Tighter Expected Generalization Error Bounds via Convexity of Information Measures

Gholamali Aminian*†, Yuheng Bu*‡, Gregory W. Wornell‡, Miguel R. D. Rodrigues†,

†University College London, ‡Massachusetts Institute of Technology
Email: {g.aminian, m.rodrigues}@ucl.ac.uk, {buyuheng, gww}@mit.edu

Abstract—Generalization error bounds are essential to understanding machine learning algorithms. This paper presents novel expected generalization error upper bounds based on the average joint distribution between the output hypothesis and each input training sample. Multiple generalization error upper bounds based on different information measures are provided, including Wasserstein distance, total variation distance, KL divergence, and Jensen-Shannon divergence. Due to the convexity of the information measures, the proposed bounds in terms of Wasserstein distance and total variation distance are shown to be tighter than their counterparts based on individual samples in the literature. An example is provided to demonstrate the tightness of the proposed generalization error bounds.

I. INTRODUCTION

Machine learning algorithms are increasingly adopted to solve various problems in a wide range of applications. Understanding the generalization behavior of a learning algorithm is one of the most important challenges in statistical learning theory. Various approaches have been developed to bound the generalization error [1], including VC dimension-based bounds [2], algorithmic stability-based bounds [3], algorithmic robustness-based bounds [4], PAC-Bayesian bounds [5].

More recently, approaches leveraging information-theoretic tools have been developed to characterize the generalization error of a learning algorithm. Such approaches incorporate various ingredients associated with a supervised learning problem, including the data generating distribution, the hypothesis space, and the learning algorithm itself, expressing expected generalization error in terms of specific information measures between the input of training dataset and output hypothesis.

In particular, building upon pioneering work by Russo and Zou [6], an expected generalization error upper based on the mutual information between the training set and the hypothesis is proposed by Xu and Raginsky [7]. Bu *et al.* [8] have derived tighter generalization error bounds based on individual sample mutual information. The generalization error bounds based on other information measures such as α-Réyni divergence [9], maximal leakage [10], Jensen-Shannon divergence [11], Wasserstein distances [12], [13] and individual sample Wasserstein distance [14] are also considered. Chaining mutual information technique is proposed in [15] and [16] to further improve the mutual information-based bound. The upper bounds based on conditional mutual information and

individual sample conditional mutual information are proposed in [17] and [18], respectively. It is shown in [19], [20], the combination of conditioning and processing techniques could provide tighter generalization error upper bounds. Using rate-distortion theory, [21]–[23] provide information-theoretic generalization error upper bounds for model misspecification and model compression, respectively. An exact characterization of the generalization error for the Gibbs algorithm in terms of symmetrized KL information is provided in [24].

In this paper, we introduce the notion of *average joint distribution*, which is the average of the distribution between the output hypothesis and each training sample. We aspire to provide a more refined analysis of the generalization ability of randomized learning algorithms by representing the expected generalization error using the aforementioned average joint distribution. The merit of this representation is that it directly leads to some tighter generalization error upper bounds based on the convexity of the information measures, including Wasserstein distance and total variation distance. The proposed bound finds its application when the importance of each training sample is not the same in the learning algorithm, e.g., imbalanced classification or learning under noisy data samples.

More specifically, our contributions are as follows:

- We provide novel expected generalization error upper bounds based on the average joint distribution between the output hypothesis and each input training sample, in terms of Wasserstein distance, total variation distance, KL divergence, and Jensen-Shannon divergence.
- We offer an upper bound on the difference between the empirical risk of two learning algorithms using the KL divergence between average joint distributions.
- We construct a simple numerical example to demonstrate the improvement of the proposed upper bound based on the average joint distribution in comparison to individual sample mutual information bound [8].

Notations: A random variable is denoted by an upper-case letter (e.g., Z), its alphabet is denoted by the corresponding calligraphic letter (e.g., $\mathcal{Z}$), and the realization of the random variable is denoted with a lower-case letter (e.g., z). The probability distribution of the random variable Z is denoted by P_Z. The joint distribution of a pair of random variables (Z_1, Z_2) is denoted by P_{Z_1, Z_2}.

