb{E}\left[\Psi^{*-1}_{\tilde{E}|Z_{[n]}^{\pm}}\left(I_{Z_{[n]}^{\pm}}(W;R_{[n]})\right)\right], \tag{36}$$

$$\mathrm{gen}\left(\xi, P_{W|Z_{[n]}}\right) \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\Psi^{*-1}_{\tilde{E}_i|Z_{[n]}^{\pm}}\left(I_{Z_{[n]}^{\pm}}(W;R_i)\right)\right]. \tag{37}$$

E. Comparison of the bounds

We first consider the special case where the loss function is bounded, i.e., $\ell \in [0,1]$. For this case, it was shown in [11] that the CIMI bound (14) is tighter than the CMI bound (11). We next show that the proposed bound (34) is tighter than the CIMI bound (14) when $\ell \in [0,1]$.

Lemma 2. *For any* $i = 1, \ldots, n$*, we have*

$$I(W;R_i|Z_i^{\pm}) \leq I(W;R_i|Z_{[n]}^{\pm}).$$

Proof of Lemma 2. By the independence of R_i and $Z_{[n]}^{\pm}$, we have

$$\begin{aligned}I(W;R_i|Z_{[n]}^{\pm}) &= H(R_i) - H(R_i|W, Z_{[n]}^{\pm}),\\ I(W;R_i|Z_i^{\pm}) &= H(R_i) - H(R_i|W, Z_i^{\pm}).\end{aligned}$$

It follows that

$$I(W;R_i|Z_{[n]}^{\pm}) - I(W;R_i|Z_i^{\pm}) = I(R_i; Z_{[n]}^{\pm}|W, Z_i^{\pm}) \geq 0,$$

which concludes the proof. □

To further understand the relation among these bounds under more general conditions when the loss function may

TABLE I
A DICHOTOMY OF SEVERAL GENERALIZATION BOUNDS USING THE CD LEMMA

Approach	X	Y	U	F	Generalization bound	Special case $\ell \in [0,1]$
MI [8]	W	$Z_{[n]}$		$-\frac{1}{n}\sum_{i=1}^{n}\ell(W,Z_i)+L_\xi(W)$	$\bar{\psi}^{*-1}_{\tilde{F}}\left(I\left(W;Z_{[n]}\right)\right)$	$\sqrt{\frac{1}{2n}I(W;Z_{[n]})}$
IMI [9]	W	Z_i		$F_i=-\ell(W,Z_i)$	$\frac{1}{n}\sum_{i=1}^{n}\bar{\psi}^{*-1}_{\tilde{F}_i}\left(I\left(W;Z_i\right)\right)$	$\frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{1}{2}I(W;Z_i)}$
CMI [10]	W	$R_{[n]}$	$Z^{\pm}_{[n]}$	$\frac{1}{n}\sum_{i=1}^{n}R_i\left(\ell(W,Z_i^-)-\ell(W,Z_i^+)\right)$	$\bar{\psi}^{*-1}_{\tilde{F}|Z^{\pm}_{[n]}}\left(I\left(W;R_{[n]}|Z^{\pm}_{[n]}\right)\right)$	$\sqrt{2I(W;R_{[n]}\|Z^{\pm}_{[n]})}$
CIMI [11]	W	R_i	$Z^{\pm}_{[n]}$	$F_i=R_i\left(\ell(W,Z_i^-)-\ell(W,Z_i^+)\right)$	$\frac{1}{n}\sum_{i=1}^{n}\bar{\psi}^{*-1}_{\tilde{F}_i|Z^{\pm}_{[n]}}\left(I\left(W;R_i|Z^{\pm}_{[n]}\right)\right)$	$\frac{1}{n}\sum_{i=1}^{n}\sqrt{2I(W;R_i\|Z^{\pm}_{[n]})}$
ICIMI (new)	W	R_i	$Z^{\pm}_i$	$F_i=R_i\left(\ell(W,Z_i^-)-\ell(W,Z_i^+)\right)$	$\frac{1}{n}\sum_{i=1}^{n}\bar{\psi}^{*-1}_{\tilde{F}_i|Z^{\pm}_i}\left(I\left(W;R_i|Z^{\pm}_i\right)\right)$	$\frac{1}{n}\sum_{i=1}^{n}\sqrt{2I(W;R_i\|Z^{\pm}_i)}$

not be bounded, let us assume the inverse Fenchel conjugate functions, which roughly capture the geometry induced by the expected loss, are the same (denoted as $\bar{\psi}^{*-1}$) for all the five approaches, i.e.,

$$\bar{\psi}^{*-1}=\bar{\psi}^{*-1}_{-\tilde{F}}=\bar{\psi}^{*-1}_{-\tilde{F}_i}=\bar{\psi}^{*-1}_{\tilde{F}|Z^{\pm}_{[n]}}=\bar{\psi}^{*-1}_{\tilde{F}_i|Z^{\pm}_{[n]}}=\bar{\psi}^{*-1}_{\tilde{F}_i|Z^{\pm}_i}.$$

Then we can focus on the information measure quantities, and compare these bounds as shown in Fig. 1. Here the inequalities given in black were proved previously (see [9] and [11]). Since the common function $\bar{\psi}^{*-1}$ is non-decreasing, the inequality "CIMI $\geq$ ICIMI" follows from Lemma 2. The inequality "IMI $\geq$ ICIMI" is implied by the following lemma for the same reason.

