Learning Algorithm Generalization Error Bounds via Auxiliary Distributions

Gholamali Aminian, Member, IEEE, Saeed Masiha, Member, IEEE, Laura Toni, Senior Member, IEEE, and Miguel R. D. Rodrigues, Fellow, IEEE

Abstract—Generalization error bounds are essential for comprehending how well machine learning models work. In this work, we suggest a novel method, i.e., the Auxiliary Distribution Method, that leads to new upper bounds on expected generalization errors that are appropriate for supervised learning scenarios. We show that our general upper bounds can be specialized under some conditions to new bounds involving the α-Jensen-Shannon, α-Rényi (0 < α < 1) information between a random variable modeling the set of training samples and another random variable modeling the set of hypotheses. Our upper bounds based on α-Jensen-Shannon information are also finite. Additionally, we demonstrate how our auxiliary distribution method can be used to derive the upper bounds on excess risk of some learning algorithms in the supervised learning context and the generalization error under the distribution mismatch scenario in supervised learning algorithms, where the distribution mismatch is modeled as α-Jensen-Shannon or α-Rényi divergence between the distribution of test and training data samples distributions. We also outline the conditions for which our proposed upper bounds might be tighter than other earlier upper bounds.

Index Terms—Expected generalization error bounds, population risk upper bound, mutual information, α-Jensen-Shannon information, α-Rényi information, distribution mismatch.

I. INTRODUCTION

NUMEROUS methods have been proposed in order to describe the generalization error of learning algorithms. These include VC-based bounds [2], algorithmic stability-based bounds [3], algorithmic robustness-based bounds [4], PAC-Bayesian bounds [5]. Nevertheless, for a number of reasons, many of these generalization error bounds are unable to describe how different machine-learning techniques can generalize: some of the bounds depend only on the hypothesis class and not on the learning algorithm; existing bounds do not easily exploit dependencies between different hypotheses; or do not exploit dependencies between the learning algorithm input and output.

More recently, methods that use information-theoretic tools have also been developed to describe the generalization of learning techniques. Such methods frequently incorporate the many components related to the learning problem by expressing the expected generalization error in terms of certain information measurements between the learning algorithm input (the training dataset) and output (the hypothesis). In particular, building upon pioneering work by Russo and Zou [6], Xu and Raginsky [7] have derived expected generalization error bounds involving the mutual information between the training set and the hypothesis. Bu et al. [8] have derived tighter expected generalization error bounds involving the mutual information between each individual sample in the training set and the hypothesis. Meanwhile, bounds using chaining mutual information have been proposed in [9], [10]. Other authors have also constructed information-theoretic based expected generalization error bounds based on other information measures such as α-Rényi divergence for 0 < α < 1, f-divergence, and maximal leakage [11]. In [12], an upper bound based on α-Rényi divergence for 0 < α < 1 is derived by using the variational representation of α-Rényi divergence. Bounds based on the Wasserstein distance between the training sample data and the output of a randomized learning algorithm [13], [14] and Wasserstein distance between distributions of an individual sample data and the output of the learning algorithm is proposed in [15], and tighter upper bounds via convexity of Wasserstein distance are proposed in [16]. Upper bounds based on conditional mutual information and individual sample conditional mutual information are proposed in [17] and [18], respectively. The combination of conditioning and processing techniques can provide tighter expected generalization error upper bounds [19]. An exact characterization of the expected generalization error for the Gibbs algorithm in terms of symmetrized KL information is provided in [20]. Reference [21] provides information-theoretic expected generalization error upper bounds in the presence of training/test data distribution mismatch, using rate-distortion theory.

Generalization error bounds have also been developed to address scenarios where the training data distribution differs from the test data distribution, known as
II. PROBLEM FORMULATION

A. Notations

In this work, we adopt the following notation in the sequel. Calligraphic letters denote spaces (e.g., \( \mathcal{Z} \)), Upper-case letters denote random variables (e.g., \( Z \)), and lower-case letters denote a realization of random variable (e.g., \( z \)). We denote the distribution of the random variable \( Z \) by \( P_Z \), the joint distribution of two random variables \((Z_1, Z_2)\) by \( P_{Z_1,Z_2} \), and the \( \alpha \)-convex combination of the joint distribution \( P_{Z_1,Z_2} \) and the product of two marginals \( P_{Z_1} \otimes P_{Z_2} \), i.e., \( \alpha P_{Z_1} \otimes P_{Z_2} + (1-\alpha)P_{Z_1,Z_2} \) for \( \alpha \in (0, 1) \), by \( P_{Z_1,Z_2}^{(\alpha)} \). The set of distributions (measures) over a space \( \mathcal{X} \) with \( \mu \) is denoted \( \mathcal{P}(\mathcal{X}) \). We denote the derivative of a real-valued function \( f(x) \) with respect to its argument \( x \) by \( f'(\cdot) \). We also adopt the notion \( \log(\cdot) \) for the natural logarithm. The function \( f(x) \) is \( L_f \)-Lipschitz if \( |f(x_1) - f(x_2)| \leq L_f \|x_1 - x_2\|_2 \), where \( \| \cdot \|_2 \) is \( L_2 \)-norm. Let \( \mathcal{N}(\alpha, B) \) denote the Gaussian distribution over \( \mathbb{R}^d \) with mean \( \alpha \in \mathbb{R}^d \) and covariance matrix \( B \in \mathbb{R}^{d \times d} \).

B. Framework of Statistical Learning

We analyze a standard supervised learning setting where we wish to learn a hypothesis given a set of input-output examples that can then be used to predict a new output given a new input.

In particular, in order to formalize this setting, we model the input data (also known as features) using a random variable \( X \in \mathcal{X} \) where \( \mathcal{X} \) is the input space, and we model the output data (also known as predictors or labels) using a random variable \( Y \in \mathcal{Y} \) where \( \mathcal{Y} \) is the output space. We also model input-output data pairs using a random variable \( Z = (X, Y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y} \) where \( Z \) is drawn from \( \mathcal{P} \) over some unknown distribution \( \mu \). We also let \( Z_{\nu|\nu,1} \) be a training set consisting of \( n \) training examples drawn i.i.d. from \( \mathcal{Z} \) according to \( \mu \).

Our goal is to learn a parameterized function, \( f_W : \mathcal{X} \rightarrow \mathcal{Y} \), where the parameters are a random variable \( W \in \mathcal{W} \subset \mathbb{R}^d \) and \( \mathcal{W} \) is a parameter space. Finally, we represent a learning algorithm via a Markov kernel that maps a given training set \( S \) onto parameter \( W \) defined on the parameter space \( \mathcal{W} \) according to the probability law \( P_{W|S} \).

We introduce a (non-negative) loss function \( \ell : \mathcal{W} \times \mathcal{Z} \rightarrow \mathbb{R}^+ \) that measures how well a hypothesis (parameterized function) predicts an output given an input. We can define the population risk and the empirical risk associated with a given hypothesis as follows:

\[
\ell_W(w) := \mathbb{E}_Z \ell(w, z) \, d\mu(z),
\]

\[
\mathbb{E}[\ell_W(s)] := \frac{1}{n} \sum_{i=1}^{n} \ell(w_i),
\]

respecively. We can also define the (expected) generalization error,

\[
\mathbb{E}[\mathbb{E}_W, S] \mathbb{E}_{P_W,S} \mathbb{E}[\mathbb{E}(W, S, \mu)],
\]

where \( \mathbb{E}(W, s, \mu) := \ell_W(w) - L_E(W, s) \). This (expected) generalization error quantifies by how much the population risk deviates from the empirical risk. This quantity cannot be
computed directly because $\mu$ is unknown, but it can often be (upper) bounded, thereby providing a means to gauge various learning algorithms’ performance. We are also interested in excess risk under the learning algorithm $P_{W|S}$:

$$\mathcal{E}_r(P_{W|S}, \mu) := \mathbb{E}_{P_{W|S}}[L_{\mu}(W)] - \inf_{w \in \mathcal{W}} L_{\mu}(w).$$  \hspace{1cm} (4)

Note that the excess risk can be decomposed as follows,

$$\mathcal{E}_r(P_{W|S}, \mu) = \mathcal{E}(P_{W|S}, \mu) + \mathbb{E}_{P_{W|S}}[L_{E}(W, S)] - \inf_{w \in \mathcal{W}} L_{\mu}(w),$$

where the first term is expected generalization error and the second is statistical excess risk.