Information Measures: The differential entropy of a continuous probability measure P defined over space $\mathcal{Z}$ is given

by $h(P) \triangleq \int_{\mathcal{Z}} -dP \log(dP)$. If P and Q are probability measures defined over space $\mathcal{Z}$, and P is absolutely continuous with respect to Q, the Kullback-Leibler (KL) divergence between P and Q is given by $D(P\|Q) \triangleq \int_{\mathcal{Z}} \log\left(\frac{dP}{dQ}\right) dP$. The Donsker-Varadhan variational representation of the KL divergence is as follows [25],

$$D(P\|Q) = \sup_{g \in \mathcal{G}} \left\{ \mathbb{E}_P[g(Z)] - \log(\mathbb{E}_Q[e^{g(Z)}]) \right\}, \quad (1)$$

where the supremum is over all measurable functions, i.e., $\mathcal{G} = \{g : \mathcal{Z} \to \mathbb{R}, \text{ s.t. } \mathbb{E}_Q[e^{g(Z)}] < \infty\}$.

The Jensen-Shannon divergence [26] is defined as

$$D_{JS}(P\|Q) \triangleq \frac{D(P\|\frac{P+Q}{2})}{2} + \frac{D(Q\|\frac{P+Q}{2})}{2}, \quad (2)$$

and it can be verified that $D_{JS}(P\|Q) \leq \log(2)$.

The mutual information between two random variables X and Y is defined as the KL divergence between their joint distribution and the product of the marginals, i.e., $I(X;Y) \triangleq D(P_{X,Y}\|P_X \otimes P_Y)$. Similarly, the Lautum information introduced in [27] is defined as the KL divergence between the product of the marginals and the joint distribution, i.e., $L(X;Y) \triangleq D(P_X \otimes P_Y\|P_{X,Y})$.

The Wasserstein distance between P and Q is defined using a metric $\rho : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_0^+$, and it is given by:

$$\mathbb{W}(P,Q) = \inf_{\pi \in \Pi(P,Q)} \int_{\mathcal{Z} \times \mathcal{Z}} \rho(z,z') d\pi(z,z'), \quad (3)$$

where $\Pi(P,Q)$ is the set of all joint distributions π over the product space $\mathcal{Z} \times \mathcal{Z}$ with marginal distributions P and Q. When $\mathcal{Z}$ is a normed space with norm $\|\cdot\|$, simply taking $\rho(z,z') = \|z - z'\|$ leads to

$$\mathbb{W}(P,Q) \triangleq \inf_{\pi \in \Pi(P,Q)} \int_{\mathcal{Z} \times \mathcal{Z}} \|z - z'\| \, d\pi(z,z'). \quad (4)$$

Another representation for the Wasserstein distance is given by the Kantorovich-Rubinstein duality [28], i.e.,

$$\mathbb{W}(P,Q) = \sup_{g \in \{g: \text{Lip}(g) \leq 1\}} \left\{ \mathbb{E}_P[g(Z)] - \mathbb{E}_Q[g(Z)] \right\}, \quad (5)$$

where $\text{Lip}(g)$ denotes the Lipschitz constant of function $g : \mathcal{Z} \to \mathbb{R}$, namely

$$\text{Lip}(g) \triangleq \inf \{ L > 0 : \quad\quad\quad\quad (6)$$
$$|g(z_1) - g(z_2)| \leq L\|z_1 - z_2\|, \; z_1, z_2 \in \mathcal{Z}\}.$$

The total variation distance between P and Q is given by

$$\mathbb{TV}(P,Q) \triangleq \frac{1}{2} \int |dP - dQ|. \quad (7)$$

Note that total variation distance also arises from Wasserstein distance [28], i.e., $\mathbb{TV}(P,Q) = \mathbb{W}(P,Q)$, when $\rho(z,z') = \mathbb{1}\{z \neq z'\}$ where $\mathbb{1}$ is an indicator function.