Lemma 3. *For any $i=1,\ldots,n$, we have*

$$I(W;R_i|Z_i^{\pm})\leq I(W;Z_i).$$

Proof of Lemma 3. First Z_i and $Z_i^{R_i}$ are both the i^{th} training sample for the input of the algorithm, thus

$$I(W;Z_i)=I(W;Z_i^{R_i}). \tag{38}$$

Then since $Z_i^{-R_i}, R_i$ and W are independent given $Z_i^{R_i}$,

$$I(W;Z_i^{\pm},R_i)=I(W;Z_i^{R_i},Z_i^{-R_i},R_i) \tag{39}$$

$$=I(W;Z_i^{R_i})+I(W;Z_i^{-R_i},R_i|Z_i^{R_i})=I(W;Z_i^{R_i}). \tag{40}$$

It follows that

$$I(W;Z_i)=I(W;Z_i^{\pm},R_i)\geq I(W;R_i|Z_i^{\pm}), \tag{41}$$

which concludes the proof. □

The inverse Fenchel conjugate functions may indeed be different for different bounds, thus although the above comparison suggests certain dominant relations, it is not clear for any specific problem, whether any particular bound is tighter than the other. This is particularly true if we use the bounds based on the inverse Fenchel conjugate, however, even for the special case of $\ell\in[0,1]$, the different multiplicative factors and the sum-square-root forms imply that the relation can be less clear.

F. Revisiting the example

We now return to the problem of estimating the Gaussian mean, and show that the proposed ICIMI bound can provide scaling behavior similar to that of IMI, thus order-wise stronger than the CMI and CIMI bounds. In fact, the bound is also strictly better than the IMI bound given in [9] asymptotically in this setting.

We first formally establish, as suspected previously, that the CMI and CIMI bounds are at least of constant order for this setting.

Proposition 1. *The strengthened CMI and CIMI bounds, i.e., (36) and (37), are at least $\frac{\sigma^2}{\pi\sqrt{\log e}}$ in the problem of estimating the Gaussian mean.*

The proof of this proposition can be found in [16]. The next proposition establishes a generalization error bound based on the ICIMI bound in this setting.

Proposition 2. *For the the problem of estimating the mean of the Gaussian distribution, the ICIMI bound gives*

$$\mathrm{gen}\left(\xi,P_{W|Z_{[n]}}\right)\leq\frac{2\sigma^2}{\sqrt{\pi}}\sqrt{\frac{1}{n-1}}+o\left(\frac{1}{\sqrt{n}}\right). \tag{42}$$

Remark: This bound scales as $\Theta(\sqrt{\frac{1}{n}})$. Compared to the IMI bound in (17), the new ICIMI based bound is asymptotically tighter by a factor of $\sqrt{\frac{\pi}{2}}\approx 1.25$.

Proposition 2 is proved by studying separately the sample-conditioned individual mutual information $I_{Z_i^{\pm}}(W;R_i)$ and the inverse Fenchel conjugate functions $\Psi^{*-1}_{\tilde{G}_i|Z^{\pm}}$. For the former, since the algorithm here is averaging the samples without any prior of the Gaussian distribution, without loss of generality, we can assume the mean of the Gaussian distribution to be 0, i.e., $\mu=0$. Therefore, given $Z_i^{\pm}=z_{\pm}\in\mathbb{R}^2$, W is mixed-Gaussian distributed, which follows $N(\frac{z_+}{n},\frac{n-1}{n^2}\sigma^2)$ when $R_i=1$ and follows $N(\frac{z_-}{n},\frac{n-1}{n^2}\sigma^2)$ when $R_i=-1$. The term $I_{Z_i^{\pm}}(W;R_i)$ is thus related to the scaling behavior of the differential entropy of a mixed Gaussian distribution. The proof of the proposition relies on a detailed analysis of this behavior, which can be found in [16].

V. CONCLUSION

We propose a new information theoretic generalization error bound, referred to as the ICIMI bound, based on a combination of the error decomposition technique and the conditional mutual information structure. Due to the reduced information content in the conditioning term, the proposed bound can be significantly tighter than several existing bounds. Particularly, when the loss function is bounded, it can be shown that the proposed bound is always tighter than the CMI and the CIMI bounds. A conditional decoupling lemma is provided which leads to a unified framework to study and compare these bounds, and it may be of independent interest.

REFERENCES

[1] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[2] D. Russo and J. Zou, "Controlling bias in adaptive data analysis using information theory," in *Artificial Intelligence and Statistics*, 2016, pp. 1232–1240.

[3] A. Asadi, E. Abbe, and S. Verdú, "Chaining mutual information and tightening generalization bounds," in *Advances in Neural Information Processing Systems*, 2018, pp. 7234–7243.

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

[5] 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*, 2019, pp. 11 015–11 025.

[6] S. T. Jose and O. Simeone, "Information-theoretic generalization bounds for meta-learning and applications," *arXiv preprint arXiv:2005.04372*, 2020.

[7] X. Wu, J. H. Manton, U. Aickelin, and J. Zhu, "Information-theoretic analysis for transfer learning," *arXiv preprint arXiv:2005.08697*, 2020.

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

[9] Y. Bu, S. Zou, and V. V. Veeravalli, "Tightening mutual information based bounds on generalization error," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 121–130, 2020.

[10] T. Steinke and L. Zakynthinou, "Reasoning about generalization via conditional mutual information," *arXiv preprint arXiv:2001.09122*, 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," *arXiv preprint arXiv:2004.12983*, 2020.

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

[13] F. Hellström and G. Durisi, "Generalization bounds via information density and conditional information density," in *IEEE Journal on Selected Areas in Information Theory*, 2020.

[14] B. Rodríguez-Gálvez, G. Bassi, R. Thobaben, and M. Skoglund, "On random subset generalization error bounds and the stochastic gradient Langevin dynamics algorithm," *arXiv preprint arXiv:2010.10994*, 2020.

[15] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[16] R. Zhou, C. Tian, and T. Liu, "Individually conditional individual mutual information bound on generalization error," *arXiv preprint arXiv:2012.09922*, 2020.