Furthermore, we analyse a supervised learning scenario under distribution mismatch (a.k.a. out-of-distribution), where training and test data are drawn from different distributions ($\mu$ and $\mu'$, respectively). In particular, we define the population risk based on test distribution $\mu'$ as

$$L_{P}(w, \mu') \triangleq \int \ell(w, z) d\mu'(z).$$  \hspace{1cm} (5)

We define the mismatched (expected) generalization error as

$$\mathcal{E}(P_{W|S}, \mu, \mu') \triangleq \mathbb{E}_{P_{W,S}}[\mathcal{E}(W, S, \mu, \mu')],$$

where $\mathcal{E}(w, s, \mu, \mu') \triangleq L_{P}(w, \mu') - L_{E}(w, s)$.

Our goal in the sequel will be to derive (upper) bounds on the expected generalization errors (3) and the excess risk (4) in terms of various information-theoretic measures.

C. Auxiliary Distribution Method

We describe our main method to derive upper bounds on the expected generalization error, i.e., the ADM. Consider $P$ and $Q$ as two distributions defined on a measurable space $\mathcal{X}$ and let $f : \mathcal{X} \rightarrow \mathbb{R}$ be a measurable function. Assume that we can use an asymmetric information measure $T(P||Q)$ between $P$ and $Q$ to construct the following upper bound:

$$|\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)]| \leq F(T(P||Q)),$$  \hspace{1cm} (7)

where $F(\cdot)$ is a given non-decreasing concave function.

Consider $R$ as an auxiliary distribution on the same space $\mathcal{X}$. We can use the following upper bound instead of (7):

$$|\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)]| \leq |\mathbb{E}_P[f(X)] - \mathbb{E}_R[f(X)]| + |\mathbb{E}_Q[f(X)] - \mathbb{E}_R[f(X)]| \leq F(T(P||R)) + F(T(Q||R))$$  \hspace{1cm} (8)

From concavity of $F$, we have

$$F(T(P||R)) + F(T(Q||R)) \leq 2F(T(P||R)/2 + T(Q||R)/2)$$  \hspace{1cm} (9)

We assume that $T$ satisfies a reverse triangle inequality as follows:

$$\min_{R \in \mathcal{P}(\mathcal{X})} T(P||R) + T(Q||R) \leq T(P||Q).$$  \hspace{1cm} (10)

Considering $R^* \in \arg\min_{R} T(P||R) + T(Q||R)$, we have

$$|\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)]| \leq 2F(T(P||R^*)/2 + T(Q||R^*)/2)$$  \hspace{1cm} (11)

We can also provide another upper bound based on $T(R||P)$ and $T(R||Q)$ instead of $T(P||R)$ and $T(Q||R)$:

$$|\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)]| \leq |\mathbb{E}_R[f(X)] - \mathbb{E}_P[f(X)]| + |\mathbb{E}_R[f(X)] - \mathbb{E}_Q[f(X)]| \leq F(T(R||P)) + F(T(R||Q)).$$  \hspace{1cm} (12)

Considering $\tilde{R} \in \arg\min_{R \in \mathcal{P}(\mathcal{X})} T(R||P) + T(R||Q)$, we have

$$|\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)]| \leq 2F(T(\tilde{R}||P)/2 + T(\tilde{R}||Q)/2).$$  \hspace{1cm} (13)

Via this ADM approach – taking $T(\cdot||\cdot)$ to be a KL divergence – we can derive expected generalization error upper bounds involving KL divergences as follows:

$$\alpha \text{KL}(P_{W,Z}, \hat{P}_{W,Z}) + (1-\alpha) \text{KL}(P_W \otimes \mu || \hat{P}_{W,Z}),$$  \hspace{1cm} (14)

$$\alpha \text{KL}(\hat{P}_{W,Z} || P_W \otimes \mu) + (1-\alpha) \text{KL}(\hat{P}_{W,Z} \otimes P_W \otimes \mu),$$  \hspace{1cm} (15)

where $\hat{P}_{W,Z}$, $P_{W,Z}$, and $P_W \otimes \mu$ are an auxiliary joint distribution over the space $\mathcal{Z} \times \mathcal{W}$, the true joint distribution of the random variables $W$ and $Z$, and the product of marginal distributions of random variables $W$ and $Z$, respectively. Inspired by the ADM, we use the fact that KL divergence is asymmetric and satisfies the reverse triangle inequality [30]. Hence, we can choose the auxiliary joint distribution, $\hat{P}_{W,Z}$, to derive new upper bounds which are finite or tighter under some conditions.

D. Information Measures

In our characterization of the expected generalization error upper bounds, we will use the information measures between two distributions $P_X$ and $P_{X'}$ on a common measurable space $\mathcal{X}$, summarized in Table I. The last two divergences are $\alpha$-JS divergence, $\alpha$-Rényi divergence, which can be characterized by (14) and (15), respectively (See their characterizations as a convex combination of KL-divergences in Lemmas 2 and 3). They are the main divergences discussed in this paper and defined in Table I. KL divergence, Symmetrized KL divergence, Bhattacharyya distance, and Jensen-Shannon divergence can be obtained as special cases of the first three divergences in Table I.

In addition, in our expected generalization error characterizations, we will also use various information measures between two random variables $X$ and $X'$ with joint distribution $P_{XX'}$ and marginals $P_X$ and $P_{X'}$. These information measures are summarized in Table II. Note that all these information measures are zero if and only if the random variables $X$ and $X'$ are independent.

E. Definitions

We offer some standard definitions that will guide our analysis in the sequel.

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

$$\Lambda_X(\lambda) := \log \mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right].$$  \hspace{1cm} (16)

1.a.k.a. capacity discrimination [31] for $\alpha = 1/2$.  

---

*Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.*
TABLE I
DIVERGENCE MEASURES DEFINITIONS

| Divergence Measure | Definition |
|--------------------|------------|
| KL divergence [32] | $\text{KL}(P_X \| P_{X'}) := \int_X P_X(x) \log \left( \frac{P_X(x)}{P_{X'}(x)} \right) dx$ |
| $\alpha$-JS divergence [33], [34] | $\text{JS}_\alpha(P_X \| P_{X'}) := \alpha \text{KL}(P_X \| \alpha P_X + (1-\alpha)P_{X'}) + (1-\alpha)\text{KL}(P_{X'} \| \alpha P_X + (1-\alpha)P_X)$ |
| Jensen-Shannon divergence [34] | $\text{JSD}(P_X \| P_{X'}) := \frac{1}{2} \text{KL}(P_X \| P_{X'}) + \frac{1}{2} \text{KL}(P_{X'} \| P_X)$ |
| $\alpha$-Rényi divergence for $\alpha \in [0, \infty)$ [35] | $R_\alpha(P_X \| P_{X'}) := \frac{1}{1-\alpha} \log \left( \int_X P_X(x)^\alpha P_{X'}(x)^{1-\alpha} dx \right)$ |
| Bhattacharyya distance [36] | $D_B(P_X \| P_{X'}) := R_{1/2}(P_X \| P_{X'})$ |

$\hat{D}_B(P_X \| P_{X'}) := -\log \left( \int_X \sqrt{P_X(x)P_{X'}(x)} dx \right)$

TABLE II
INFORMATION MEASURES DEFINITIONS

| Information Measure | Definition |
|---------------------|------------|
| Mutual information  | $I(X;X') := \text{KL}(P_{X,X'} \| P_X \otimes P_{X'})$ |
| Lautum information [37] | $L(X;X') := \text{KL}(P_X \| P_{X'} \otimes P_X)$ |
| $\alpha$-JS information ($0 < \alpha < 1$) | $I_{JS}^\alpha(X;X') := \text{JS}_\alpha(P_{X,X'} \| P_X \otimes P_{X'})$ |
| Jensen-Shannon information [38] | $I_{JS}(X;X') := \text{JSD}(P_{X,X'} \| P_X \otimes P_{X'})$ |
| $\alpha$-Rényi information | $I_R^\alpha(X;X') := R_\alpha(P_{X,X'} \| P_X \otimes P_{X'})$ |
| Sibson’s $\alpha$-Mutual information [39] | $I^*_\alpha(X;X') := \min_{Q_{X'}} R_\alpha(P_{X,X'} \| P_X \otimes Q_{X'})$ |

III. UPPER BOUNDS ON THE EXPECTED GENERALIZATION ERROR VIA ADM

We provide a series of bounds on the expected generalization error of supervised learning algorithms based on different information measures using the ADM coupled with KL divergence.