II. PROBLEM FORMULATION

Let $S = \{Z_i\}_{i=1}^n$ be the training set, where each Z_i is defined on the same alphabet $\mathcal{Z}$. Note that Z_i is not required to be i.i.d generated from the same data-generating distribution P_Z, and we denote the joint distribution of all the training samples as P_S. We denote the hypotheses by $w \in \mathcal{W}$, where $\mathcal{W}$ is a hypothesis class. The performance of the hypothesis is measured by a non-negative loss function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}_0^+$, and we can define the empirical risk and the population risk associated with a given hypothesis w as

$$L_E(w,s) \triangleq \frac{1}{n} \sum_{i=1}^n \ell(w,z_i), \quad (8)$$

$$L_P(w,P_S) \triangleq \mathbb{E}_{P_S}[L_E(w,S)], \quad (9)$$

respectively. A learning algorithm can be modeled as a randomized mapping from the training set S onto an hypothesis $W \in \mathcal{W}$ according to the conditional distribution $P_{W|S}$. Thus, the expected generalization error quantifying the degree of over-fitting can be written as

$$\overline{\text{gen}}(P_{W|S}, P_S) \triangleq \mathbb{E}_{P_{W,S}}[L_P(W,P_S) - L_E(W,S)], \quad (10)$$

where the expectation is taken over the joint distribution $P_{W,S} = P_{W|S} \otimes P_S$.

In this paper, we construct different upper bounds for generalization error using the average joint distribution, which is defined as

$$\overline{P}_{W,\overline{Z}}(w,z) \triangleq \frac{1}{n} \sum_{i=1}^n P_{W,Z_i}(w,z). \quad (11)$$

Note that the average sample distribution is defined as

$$\overline{P}_{\overline{Z}}(z) \triangleq \frac{1}{n} \sum_{i=1}^n P_{Z_i}(z). \quad (12)$$

It is worthwhile to mention that under i.i.d assumption we have $\overline{P}_{\overline{Z}} = P_Z$. Similarly, the average conditional distribution is defined as

$$\overline{P}_{W|\overline{Z}=z}(w) = \frac{1}{n} \sum_{i=1}^n P_{W|Z_i=z}(w). \quad (13)$$

A learning algorithm is said to be *symmetric*, if the conditional distributions between each sample Z_i and hypothesis W are the same, i.e., $P_{W|Z_i} = P_{W|Z}, \; \forall i \in \{1, \cdots, n\}$.

III. GENERALIZATION ERROR UPPER BOUNDS

This section provides expected generalization error upper bounds in terms of different information measures, including Wasserstein distance, total variation distance, KL divergence, and Jensen-Shannon divergence. In the case of Wasserstein distance and total variation distance, our upper bounds are shown to be tighter than existing upper bounds based on these information measures.

To present our result, we first show that the expected generalization error can be expressed in terms of the average joint distribution (11).

Proposition 1. *The expected generalization error of a learning algorithm $P_{W|S}$ can be written as*

$$\overline{\text{gen}}(P_{W|S}, P_S) = \mathbb{E}_{P_W \otimes \overline{P}_{\overline{Z}}}[\ell(W, Z)] - \mathbb{E}_{\overline{P}_{W, \overline{Z}}}[\ell(W, Z)]. \tag{14}$$

Proof. By the definition of generalization error, we have

$$\overline{\text{gen}}(P_{W|S}, P_S) = \mathbb{E}_{P_{W,S}}[L_P(W, P_S) - L_E(W, S)]$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{P_W \otimes P_{Z_i}}[\ell(W, Z)] - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{P_{W,Z_i}}[\ell(W, Z)] \tag{15}$$

$$= \mathbb{E}_{P_W \otimes \frac{1}{n}\sum_{i=1}^{n} P_{Z_i}}[\ell(W, Z)] - \mathbb{E}_{\frac{1}{n}\sum_{i=1}^{n} P_{W,Z_i}}[\ell(W, Z)],$$

where the last line follows by the linearity of expectation. $\square$

This characterization embodied in Proposition 1 leads directly to various generalization error bounds in terms of different information measures.

A. Wasserstein Distance-based Upper Bound

In the following theorem, we provide a generalization error upper bound based on Wasserstein distance using (14) under Lipschitz condition.