A. $\alpha$-Jensen-Shannon- Based Upper Bound

In the following Theorem, we provide a new expected generalization error upper bound based on KL divergence by applying ADM and using KL divergences terms, $\text{KL}(P_W \otimes \mu \| \hat{P}_{W,Z})$ and $\text{KL}(P_{W,Z} \| \hat{P}_{W,Z})$. All the proof details are deferred to Appendix A in the supplementary material.

**Theorem 1:** Assume that under an auxiliary joint distribution $\hat{P}_{W,Z} \in \mathcal{P}(\mathcal{W} \times \mathcal{Z}) - \Lambda_{l(W,Z)}(\lambda)$ exists, it is upper bounded by $\psi_+(\lambda)$ for $\lambda \in [0,b_+)$, $0 < b_+ < +\infty$, and it is also upper bounded by $\psi_-(\lambda)$ for $\lambda \in (b_-,0]$, $\forall i = 1, \ldots, n$. Also assume that $\psi_+(\lambda)$ and $\psi_-(\lambda)$ are convex functions and $\psi_-(0) = \psi_+(0) = \psi'_+(0) = \psi'_-(0) = 0$. Then, it holds that:

\[
\begin{align*}
\text{gen}(P_W \| \mu) &\leq \frac{1}{n} \sum_{i=1}^{n} \left( \psi'^-_i(A_i) + \psi'^+_i(B_i) \right), \\
-\text{gen}(P_W \| \mu) &\leq \frac{1}{n} \sum_{i=1}^{n} \left( \psi'^-_i(A_i) + \psi'^+_i(B_i) \right),
\end{align*}
\]

where $A_i = \text{KL}(P_{W} \| \mu \| \hat{P}_{W,Z})$, $B_i = \text{KL}(P_{W,Z} \| \hat{P}_{W,Z})$, $\psi'^-_i(x) = \inf_{x \in [0,b_+)} \frac{x + \psi_+(\lambda)}{\lambda}$ and $\psi'^+_i(x) = \inf_{x \in [0,b_-]} \frac{x + \psi_-(\lambda)}{\lambda}$.

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

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

$\psi^*(x) := \sup_{\lambda \in [0,b)} (\lambda x - \psi(\lambda))$. (17)

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

**Lemma 1 [40, Lemma 2.4]:** Assume that $\psi(0) = \psi'(0) = 0$. Then, the Legendre dual $\psi^*(x)$ of $\psi(x)$ defined above is a non-negative convex and non-decreasing function on $[0, \infty)$ with $\psi^*(0) = 0$. Moreover, its inverse function $\psi'^{-1}(y) = \inf_{x \geq 0} : \psi^*(x) \geq y$ is concave, and can be written as

$\psi'^{-1}(y) = \inf_{\lambda \in (0,b)} \left( \frac{y + \psi(\lambda)}{\lambda} \right), \quad b > 0$. (18)

Importantly, using these results, we can characterize the tail behaviour of Sub-Gaussian random variables. A random variable $X$ is $\sigma$-sub-Gaussian, if $\psi(\lambda) = \sigma^2 \lambda^2$ is an upper bound on $\Lambda_X(\lambda)$, for $\lambda \in \mathbb{R}$. Then by Lemma 1,

$\psi'^{-1}(y) = \sqrt{2\sigma^2} y$. (19)

The tail behaviour of sub-Exponential and sub-Gamma random variables are introduced in [20].
Note that Theorem 1 can be applied to sub-Gaussian [19]. It can also sub-Exponential and sub-Gamma assumptions on loss function CGF, introduced in [20].

We can utilize Theorem 1 to recover existing expected generalization error bounds and offer new ones. For example, we can immediately recover the mutual information bound [7] from the following results.

**Example 1:** Choose \( \hat{P}_{W,Z_i} = P_W \otimes \mu \) for \( i = 1, \ldots, n \). It follows immediately from Theorem 1 that:

\[
\gen(P_{W|S}, \mu) \leq \frac{1}{n} \sum_{i=1}^{n} \psi_{-1}^{-1}(I(W; Z_i)),
\]

\[\text{Equation (22)}\]

\[\gen(P_{W|S}, \mu) \leq \frac{1}{n} \sum_{i=1}^{n} \psi_{-1}^{-1}(I(W; Z_i)).\]

\[\text{Equation (23)}\]

**Example 2:** Choose \( \hat{P}_{W,Z_i} = P_{W,Z_i} \) for \( i = 1, \ldots, n \). It also follows immediately from Theorem 1 that:

\[
\gen(P_{W|S}, \mu) \leq \frac{1}{n} \sum_{i=1}^{n} \psi_{-1}^{-1}(L(W; Z_i)),
\]

\[\text{Equation (24)}\]

\[\gen(P_{W|S}, \mu) \leq \frac{1}{n} \sum_{i=1}^{n} \psi_{-1}^{-1}(L(W; Z_i)).\]

\[\text{Equation (25)}\]

The result in Example 1 is the same as a result appearing in [8] whereas the result in Example 2 extends the result appearing in [41].

The conclusion in Theorem 1 can be extended to many auxiliary distributions by repeatedly using ADM. In this study, we take into account just one auxiliary distribution and use ADM just once.

Building upon Theorem 1, we are also able to provide an expected generalization error upper bound based on a convex combination of KL terms, i.e.,

\[
\alpha KL(P_W \otimes \mu \| \hat{P}_{W,Z_i}) + (1 - \alpha)KL(P_{W,Z_i} \| \hat{P}_{W,Z_i}),
\]

that relies on a certain \( \sigma \)-sub-Gaussian tail assumption.

**Proposition 1:** Assume that the loss function is \( \hat{\sigma} \)-sub-Gaussian—under the distribution \( P_{W,Z_i} \) \( \forall i = 1, \ldots, n \)—Then, it holds \( \forall \alpha \in (0, 1) \) that:

\[
\gen(P_{W|S}, \mu) \leq \frac{1}{n} \sum_{i=1}^{n} \left\{ 2 \hat{\sigma}^2 \frac{(\alpha A_i + (1 - \alpha) B_i)}{\alpha(1 - \alpha)} \right\},
\]

\[\text{Equation (26)}\]

where \( A_i = KL(P_{W,Z_i} \| \hat{P}_{W,Z_i}) \) and \( B_i = KL(P_W \otimes \mu \| \hat{P}_{W,Z_i}). \)

We propose a Lemma connecting certain KL divergences to the \( \alpha \)-JS information.

**Lemma 2:** Consider an auxiliary distribution \( \hat{P}_{W,Z} \in \mathcal{P}(W \times Z) \). Then, the following equality holds:

\[
\alpha KL(P_W \otimes \mu \| \hat{P}_{W,Z}) + (1 - \alpha)KL(P_{W,Z} \| \hat{P}_{W,Z}) = I_{JS}^{(\alpha)}(W; Z_i) + KL \left( P_{W,Z}^{(\alpha)} \| \hat{P}_{W,Z} \right).
\]

\[\text{Equation (27)}\]

Note that the proof is inspired by [42].

Using the result in Proposition 1 and ADM we can provide a tighter upper bound. For this purpose, Lemma 2 paves the way to apply ADM and offer a tighter version of the expected generalization error bound appearing in Proposition 1 based on choosing an appropriate auxiliary distribution, as well as recover existing ones.