Theorem 1. *Suppose that for all $z \in \mathcal{Z}$, the loss function $\ell(\cdot, z)$ is L-Lipschitz, and we have i.i.d. training samples $S = \{Z_i\}_{i=1}^{n}$. Then, we have the following upper bound*

$$|\overline{\text{gen}}(P_{W|S}, P_S)| \leq L\mathbb{E}_{P_Z}[\mathbb{W}(\overline{P}_{W|\overline{Z}}, P_W)]. \tag{16}$$

Proof. We have $\overline{P}_{\overline{Z}} = P_Z$ for i.i.d. data samples, then $\overline{P}_{W,\overline{Z}} = \overline{P}_{W|\overline{Z}} \otimes P_Z$. By (14), we have

$$|\overline{\text{gen}}(P_{W|S}, P_S)| = |\mathbb{E}_{P_Z}[\mathbb{E}_{P_W}[\ell(W, Z)] - \mathbb{E}_{\overline{P}_{W|\overline{Z}}}[\ell(W, Z)]]|$$

$$\leq L\mathbb{E}_{P_Z}[\mathbb{W}(\overline{P}_{W|\overline{Z}}, P_W)], \tag{17}$$

where the last inequality follows from Kantorovich-Rubinstein duality (5). $\square$

In the following, we show that our upper bound in Theorem 1 would be tighter than the individual sample Wasserstein distance upper bound in [14].

Proposition 2. *Under the same assumption as in Theorem 1, the upper bound in Theorem 1 is always no worse than the upper bound in [14, Theorem 1], i.e.,*

$$|\overline{\text{gen}}(P_{W|S}, P_S)| \leq L\mathbb{E}_{P_Z}[\mathbb{W}(\overline{P}_{W|\overline{Z}}, P_W)]$$

$$\leq \frac{L}{n}\sum_{i=1}^{n} \mathbb{E}_{P_Z}[\mathbb{W}(P_{W|Z_i}, P_W)]. \tag{18}$$

Proof. By Kantorovich-Rubinstein duality (5), we have

$$L\mathbb{W}(\overline{P}_{W|\overline{Z}}, P_W) = \sup_{g \in \{g: \text{Lip}(g) \leq 1\}} \{\mathbb{E}_{\overline{P}_{W|\overline{Z}}}[g] - \mathbb{E}_{P_W}[g]\}$$

$$\leq \frac{1}{n}\sum_{i=1}^{n} \sup_{g \in \{g: \text{Lip}(g) \leq 1\}} \{\mathbb{E}_{P_{W|Z_i}}[g] - \mathbb{E}_{P_W}[g]\}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{W}(P_{W|Z_i}, P_W), \tag{19}$$

where the inequality follows from convexity of supremum function. $\square$

Remark 1. *The upper bound based on average conditional distribution in Theorem 1 will reduce to the individual sample Wasserstein distance based upper bound in [14, Theorem 1] when the learning algorithm $P_{W|S}$ is symmetric, i.e., $P_{W|Z_i} = P_{W|Z}$ for all i.*

B. Total Variation Distance-based Upper Bound

In the following result, we provide a tighter expected generalization error upper bound in terms of total variation distance for bounded loss functions.

Proposition 3. *Suppose that the loss function is bounded, i.e., $\ell \in [a, b]$, and we have i.i.d. training samples $S = \{Z_i\}_{i=1}^{n}$. Then, the following upper bound holds*

$$|\overline{\text{gen}}(P_{W|S}, P_S)| \leq (b - a)\mathbb{E}_{P_Z}[\mathbb{TV}(\overline{P}_{W|\overline{Z}}, P_W)]$$

$$= \mathbb{TV}(\overline{P}_{W,\overline{Z}}, P_W \otimes P_Z). \tag{20}$$

Proof. The bounded condition implies that the loss function $\ell(\cdot, z)$ is $(b - a)$-Lipschitz for all $z \in \mathcal{Z}$. Recall that total variation distance is a special case of Wasserstein distance with $\rho(z, z') = \mathbb{1}\{z \neq z'\}$, then the inequality can be proved by applying Theorem 1 directly.

By the assumption of i.i.d. training samples and the definition of total variation in (7), we have

$$\mathbb{E}_{P_Z}[\mathbb{TV}(\overline{P}_{W|\overline{Z}}, P_W)] = \mathbb{TV}(\overline{P}_{W,\overline{Z}}, P_W \otimes P_Z), \tag{21}$$

which completes the proof for the equality. $\square$

Next, we compare our upper bound in terms of total variation distance with the individual sample total variation distance based upper bound in [14, Corollary 1].