**Theorem 2:** Assume that the loss function is \( \sigma(\alpha) \)-sub-Gaussian—under the distribution \( P_{W,Z_i} \) \( \forall i = 1, \ldots, n \). Then, it holds \( \forall \alpha \in (0, 1) \) that:

\[
\gen(P_{W|S}, \mu) \leq \frac{1}{n} \sum_{i=1}^{n} \left\{ 2 \hat{\sigma}^2 \frac{I_{JS}^{(\alpha)}(W; Z_i)}{\alpha(1 - \alpha)} \right\},
\]

\[\text{Equation (28)}\]

The bound in Theorem 2 results from minimizing the term \( \alpha KL(P_W \otimes \mu \| \hat{P}_{W,Z}) + (1 - \alpha)KL(P_{W,Z_i} \| \hat{P}_{W,Z_i}) \), in the upper bound (26), presented in Proposition 1, over the joint auxiliary distribution \( \hat{P}_{W,Z} \). Such an optimal joint auxiliary distribution is \( P_{W,Z_i}^{(\alpha)} = \alpha P_W + (1 - \alpha)P_{W,Z_i} \). Note that, the parameter of sub-Gaussianity, denoted as \( \hat{\sigma} \) in Proposition 1, relies on \( \hat{P}_{W,Z} \). Consequently, the upper bound mentioned in Theorem 2 is not the minimum of the upper bound presented in Proposition 1. However, assuming a bounded loss function, the upper bound in Theorem 2 becomes the minimum of the upper bound in Proposition 1.

It turns out that we can immediately recover existing bounds from Theorem 2 depending on how we choose \( \alpha \).

**Remark 1 (Recovering Upper Bound Based on Jensen-Shannon Information):** The expected generalization error upper bound based on Jensen-Shannon information in [1] can be immediately recovered by considering \( \alpha = \frac{1}{2} \) in Theorem 2.

**Remark 2 (Recovering Upper Bounds Based on Mutual Information and Lautum Information):** The expected generalization error upper bound based on mutual information in [8] and lautum information in [41] can be immediately recovered by considering \( \alpha \to 1 \) and \( \alpha \to 0 \) in Theorem 2, respectively.

Note that we can also establish how the bound in Theorem 2 behaves as a function of the number of training samples. This can be done by using \( \hat{P}_{W,Z} = P_W \otimes \mu \) in Lemma 2, leading up to:

\[
(1 - \alpha)I(W; Z_i) = I_{JS}^{(\alpha)}(W; Z_i) + KL \left( P_{W,Z}^{(\alpha)} \| P_W \otimes \mu \right).
\]

\[\text{Equation (29)}\]

and in turn to the following inequality:

\[
I_{JS}^{(\alpha)}(W; Z_i) \leq (1 - \alpha)I(W; Z_i), \quad \forall \alpha \in (0, 1).
\]

\[\text{Equation (30)}\]

We prove the convergence rate of the upper bound in Theorem 2 using (28).

**Proposition 2:** Assume the hypothesis space is finite and the data samples, \( \{Z_i\}_{i=1}^{n} \), are i.i.d. Then, the bound in Theorem 2 has a convergence rate of \( O(\frac{1}{\sqrt{n}}) \).

The value of this new proposed bound presented in Theorem 2 in relation to existing bounds can also be further appreciated by offering two additional results.

**Proposition 3:** Consider the assumptions in Theorem 2. Then, it follows that:

\[
\gen(P_{W|S}, \mu) \leq \sigma(\alpha) \sqrt{\frac{2}{(1 - \alpha)}} h(\alpha), \quad \forall \alpha \in (0, 1),
\]

\[\text{Equation (31)}\]

where \( h(\alpha) = -\alpha \log(\alpha) - (1 - \alpha) \log(1 - \alpha). \)

This proposition shows that, unlike the mutual information-based and lautum information-based generalization bounds that currently exist (e.g. [7], [8], [9], and [11]) the proposed
\(\alpha\)-JS information generalization bound is always finite. We can also optimize the bound in (29) with respect to \(\alpha\), where the minimum is achieved at \(\alpha = 1/2\).

**Corollary 1**: Consider the assumptions in Theorem 2. Then, it follows that:

\[
\left|\mathbb{E}(P_{W|S}, \mu)\right| \leq 2\sigma(1/2)\sqrt{2\log(2)}.
\]

Also, this result applies independently of whether the loss function is bounded or not. Naturally, it is possible to show that the absolute value of the expected generalization error is always upper bounded as \(|\mathbb{E}(P_{W|S}, \mu)| \leq (b-a)\) for any bounded loss function within the interval \([a, b]\). If we consider the bounded loss functions in the interval \([a, b]\), then our upper bound (30) would be \(\sqrt{2\log(2)}(b-a)\) which is less than total variation constant upper bound, \(2(b-a)\) presented in [15], [43].

It is worthwhile to mention that our result cannot be immediately recovered from existing approaches such as [11, Th. 2]. For example, if we consider the upper bound based on Jensen-Shannon information, then there exist \(f\)-divergence based representations of the Jensen-Shannon information as follows:

\[
\text{JSD}(P_X, P_{Y_X}) = \int dP_X f\left(\frac{dP_Y}{dP_X}\right).
\]

with \(f(t) = t \log(t) - (1 + t) \log(1 + t)\). However, [11, Th. 2] requires that the function \(f(t)\) associated with the \(f\)-divergence is non-decreasing within the interval \([0, +\infty)\), but such a requirement is naturally violated by the function \(f(t) = t \log(t) - (1 + t) \log(1 + t)\) associated with the Jensen-Shannon divergence.

### B. \(\alpha\)-Rényi-Based Upper Bound

Next, we provide a new expected generalization error upper bound based on KL divergence by applying ADM and using the following KL divergences terms, \(\text{KL}(\hat{P}_{W|Z}||P_W \otimes \mu)\) and \(\text{KL}(\hat{P}_{W|Z}||P_{W,Z})\). All the proof details are deferred to Appendix B in the supplementary material.

**Proposition 4**: Suppose that \(\Lambda_{i,(W|Z)}(\lambda) \leq \gamma^\pm(\lambda)\) and \(\Lambda_{i,(W,Z)}(\lambda) \leq \phi^\pm,\lambda)\), \(i = 1, \ldots, n\) for \(\lambda \in [0, a_+), 0 < a_+ < +\infty\) and \(\lambda \in [0, c_+), 0 < c_+ < +\infty\), under \(P_W \otimes \mu\) and \(P_{W,Z}\), resp. We also have \(\Lambda_{i,(W|Z)}(\lambda) \leq \gamma^\pm(\lambda)\) and \(\Lambda_{i,(W,Z)}(\lambda) \leq \phi^\pm,\lambda)\), \(i = 1, \ldots, n\) for \(\lambda \in (a_-, 0]\), \(-\infty < a_- < 0\) and \(\lambda \in (c_-, 0]\), \(-\infty < c_- < 0\) under \(P_W \otimes \mu\) and \(P_{W,Z}\), resp. Assume that \(\gamma^\pm(\lambda), \phi^\pm,\lambda)\) and \(\gamma_\lambda(\lambda, i)\) are convex functions, \(\gamma^\pm(0) = \gamma^\pm_\lambda(0) = \gamma^\pm_\lambda(0) = 0\) and \(\phi^\pm(0) = \phi^\pm_\lambda(0) = \phi^\pm_\lambda(0) = 0\). Then, the following upper bounds hold,

\[
\text{gen}(P_{W|S}, \mu) \leq \frac{1}{n} \sum_{i=1}^{n} \left(\gamma^\pm_\lambda(D_i) + \phi^\pm_\lambda(C_i)\right),
\]

\[
\text{gen}(P_{W|S}, \mu) \geq \frac{1}{n} \sum_{i=1}^{n} \left(\phi^\pm_\lambda(D_i) + \gamma^\pm_\lambda(C_i)\right).
\]

where \(D_i = \text{KL}(\hat{P}_{W|Z}||P_W \otimes \mu), C_i = \text{KL}(\hat{P}_{W|Z}||P_{W,Z}), \gamma^\pm_\lambda(x) = \inf_{\lambda \in [0, a_+)} \frac{1 + \gamma^\pm(\lambda)}{\lambda} x\).

**Proof**: The proof approach is similar to Theorem 1 by considering different cumulant generating functions and their upper bounds.

Inspired by the upper bound in Proposition 4, we can provide an upper bound on expected generalization error instantly that is dependent on the convex combination of KL divergence terms, i.e.,

\[
\alpha\text{KL}(\hat{P}_{W|Z}||P_{W,Z}) + (1 - \alpha)\text{KL}(\hat{P}_{W|Z}||P_W \otimes \mu),
\]

and assuming \(\sigma\)-sub-Gaussian tail distribution.