Corollary 1. *Under the same assumptions as in Proposition 3, the upper bound in Proposition 3 is always no worse than the individual sample total variation distance bound in [14, Corollary 1], i.e.,*

$$|\overline{\text{gen}}(P_{W|S}, P_S)| \leq (b - a)\mathbb{E}_{P_Z}[\mathbb{TV}(\overline{P}_{W|\overline{Z}}, P_W)]$$

$$\leq \frac{(b - a)}{n}\sum_{i=1}^{n} \mathbb{E}_{P_Z}[\mathbb{TV}(P_{W|Z_i}, P_W)]. \tag{22}$$

Proof. As the total variation is an f-divergence, it has the joint convexity property with respect to its input [25]. Thus, the result follows by applying the convexity of the total variation distance in (22). $\square$

Remark 2. *Under the same assumptions as in Proposition 3, it is shown in [14, Corollary 1] that the upper bound based on individual sample total variation distance is tighter than the Individual sample mutual information (ISMI) [8]. Therefore, our upper bound in Proposition 2 and Corollary 1 would also be tighter than the ISMI bound.*

The proposed bound in Proposition 3 will reduce to the individual sample total variation distance-based bound in [14, Corollary 1], when the learning algorithm is symmetric. However, we may want to use *non-symmetric* learning algorithm in practice since the importance of each training sample is

not the same, e.g., imbalanced classification or learning under noisy data samples. As we will show in Section V, for a non-symmetric learning algorithm, our proposed upper bound will be strictly tighter than the bound in [14, Corollary 1].

C. KL Divergence-based Upper Bound

In the following theorem, we provide an upper bound in terms of KL divergence using (14) under sub-Gaussian condition.

Theorem 2. *Suppose that the loss function $\ell(w, z)$ is σ-Gaussian[1] under distribution $P_W \otimes \overline{P_{\overline{Z}}}$. The following upper bound holds on the expected generalization error*

$$\overline{gen}(P_{W|S}, P_S) \leq \sqrt{2\sigma^2 D(\overline{P}_{W,\overline{Z}} \| P_W \otimes \overline{P_{\overline{Z}}})}. \quad (23)$$

Sketch of Proof. Applying the Donsker-Varadhan representation of KL divergence (1) to the generalization error expressed in (14) and using the σ-sub-Gaussianity in a similar approach to [7, Lemma 1], it completes the proof. □

In the following, we compare our KL divergence based upper bound with the mutual information based bound in [7, Theorem 1].

Corollary 2. *Under the same assumption as in Theorem 2, and further assume that training samples $S = \{Z_i\}_{i=1}^n$ are i.i.d., the upper bound in Theorem 2 is no worse than the mutual information-based upper bound in [7, Theorem 1], i.e.,*

$$\overline{gen}(P_{W|S}, P_S) \leq \sqrt{2\sigma^2 D(\overline{P}_{W,\overline{Z}} \| P_W \otimes \overline{P_{\overline{Z}}})}$$
$$\leq \sqrt{\frac{2\sigma^2}{n} I(W; S)}. \quad (24)$$

Proof. Under i.i.d assumption, $P_Z = P_{\overline{Z}}$. Then, we have

$$\overline{\text{gen}}(P_{W|S}, P_S) \leq \sqrt{2\sigma^2 D(\overline{P}_{W,\overline{Z}} \| P_W \otimes P_Z)} \quad (25)$$

$$\leq \sqrt{\frac{2\sigma^2}{n} \sum_{i=1}^n D(P_{W,Z_i} \| P_W \otimes P_Z)} \quad (26)$$

$$= \sqrt{\frac{2\sigma^2}{n} \sum_{i=1}^n I(W; Z_i)} \quad (27)$$

$$\leq \sqrt{\frac{2\sigma^2}{n} I(W; S)}, \quad (28)$$

where the second inequality follows from the convexity of KL divergence, and the last inequality is due to the chain rule of mutual information and the i.i.d assumption [8, Proposition 2]. □

Remark 3. *Under the same assumption as in Theorem 2, our upper bound in Theorem 2 will reduce to the ISMI bound proposed in [8, Proposition 1], when the learning algorithm $P_{W|S}$ is symmetric.*

We can also provide the following generalization error upper bound in terms of the reversed KL divergence using the average joint distribution as in (14).

---

[1]A random variable X is σ-sub-Gaussian if $E[e^{\lambda(X - E[X])}] \leq e^{\frac{\lambda^2 \sigma^2}{2}}$ for all $\lambda \in \mathbb{R}$.

Proposition 4. *Suppose that the loss function $\ell(w, z)$ is σ-sub-Gaussian under $\overline{P}_{W,\overline{Z}}$ distribution. Then, the following upper bound holds*

$$\overline{gen}(P_{W|S}, P_S) \leq \sqrt{2\sigma^2 D(P_W \otimes \overline{P_{\overline{Z}}} \| \overline{P}_{W,\overline{Z}})}. \quad (29)$$

Similar to Corollary 2, we have the following result.

Corollary 3. *Under the same assumption as in Proposition 4, the upper bound in Proposition 4 is always no worse than the upper bound based on individual sample Lautum Information,*

$$\overline{gen}(P_{W|S}, P_S) \leq \sqrt{2\sigma^2 D(P_W \otimes P_Z \| \overline{P}_{W,\overline{Z}})}$$
$$\leq \sqrt{\frac{2\sigma^2}{n} \sum_{i=1}^n L(W; Z_i)}. \quad (30)$$

D. Jensen-Shannon Divergence Based Upper Bound

We can also apply the average joint distribution approach to the Jensen-Shannon divergence based upper bound in [11].

Theorem 3. *Suppose that the loss function $\ell(w, z)$ is σ-sub-Gaussian under distribution $\frac{P_W \otimes \overline{P_{\overline{Z}}} + \overline{P}_{W,\overline{Z}}}{2}$. The following upper bound holds on the expected generalization error*

$$|\overline{gen}(P_{W|S}, P_S)| \leq 2\sqrt{2\sigma^2 D_{JS}(\overline{P}_{W,\overline{Z}} \| P_W \otimes \overline{P_{\overline{Z}}})}. \quad (31)$$

Sketch of Proof. The theorem can be proved by using the auxiliary distribution technique in [11] and considering the generalization error representation in terms of average joint distribution in (14). □

As discussed in [25], Jensen-Shannon is a f-divergence and it is a jointly convex function. Thus, we have:

$$|\overline{\text{gen}}(P_{W|S}, P_S)| \leq 2\sqrt{2\sigma^2 D_{JS}(\overline{P}_{W,\overline{Z}} \| P_W \otimes \overline{P_{\overline{Z}}})}$$
$$\leq 2\sqrt{\frac{2\sigma^2}{n} \sum_{i=1}^n D_{JS}(P_{W,Z_i} \| P_W \otimes P_Z)}, \quad (32)$$

where (32) is an upper bound based on per sample Jensen-Shannon divergence.

IV. THE DIFFERENCE OF EMPIRICAL RISKS

We now consider a slightly different setting. Suppose one has access two different learning algorithms A and B, i.e. $P_{W_A|S}$ and $P_{W_B|S}$. And the goal is to quantify the difference between the empirical risk associated with each of the learning algorithms, i.e.,

$$\Delta_E(A, B) = \mathbb{E}_{P_{W_A, W_B, S}}[L_E(W_A, S) - L_E(W_B, S)]. \quad (33)$$

Using the average joint distribution, we can provide an upper bound on the absolute value of the difference between the empirical risks of these algorithms.

Proposition 5. *Suppose that the loss, $\ell(w, z)$, is σ-sub-Gaussian under $\overline{P}_{W_B, \overline{Z}}$ distribution. The following upper bound holds on the expected difference between empirical risks of two learning algorithms,*

$$|\Delta_E(A, B)| \leq \sqrt{2\sigma^2 D(\overline{P}_{W_A, \overline{Z}} \| \overline{P}_{W_B, \overline{Z}})} \quad (34)$$

Proof. $\Delta_E(A, B)$ can be written as

$$\Delta_E(A, B) = \mathbb{E}_{P_{W_A, W_B, S}}[L_E(W_A, S) - L_E(W_B, S)]$$
$$= \mathbb{E}_{\overline{P}_{W_A, \overline{Z}}}[\ell(W, Z)] - \mathbb{E}_{\overline{P}_{W_B, \overline{Z}}}[\ell(W, Z)]. \quad (35)$$

The final result holds by applying Donsker-Varadhan (1) to (35) and using σ-sub-Gaussian in a similar way as in [7, Lemma 1]. □

In a similar way to Proposition 5, we could provide an upper bound on the difference of two empirical risks achieved using a different number of training samples. Let W' denote the output of the learning algorithm trained with S'_m, which contains m samples, and W is learned using S_n with n samples.