**Proposition 5 (Upper Bound With Sub-Gaussian Assumption)**: Assume that the loss function is \(\sigma\)-sub-Gaussian under distribution \(P_W \otimes \mu\) and \(\gamma\)-sub-Gaussian under \(P_{W,Z}\), \(\forall i = 1, \ldots, n\). Then, it holds for \(\forall \alpha \in (0, 1)\) that,

\[
\text{gen}(P_{W|S}, \mu) \leq \frac{1}{n} \sum_{i=1}^{n} \left(\gamma^\pm(0) = \gamma^\pm_\lambda(0) = \gamma^\pm_\lambda(0) = 0\right.
\]

\[
\phi^\pm(0) = \phi^\pm_\lambda(0) = \phi^\pm_\lambda(0) = 0\)\).

A tighter version of the expected generalization error bound appears in Proposition 5 via ADM and using Lemma 3.

**Theorem 3 (Upper Bound Based on \(\alpha\)-Rényi Information)**: Consider the same assumptions in Proposition 5. The following upper bound for \(\forall \alpha \in (0, 1)\) holds,

\[
\text{gen}(P_{W|S}, \mu) \leq \frac{1}{n} \sum_{i=1}^{n} \left(2(\sigma^2 + (1 - \alpha)\gamma^2)\right) \frac{\text{KL}(P_W; Z_i)}{\alpha}.
\]

The bound in Theorem 3 results from minimizing the bound in Proposition 5 over the joint auxiliary distribution \(P_{W,Z} \in \mathcal{P}(W \times Z)\). Such an optimal joint auxiliary distribution is

\[
\hat{P}_{W,Z} = \frac{(P_{Z_i} \otimes P_W)^{\alpha}}{\int_{W \times Z} (dP_{Z_i} \otimes dP_W)^{\alpha} (dP_{Z_i})^{(1-\alpha)}}.
\]

**Remark 3 (Deterministic Algorithms per Sample)**: If the parameter, \(W\), is a deterministic function of data sample \(Z_i\), then \(I(W; Z_i)\) is not well-defined as \(P_{W,Z}\) is not absolutely continuous with respect to \(P_W P_Z\). However, by considering

\[
\text{gen}(P_{W|S}, \mu) \leq \frac{1}{n} \sum_{i=1}^{n} \left(2(\sigma^2 + (1 - \alpha)\gamma^2)\right) \frac{\text{KL}(P_W; Z_i)}{\alpha}.
\]

We say \(\mu \ll \nu\), i.e., \(\mu\) is absolutely continuous with respect to \(\nu\) if \(\nu(A) = 0\) for some \(A \in \mathcal{X}\), then \(\mu(A) = 0\).

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
the $\alpha$-Rényi information for $\alpha \in [0, 1)$, we do not need to assume the absolute continuous.

Remark 4 (Upper Bound Based on the Bhattacharyya Distance): We can derive the expected generalization error upper bound based on Bhattacharyya distance by considering $\alpha = 1/2$ in Theorem 3,

$$\left|\mathbb{E}(P_{W|S}, \mu)\right| \leq \frac{2}{n} \sum_{i=1}^{n} \sqrt{\frac{\alpha^2 + 2}{\alpha^2}} D_B(P_{W,Z_i} \| P_W \otimes \mu),$$

Remark 5 (Recovering the Upper Bound Based on Mutual Information and Lautum Information): We can recover the expected generalization error upper bound based on mutual information in [7] and lautum information in [41] by considering $\alpha \to 1$ and $\alpha \to 0$ in Theorem 3, respectively.

By considering $P_{W,Z_i} = P_{W,Z}$, we have,

$$\alpha I(W; Z_i) = \left(1 - \alpha\right) I_{\alpha}(W; Z_i) + \text{KL} \left(\frac{P_{W,Z}}{P_{W} \otimes P_W}\right)^\alpha (P_{W,Z})^{\left(1 - \alpha\right)} \frac{\int_{W \times Z} (dP_{Z_i} \otimes dP_{W})^\alpha (dP_{W,Z})^{\left(1 - \alpha\right)}}{\int_{W \times Z} (dP_{W,Z})^\alpha (dP_{W,Z})^{\left(1 - \alpha\right)}}.$$  

(37)

Since that KL divergence is non-negative, based on Lemma 3 and the monotonicity of $R_\sigma$ with respect to $\alpha$, we have,

$$I_{\alpha}(W; Z_i) \leq \min \left\{ 1, -\frac{\alpha}{1 - \alpha} \right\} I(W; Z_i).$$  

(38)

The result in (38) implies that our expected generalization error bound based on $\alpha$-Rényi information in Theorem 3 exhibits the same convergence rate as upper bound based on mutual information [7].

Proposition 6 (Convergence Rate of Upper Bound Based on $\alpha$-Rényi Information): Assume the hypothesis space is finite and the data samples are i.i.d. Then, the upper bounds based on $\alpha$-Rényi information in Theorem 3 have a convergence rate of $O\left(\frac{1}{n}\right)$.

We can also provide an upper bound based on Sibson’s $\alpha$-mutual information.

Theorem 4 (Upper Bound Based on Sibson’s $\alpha$ Mutual Information): Assume the loss function is $\sigma$-sub-Gaussian under distribution $\mu$ for all $w \in W$ and $\gamma$-sub-Gaussian under $P_{W,Z}$, $\forall i = 1, \ldots, n$. Then, it holds that:

$$\left|\mathbb{E}(P_{W|S}, \mu)\right| \leq \frac{1}{n} \sum_{i=1}^{n} 2\left(\alpha \sigma^2 + (1 - \alpha) \gamma^2\right) I_{\alpha}(W; Z_i).$$

The upper bound based on $\alpha$-Rényi divergence could also be derived using the variational representation of $\alpha$-Rényi divergence in [44]. This approach is applied in [12] by considering the sub-Gaussianity under $P_{Z_i}$ and $P_{Z|W}$. Our approach is more general, paving the way to offer an upper bound based on $\alpha$-Sibson’s mutual information in Theorem 4, which is derived via ADM. Since that,

$$I_{\alpha}(W; Z_i) = \min_{Q_{W} \in P(W)} R_\sigma(P_{W,Z_i} \| Q_{W} \otimes \mu) \leq R_\sigma(P_{W,Z_i} \| P_W \otimes \mu) = I_{\alpha}(W; Z_i),$$

(39)

the upper bound in Theorem 4 is tighter than the upper bound in Theorem 3. It is worthwhile mentioning that we assume the loss function is $\sigma$-sub-Gaussian under $P_W \otimes \mu$ distribution in Theorem 3. However, in Theorem 4, we consider the loss function is $\sigma$-sub-Gaussian under $\mu$ distribution for all $w \in W$.

We can also apply generalized Pinsker’s inequality [35] to bounded loss functions for bounding the expected generalization error using the $\alpha$-Rényi information between data samples, $S$, and hypothesis, $W$.

Proposition 7: Consider $\ell(w, z)$ be a bounded loss function, i.e., $|\ell(w, z)| \leq b$. Then

$$\left|\mathbb{E}(P_{W|S}, \mu)\right| \leq \frac{1}{n} \sum_{i=1}^{n} 2\frac{b^2}{\alpha} I_{\alpha}(W; Z_i), \quad \forall \alpha \in (0, 1].$$

(41)

Considering the bounded loss function bound can help to provide an upper bound based on $\alpha$-Sibson’s mutual information between $S$ and $W$ in a similar approach to Proposition 7.

C. Comparison of Proposed Upper Bounds

A summary of upper bounds on expected generalization error under various $\sigma$-sub-Gaussian assumptions is provided in Table III.

Remark 6 (Bounded Loss Function): The bounded loss function $f: W \times Z \to [a, b]$ is $(\frac{b-a}{2})$-sub-Gaussian under all distributions [7]. In fact, for bounded functions, we have,

$$\sigma = \gamma = \sigma_{(a)} = \frac{(b-a)}{2}.$$  

(42)

We next compare the upper bounds based on $\alpha$-JS information, Theorem 2, with the upper bounds based on $\alpha$-Rényi information, Theorem 3. The next proposition showcases that the $\alpha$-JS information bound can be tighter than the $\alpha$-Rényi based upper bound under certain conditions. The proof details are deferred to Appendix C in the supplementary material.

Proposition 8 (Comparison of Upper Bounds Based on $\alpha$’-Jensen-Shannon and $\alpha$-Rényi Information Measures): Consider the same assumptions in Theorem 2. Then, it follows that $\alpha$’-Jensen-Shannon bound given by:

$$\left|\mathbb{E}(P_{W|S}, \mu)\right| \leq \frac{1}{n} \sum_{i=1}^{n} 2\sigma^2 \frac{b_{R}(W; Z_i)}{\alpha'(1 - \alpha')}.$$  

(43)

is tighter than the $\alpha$-Rényi based upper bound for $0 \leq \alpha' \leq 1$, given by,

$$\left|\mathbb{E}(P_{W|S}, \mu)\right| \leq \frac{1}{n} \sum_{i=1}^{n} 2\left(\alpha \sigma^2 + (1 - \alpha) \gamma^2\right) I_{\alpha}(W; Z_i).$$  

(44)

provided that $\frac{a \sigma_{(\alpha')}}{1 - \alpha'} \leq I_{\alpha}(W; Z_i)$ holds for $i = 1, \ldots, n$ and $\sigma_{(\alpha')} = \sigma = \gamma$.

Remark 7: The condition in Proposition 8, i.e., $\frac{a \sigma_{(\alpha')}}{1 - \alpha'} \leq I_{\alpha}(W; Z_i)$, could be tightened by considering $\alpha' = \frac{1}{2}$ and considering the upper bound based on Jensen-Shannon information.

Remark 8: If we consider $\alpha \to 1$ and $\alpha' = \frac{1}{2}$ in Proposition 8, then the upper bound based on Jensen-Shannon information is tighter than ones based on mutual information.
information [8] provided that $4 \log(2) \leq I(W; Z_i)$ for all $i = 1, \ldots, n$ and $\sigma = \sigma_{JS}$.

IV. UPPER BOUNDS ON EXCESS RISK

This section provides upper bounds on excess risks for regularized empirical risk minimization (ERM) by $\alpha$-Rényi divergence or $\alpha$-JS divergence.

A. $\alpha$-JS-Regularized ERM

It is interesting to consider the regularized ERM with $\alpha$-JS information between dataset $S$, and hypothesis $W$,

$$
\min_{P_W\mathcal{S}} \mathbb{E}[L_E(W, S)] + \frac{1}{\beta} J_{JS}(W; S).
$$

(45)

where $\beta > 0$ is a parameter that balances fitting and generalization. Since the optimization problem in (45) is dependent on the data generating distribution, we relax the problem and replace $\alpha$-JS information with the $\alpha$-JS divergence $JS_{\alpha}(P_W|s|Q_W|P_S)$, as follows,

$$
\min_{P_W\mathcal{S}} \mathbb{E}[L_E(W, S)] + \frac{1}{\beta} J_{JS}(P_W|s|Q_W|P_S),
$$

(46)

where $Q_W \in \mathcal{P}(W)$ is a prior distribution over parameter space.

Lemma 4 (Solution Existence of $\alpha$-JS-Regularized ERM): The optimization problem in (46) is a convex optimization problem and has a solution.

Proof: The first term in objective $\mathbb{E}[L_E(W, S)]$ is linear in terms of $P_W\mathcal{S}$ and the second term $\frac{1}{\beta} J_{JS}(P_W|s|Q_W|P_S)$ is convex in $P_W\mathcal{S}$ for $0 < \alpha < 1$ due to [45]. Therefore, a solution exists.

Let us define the solution of (45),

$$
P_{W|^S}^{*, JS_{\alpha}} := \arg \min_{P_W\mathcal{S} \in \mathcal{P}(W)} \mathbb{E}[L_E(W, S)] + \frac{1}{\beta} J_{JS}(P_W|s|Q_W|P_S).
$$

In the following, we provide an upper bound on excess risk under $P_{W|^S}^{*, JS_{\alpha}}$ as a learning algorithm.

Theorem 5 (Upper Bound on Excess Risk Under $P_{W|^S}^{*, JS_{\alpha}}$): Assume the bounded loss function, i.e., $|\ell(w, z)| \leq b$ for all $(w, z) \in W \times Z$ and $L$-Lipschitz. Then, the following upper bound holds on the excess risk under $P_{W|^S}^{*, JS_{\alpha}}$,

$$
\mathcal{E}(P_{W|^S}^{*, JS_{\alpha}}, \mu) \leq \frac{2b^2}{h(\alpha)} \sum_{i=1}^{n} P_{JS}(W; Z_i) + \frac{L\sqrt{d}}{\beta} + \frac{J_{JS}(N(w^*, \beta^{-1}I_d)\|Q)}{\beta}.
$$

Remark 9 (Comparison to the Gibbs algorithm): Our convergence rate of the upper bound on the excess risk under $P_{W|^S}^{*, JS_{\alpha}}$ is less than the convergence rate of the upper bound on excess risk under the Gibbs algorithm as the solution of KL-regularized empirical which is $O(n^{-1/4})$, [7, Corollary 3] and [46].

B. $\alpha$-Rényi-Regularized ERM

Similarly, it is interesting to consider the regularized ERM with $\alpha$-Rényi-information between dataset, $S$, and hypothesis, $W$, for $0 < \alpha < 1$,

$$
\min_{P_W\mathcal{S}} \mathbb{E}[L(E(W, S)] + \frac{1}{\beta} J_{\alpha}(W; S),
$$

(47)

where $\beta > 0$ is a parameter that balances fitting and generalization.

Since the optimization problem in (47) is dependent on the data generating distribution, $\mu$, we propose to relax
the problem in (47) by replacing \( \alpha \)-Rényi-information, i.e., \( R_R^\alpha(W; S) \), with \( Q_\alpha(P_{WS}\|Q_W|P_S) \) as follows,

\[
\min_{P_{WS}} E[L_E(W, S)] + \frac{1}{\beta} R_\alpha(P_{WS}\|Q_W|P_S),
\]

where \( Q_\alpha \in \mathcal{P}(W) \).

Lemma 5 (Solution Existence of \( \alpha \)-Rényi-Regularized ERM): The optimization problem considered in (48) is a convex optimization problem.

Proof: The first term in objective \( E[L_E(W, S)] \) is linear in term of \( P_{WS} \) and the second term \( \frac{1}{\beta} R_\alpha(P_{WS}\|Q_W|P_S) \) is convex in \( P_{WS} \) for \( 0 < \alpha < 1 \) due to [35, Th. 11]. Therefore, a solution exists.

Let us define

\[
P_{WS}^{*, \beta, R_\alpha} := \arg \min_{P_{WS} \in \mathcal{P}(W)} E[L_E(W, S)] + \frac{1}{\beta} R_\alpha(P_{WS}\|Q_W|P_S),
\]

as the solution of convex optimization problem (48).

Theorem 6 (Upper Bound on Excess Risk Under \( P_{WS}^{*, \beta, R_\alpha} \)): Assume the bounded loss function, i.e., \( |\ell(w, z)| \leq b \) for all \( (w, z) \in W \times Z \) and \( L \)-Lipschitz. Then, the following upper bound holds on the excess risk under \( P_{WS}^{*, \beta, R_\alpha} \),

\[
E_r(P_{WS}^{*, \beta, R_\alpha}, \mu) \leq \frac{2b^2}{na} \sum_{i=1}^n R_R^\beta(W; Z_i) + \frac{\sqrt{d}}{\beta} + \frac{1}{\beta} R_\alpha(N(w^*, \beta^{-1} I_d) \| Q),
\]

where \( w^* = \arg \min_{w \in W} L_\mu(w) \) and \( I_d \) is identity matrix with size \( d \).