Corollary 4. *Suppose that the loss function $\ell(w, z)$ is σ-sub-Gaussian under distribution $\overline{P}_{W, \overline{Z}}$. We have the following upper bound on the expected difference of empirical risks achieved using different number of training samples*

$$|\mathbb{E}[L_E(W', S'_m) - L_E(W, S_n)]| \leq \sqrt{2\sigma^2 D(\overline{P}_{W', \overline{Z}'} \| \overline{P}_{W, \overline{Z}})},$$

where the expectation is over the distribution P_{W', W, S'_m, S_n}.

V. NUMERICAL EXAMPLE

We illustrate that the proposed bounds can be tighter than existing ones using a simple toy example. The goal of the example is to estimate the mean of a Gaussian random variable $Z \sim \mathcal{N}(\beta, \sigma^2)$ based on two i.i.d. samples Z_1 and Z_2. We consider the estimate given by $W = tZ_1 + (1 - t)Z_2$ for $0 < t < 1$, and adopt the truncated ℓ_2 loss function $\ell(w, z) = \min((w - z)^2, c^2)$. Since the loss function is bounded within the interval $[0, c^2]$, it is $\frac{c^2}{2}$-sub-Gaussian for all w. In the following, we evaluate four generalization error upper bounds based on different information measures: 1) Individual sample mutual information proposed in [8, Proposition 1], 2) KL divergence using average joint distribution in Theorem 2, 3) individual sample total variation distance in [14, Corollary 1], and 4) total variation using average joint distribution in Proposition 3. It can be shown that $W \sim \mathcal{N}(\beta, \sigma^2(t^2 + (1 - t)^2))$, and (W, Z_1) and (W, Z_2) are jointly Gaussian with correlation coefficients $\rho_1 = \frac{t}{\sqrt{t^2 + (1-t)^2}}$ and $\rho_2 = \frac{(1-t)}{\sqrt{t^2 + (1-t)^2}}$, respectively. Note that

$$D(\overline{P}_{W, \overline{Z}} \| P_W \otimes P_Z) = h(P_W) + h(P_Z) - h(\overline{P}_{W, \overline{Z}}), \quad (36)$$

with $h(\cdot)$ denoting the differential entropy, i.e.,

$$h(P_Z) = \frac{1}{2} \log(2\pi\sigma^2 e),$$
$$h(P_W) = \frac{1}{2} \log(2\pi\sigma^2(t^2 + (1 - t)^2)e),$$

whereas $h(\overline{P}_{w, Z^2})$ can be computed numerically.

Fig.1 depicts the four generalization error bounds based on individual sample mutual information, KL divergence using average joint distribution, individual sample total variation distance, total variation distance using average joint distribution, and the true generalization error. It can be seen that
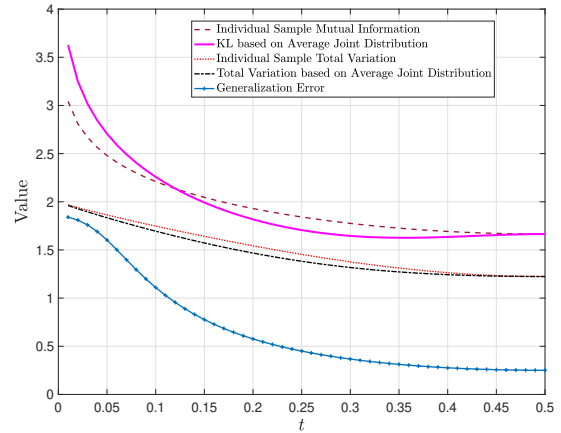

Fig. 1: Comparison of the true generalization error and four generalization error upper bounds in Gaussian mean estimation example with $\sigma = 10$ and $c = 2$, as we change t.

for $t > 0.1$, the upper bound based on KL divergence using average joint distribution is tighter than the individual sample mutual information-based upper bound. In addition, the total variation using average joint distribution gives the tightest upper bound. At $t = 0.5$, the learning algorithm would be symmetric with respect to Z_1 and Z_2. Therefore, the individual sample mutual information-based upper bound equals KL divergence-based upper bound using average join distribution. Similarly, our total variation distance-based upper bound using average joint distribution is equal to the individual sample total variation distance-based upper bound at $t = 0.5$.