Corollary 3 (Convergence Rate of Excess Risk Under \( P_{WS}^{*, \beta, R_\alpha} \)): Under the same assumptions in Theorem 6, assuming that hypothesis space is finite and \( \beta \) is of order \( \sqrt{n} \), the following upper bound holds on the excess risk of \( P_{WS}^{*, \beta, R_\alpha} \) with convergence rate of \( O(\log(n)/\sqrt{n}) \),

\[
E_r(P_{WS}^{*, \beta, R_\alpha}, \mu) \leq \frac{2b^2}{na} \sum_{i=1}^n R_R^\beta(W; Z_i) + \frac{\sqrt{d}}{\sqrt{n}} + \frac{1}{2\sqrt{n}} \|w^*\|^2_2 + \frac{d}{4\sqrt{n}} \log(n) + \frac{d}{2\sqrt{n}(1 - \alpha)} \log(\alpha).
\]

V. EXPECTED GENERALIZATION ERROR UPPER BOUNDS UNDER DISTRIBUTION MISMATCH

In this section, we extend our results in Section III under distribution mismatch, where the training data distribution differs from the test data distribution. All the proof details are deferred to Appendix E in the supplementary material.

Proposition 9: Assume that the loss function is \( \sigma_{(\alpha)} \)-sub-Gaussian – under the distributions \( p_{WS}^{(\alpha)} \), \( w, z, \forall i = 1, \ldots, n \) and \( \alpha \mu + (1 - \alpha)\mu' \) for all \( w \in W \) – Then under distribution mismatch (6), it holds \( \forall \alpha \in (0, 1) \) that:

\[
\gen(P_{WS}, \mu, \mu') \leq 2\sigma_{(\alpha)}^2 \frac{JS_\alpha(\mu' \| \mu)}{2(1 - \alpha)} \]

\[
+ \frac{1}{n} \sum_{i=1}^n 2\sigma_{(\alpha)}^2 \frac{R_R^{\alpha}(W; Z_i)}{\alpha(1 - \alpha)}, \quad \forall \alpha \in (0, 1).
\]

VI. NUMERICAL EXAMPLE

In this section, we illustrate that some of our proposed bounds can be tighter than existing ones in a simple toy example. We consider the \( \alpha \)-JS and \( \alpha \)-Rényi information only. Our example setting involves the estimation of the mean of a Gaussian random variable \( Z \sim N(\beta, \sigma^2) \) based on two i.i.d. samples \( Z_1 \) and \( Z_2 \). We consider the hypothesis (estimate) given by \( W = iZ_1 + (1 - i)Z_2 \) for \( 0 < i < 1 \). We also consider the loss function given by \( \ell(w, z) = \min(w, z)^2, c^2 \).

Due to the fact that the loss function is bounded within the interval \([0, c^2]\), then it is \( \frac{c^2}{2} \)-sub-Gaussian under all distributions. Therefore, we can apply the expected generalization error upper bounds based on mutual information, \( \alpha \)-JS information and \( \alpha \)-Rényi information \( \forall \alpha \in (0, 1) \) as follows:

\[
\gen(P_{WS}Z_1, Z_2, P_Z) \leq \frac{c^2}{4} \left( \sqrt{2I(W; Z_1)} + \sqrt{2I(W; Z_2)} \right),
\]

\[
\gen(P_{WS}Z_1, Z_2, P_Z) \leq \frac{c^2}{4} \left( \sqrt{2\frac{p_{WS}^R(W; Z_1)}{\alpha(1 - \alpha)}} + \sqrt{2\frac{p_{WS}^R(W; Z_2)}{\alpha(1 - \alpha)}} \right),
\]

\[
\gen(P_{WS}Z_1, Z_2, P_Z) \leq \frac{c^2}{4} \left( \frac{p_{WS}^R(W; Z_1)}{\alpha} + \frac{p_{WS}^R(W; Z_2)}{\alpha} \right).
\]
\( -\frac{1}{2} \log(1 - \rho_t^2) \). In contrast, the \( \alpha \)-JS information appearing above can be computed via an extension of entropic-based formulation of the Jensen-Shannon measure as follows [34]:

\[
I_{JS}(W; Z_t) = h\left( P_{W,Z_t}^{(\alpha)} \right) - \left( ah(P_W) + ah(P_{Z_t}) + (1 - \alpha)h(P_{Z_t,W}) \right),
\]

with \( h(\cdot) \) denoting the differential entropy – where

\[
\begin{align*}
    h(P_{Z_t}) &= \frac{1}{2} \log(2\pi \sigma^2), \\
    h(P_W) &= \frac{1}{2} \log(2\pi \sigma^2 (\rho_t^2 + (1 - t)^2)), \\
    h(P_{W,Z_t}) &= \log(2\pi \sigma^2 (\rho_t^2 + (1 - t)^2) (1 - \rho_t^2)),
\end{align*}
\]

whereas \( h(P_{W,Z_t}^{(\alpha)}) \) can be computed numerically.

Fig. 1 depicts the true generalization error, the mutual information based bound in (51), and the \( \alpha \)-JS information based bound for \( \alpha = 0.25, 0.5, 0.75 \) in (52) for values of \( t \in (0, 0.5] \), considering \( \sigma^2 = 1 \), \( \mu = 1 \), \( \epsilon = \frac{\pi}{7} \).

It can be seen that for \( \alpha = 0.75 \) the \( \alpha \)-JS information bound is tighter than the mutual information bound. For \( \alpha = 0.5 \), which is equal to traditional Jensen-Shannon information, if we consider \( t < 0.25 \) then the Jensen-Shannon information bound is tighter than the mutual information bound; in contrast, for \( t > 0.25 \), the mutual information bound is slightly better than the Jensen-Shannon information bound. This showcases that our proposed bounds can be tighter than existing ones in some regimes.

Fig. 2 also depicts the true generalization error, the mutual information based bound in (51), and the \( \alpha \)-Rényi information based bound for \( \alpha = 0.25, 0.5, 0.75 \) in (53). It can be seen that the \( \alpha \)-Rényi based bound is looser than the mutual information based bound. In our experiment setup, when \( t \to 0 \) (or \( t \to 1 \)), we have \( I(W; Z_t) \to \infty \) (or \( I(W; Z_t) \to \infty \)). However, the \( \alpha \)-Rényi based bound is finite.

VII. CONCLUSION AND FUTURE WORKS

We have presented the Auxiliary Distribution Method, a novel approach for deriving information-theoretic upper bounds on the generalization error within the context of supervised learning problems. Our method offers the flexibility to recover existing bounds while also enabling the derivation of new bounds grounded in the \( \alpha \)-JS and \( \alpha \)-Rényi information measures. Notably, our upper bounds, which are rooted in the \( \alpha \)-JS information measure, are finite, in contrast to mutual information-based bounds. Moreover, our upper bound based on \( \alpha \)-Rényi information, for \( \alpha \in (0, 1) \), remains finite when considering a deterministic learning process. An intriguing observation is that our newly introduced \( \alpha \)-JS information measure can, in certain regimes, yield tighter bounds compared to existing approaches. We also discuss the existence of algorithms under \( \alpha \)-JS-regularized and \( \alpha \)-Rényi-regularized empirical risk minimization problems and provide upper bounds on excess risk of these algorithms, where the upper bound on the excess risk under \( \alpha \)-JS-regularized empirical risk minimization is tighter than other well-known upper bounds on excess risk. Furthermore, we provide an upper bound on generalization error in a mismatch scenario, where the distributions of test and training datasets are different, via our auxiliary distribution method.

As a direction for future research, we propose extending our bounds to the PAC-Bayesian framework, leveraging the \( \alpha \)-JS and \( \alpha \)-Rényi divergences for \( 0 < \alpha < 1 \). Additionally, the conditional technique based on individual sample measures, as described in [18], could be applied to improve the effectiveness of our upper bounds.

REFERENCES

[1] G. Aminian, L. Toni, and M. R. D. Rodrigues, “Jensen-Shannon information based characterization of the generalization error of learning algorithms,” in Proc. IEEE Inf. Theory Workshop (ITW), 2021, pp. 1–5.
[2] V. N. Vapnik, “An overview of statistical learning theory,” IEEE Trans. Neural Netw., vol. 10, no. 5, pp. 988–999, Sep. 1999.
[3] O. Bousquet and A. Elisseeff, “Stability and generalization,” J. Mach. Learn. Res., vol. 2, pp. 999–1026, Mar. 2002.
[4] H. Xu and S. Mannor, “Robustness and generalization,” Mach. Learn., vol. 86, no. 3, pp. 391–423, 2012.
[5] D. A. McAllester, “PAC-Bayesian stochastic model selection,” Mach. Learn., 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 Trans. Inf. Theory, vol. 66, no. 1, pp. 302–323, Jan. 2020.
[7] A. Xu and M. Ragninsky, “Information-theoretic analysis of generalization capability of learning algorithms,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 2524–2533.
[8] Y. Bu, S. Zou, and V. V. Veeravalli, “Tightening mutual information-based bounds on generalization error,” *IEEE J. Select. Areas Inf. Theory*, vol. 1, no. 1, pp. 121–130, May 2020.

[9] A. Asadi, E. Abbe, and S. Verdú, “Chaining mutual information and tightening generalization bounds,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 7234–7243.

[10] A. R. Asadi and E. Abbe, “Chaining meets chain rule: Multilevel entropic regularization and training of neural networks,” *J. Mach. Learn. Res.*, vol. 21, no. 139, pp. 1–32, 2020.

[11] A. R. Esposito, M. Gastpar, and I. Issa, “Generalization error bounds via Rényi-f-divergences and maximal leakage,” *IEEE Trans. Inf. Theory*, vol. 67, no. 8, pp. 4968–5004, Aug. 2021.

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

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

[14] H. Wang, M. Diaz, J. C. S. Santos Filho, and F. P. Calmon, “An information-theoretic view of generalization via Wasserstein distance,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2019, pp. 577–581.

[15] B. R. Galván, G. Bassi, R. Thobaben, and M. Skoglund, “Tighter expected generalization error bounds via Wasserstein distance,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 1–13.

[16] G. Aminian, Y. Bu, G. Wornell, and M. Rodrigues, “Tighter expected generalization error bounds via convexity of information measures,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2022, pp. 2481–2486.

[17] T. Steinke and L. Zakynthinou, “Reasoning about generalization via conditional mutual information,” in *Proc. Conf. Learn. Theory*, 2020, pp. 3437–3452.

[18] R. Zhou, C. Tian, and T. Liu, “Individually conditional individual mutual information bound on Generalization error,” *IEEE Trans. Inf. Theory*, vol. 68, no. 5, pp. 3304–3316, May 2022.

[19] H. Hafez-Kolahi, Z. Golgooni, S. Kasaei, and M. Soleymani, “Conditioning and processing: Techniques to improve information-theoretic generalization bounds,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 16457–16467.

[20] G. Aminian, Y. Bu, L. Toni, M. Rodrigues, and G. Wornell, “An exact characterization of the generalization error for the Gibbs algorithm,” *Adv. Neural Inf. Process. Syst.*, 2021, pp. 8106–8118.

[21] M. S. Masiha, A. Gohari, M. H. Yassaei, and M. R. Aref, “Learning under distribution mismatch and model misspecification,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2021, pp. 2912–2917.

[22] Y. Mansour, M. Mohri, and A. Rostamizadeh, “Domain adaptation: Learning bounds and algorithms,” 2009, arXiv:0902.3430.

[23] Z. Wang, “Theoretical guarantees of transfer learning,” 2018, arXiv:1810.05986.

[24] X. Wu, J. H. Manton, U. Aickelin, and J. Zhu, “Information-theoretic analysis for transfer learning,” in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2020, pp. 2819–2824.

[25] T. Kailath, “The divergence and Bhattacharyya distance measures in signal selection,” *IEEE Trans. Commun. Technol.*, vol. CT-15, no. 1, pp. 52–60, Feb. 1967.

[26] I. Sason and S. Verdú, “f-divergence inequalities,” *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 5973–6006, Nov. 2016.

[27] S. Verdú, “α-mutual information,” in *Proc. Inf. Theory Appl. Workshop (ITA)*, 2015, pp. 1–6.

[28] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford, U.K.: Oxford univ. press, 2013.

[29] M. Gastpar, A. R. Esposito, and I. Issa, “Information measures, learning and generalization,” in *Proc. 5th London Symp. Inf. Theory*, 2019, pp. 1–174.

[30] T. Kailath, “The divergence and Bhattacharyya distance measures in signal selection,” *IEEE Trans. Commun. Technol.*, vol. CT-15, no. 1, pp. 52–60, Feb. 1967.

[31] D. P. Palomar and S. Verdú, “Lautum information,” *IEEE Trans. Inf. Theory*, vol. 54, no. 3, pp. 964–975, Mar. 2008.

[32] I. Sason and S. Verdú, “f-divergence inequalities,” *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 5973–6006, Nov. 2016.

[33] F. Nielsen, “On a generalization of the Jensen-Shannon divergence and related measures of discrimination,” in *Proc. Conf. Learn. Theory*, 2020, pp. 2912–2917.

[34] J. Lin, “Divergence measures based on the Shannon entropy,” *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 145–151, Jan. 1991.

[35] T. Van Erven and P. Harremos, “Rényi divergence and Kullback-Leibler divergence,” *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3797–3820, Jul. 2014.

---

**Gholamali Aminian** (Member, IEEE) received the B.Sc. degree in electrical engineering from Amirkabir University, Tehran, Iran, in 2010, and the M.Sc. and Ph.D. degrees in electrical engineering from the Sharif University of Technology, Tehran, in 2012 and 2017, respectively. He was an Honorary Research Fellow with UCL. In July 2022, he joined the Alan Turing Institute under the FAIR Project as a Research Associate, working on reinforcement learning, graph neural networks, and stability analysis. His fields of interest include information theory, measure theory, and learning theory. He was awarded the Newton International Fellowship by the Royal Society.

**Saeed Masiha** (Member, IEEE) received the B.Sc. degree in electrical engineering from the Sharif University of Technology, Tehran, Iran, in 2021. He is currently pursuing the Ph.D. degree with the Chair of Business Analytics and Information and Network Dynamics Group, École Polytechnique Fédérale Lausanne. His research interests include the theory of machine learning, nonconvex optimization, reinforcement learning, and information theory.
Laura Toni (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Bologna, Italy, in 2009. She is an Associate Professor with the Department of Electronic and Electrical Engineering, University College London. After her Ph.D., she was a Postdoctoral Fellow with the University of California at San Diego from 2011 to 2012 and with the Swiss Federal Institute of Technology, Switzerland, from 2012 to 2016. Her major contributions are in the area of large-scale signal processing for machine learning, graph signal processing, decision-making strategies under uncertainty, and multimedia processing. She has coauthored 30 high-impact journals and over 60 conference publications, and she is co-inventor of two patents on low-delay video processing and streaming. She is the recipient of 2022 TOMM Best Journal Paper Award, the Best Paper Candidate in Best Student Paper Award MMSys 2021, the IEEE Best 10% Paper Award at VCIP 2016, the IEEE Best Paper Award at IEEE ISM 2016, and the ACM Best 10% Paper Award at MMSP 2013. She is significantly involved in scientific committees. She has served as the Technical Program Chair of ACM Multimedia 2022, an Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING, EURASIP Journal on Advances in Signal Processing, and ACM Transactions on Multimedia Computing, Communications, and Applications. She is also an ELLIS Member and an Alan Turing Fellow.

Miguel R. D. Rodrigues (Fellow, IEEE) received the Licenciatura degree in electrical and computer engineering from the University of Porto, Porto, Portugal, and the Ph.D. degree in electronic and electrical engineering from the University College London (UCL), London, U.K. He is currently a Professor of Information Theory and Processing with UCL and a Turing Fellow with the Alan Turing Institute—The U.K. National Institute of Data Science and Artificial Intelligence. His work has led to more than 200 articles in leading journals and conferences in the field, a book on Information-Theoretic Methods in Data Science (Cambridge Univ. Press). His research lies in the general areas of information theory, information processing, and machine learning. He received the IEEE Communications and Information Theory Societies Joint Paper Award 2011. He is an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION THEORY and the IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY. He was an Associate Editor of the IEEE COMMUNICATIONS LETTERS, and a Lead Guest Editor of the Special Issue on “Information-Theoretic Methods in Data Acquisition, Analysis, and Processing” of the IEEE JOURNAL ON SELECTED TOPICS IN SIGNAL PROCESSING. He was the Co-Chair of the Technical Programme Committee of the IEEE Information Theory Workshop 2016, Cambridge, U.K. He is a member of the IEEE Signal Processing Society Technical Committee on “Signal Processing Theory and Methods,” and the EURASIP SAT on Signal and Data Analytics for Machine Learning.