VI. CONCLUSION

We have introduced a new approach to obtain information-theoretic bounds of the generalization error for supervised learning problems. Our upper bounds based on Wasserstein distance and total variation distance are tighter than counterparts based on individual samples. Our approach could also be combined with PAC-Bayesian upper bounds [29] and conditional information techniques [17] to tighten the result, which is left for future research.

REFERENCES

[1] M. R. Rodrigues and Y. C. Eldar, *Information-Theoretic Methods in Data Science*. Cambridge University Press, 2021.

[2] V. N. Vapnik, "An overview of statistical learning theory," *IEEE transactions on neural networks*, vol. 10, no. 5, pp. 988–999, 1999.

[3] O. Bousquet and A. Elisseeff, "Stability and generalization," *Journal of machine learning research*, vol. 2, no. Mar, pp. 499–526, 2002.

[4] H. Xu and S. Mannor, "Robustness and generalization," *Machine learning*, vol. 86, no. 3, pp. 391–423, 2012.

[5] D. A. McAllester, "Pac-bayesian stochastic model selection," *Machine Learning*, vol. 51, no. 1, pp. 5–21, 2003.

[6] D. Russo and J. Zou, "How much does your data exploration overfit? controlling bias via information usage," *IEEE Transactions on Information Theory*, vol. 66, no. 1, pp. 302–323, 2019.

[7] 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.

[8] 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.

[9] E. Modak, H. Asnani, and V. M. Prabhakaran, "Rényi divergence based bounds on generalization error," in *2021 IEEE Information Theory Workshop (ITW)*. IEEE, 2021, pp. 1–6.

[10] A. R. Esposito, M. Gastpar, and I. Issa, "Generalization error bounds via rényi-, f-divergences and maximal leakage," *IEEE Transactions on Information Theory*, 2021.

[11] G. Aminian, L. Toni, and M. R. Rodrigues, "Jensen-Shannon information based characterization of the generalization error of learning algorithms," *2020 IEEE Information Theory Workshop (ITW)*, 2020.

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

[13] H. Wang, M. Diaz, J. C. S. Santos Filho, and F. P. Calmon, "An information-theoretic view of generalization via Wasserstein distance," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 577–581.

[14] B. R. Gálvez, G. Bassi, R. Thobaben, and M. Skoglund, "Tighter expected generalization error bounds via Wasserstein distance," in *Advances in Neural Information Processing Systems*, 2021.

[15] 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.

[16] A. R. Asadi and E. Abbe, "Chaining meets chain rule: Multilevel entropic regularization and training of neural networks," *Journal of Machine Learning Research*, vol. 21, no. 139, pp. 1–32, 2020.

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

[18] R. Zhou, C. Tian, and T. Liu, "Individually conditional individual mutual information bound on generalization error," *IEEE Transactions on Information Theory*, pp. 1–1, 2022.

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

[20] 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." *Advances in Neural Information Processing Systems*, 2020.

[21] M. S. Masiha, A. Gohari, M. H. Yassaee, and M. R. Aref, "Learning under distribution mismatch and model misspecification," in *IEEE International Symposium on Information Theory (ISIT)*, 2021.

[22] Y. Bu, W. Gao, S. Zou, and V. Veeravalli, "Information-theoretic understanding of population risk improvement with model compression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 3300–3307.

[23] Y. Bu, W. Gao, S. Zou, and V. V. Veeravalli, "Population risk improvement with model compression: An information-theoretic approach," *Entropy*, vol. 23, no. 10, p. 1255, 2021.

[24] G. Aminian, Y. Bu, L. Toni, M. Rodrigues, and G. Wornell, "An exact characterization of the generalization error for the Gibbs algorithm," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[25] Y. Polyanskiy and Y. Wu, "Lecture notes on information theory," *Lecture Notes for ECE563 (UIUC) and*, vol. 6, no. 2012-2016, p. 7, 2014.

[26] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information theory*, vol. 37, no. 1, pp. 145–151, 1991.

[27] D. P. Palomar and S. Verdú, "Lautum information," *IEEE transactions on information theory*, vol. 54, no. 3, pp. 964–975, 2008.

[28] C. Villani, *Optimal transport: old and new*. Springer, 2009, vol. 338.

[29] T. van Erven, "Pac-bayes mini-tutorial: a continuous union bound," *arXiv preprint arXiv:1405.1580*, 2014.