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 ($0 < \alpha < 1$) 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 $\alpha > 1$, f-divergence, and maximal leakage [11]. In [12], an upper bound based on α-Rényi divergence for $0 < \alpha < 1$ 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]. [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 Distribution Mismatch. This scenario – which also links to out-of-distribution generalization – has attracted various contributions in recent years, such as [22]–[24]. In particular, Masiha et al. [21] provides information-theoretic generalization error upper bounds in the presence of training/test data distribution mismatch, using rate-distortion theory.

In this work, we propose an auxiliary distribution method (ADM) to characterize the expected generalization error upper bound of supervised learning algorithms in terms of novel information measures. Our new bounds offer two advantages over existing ones: (1) Some of our bounds – such as the α-JS information ones – are always finite, whereas conventional mutual information ones (e.g., [7]) may not be; (2) In contrast

* Equal Contribution.

One of the ideas of this work was presented, in part, at 2020 IEEE Information Theory Workshop (ITW), [1].

G. Aminian is with the Alan Turing Institute, London NW1 2DB, U.K. (e-mail: gaminian@turing.ac.uk), L. Toni and M. R. D. Rodrigues are with the Electronic and Electrical Engineering Department, University College London, London WC1E 6BT, U.K. (e-mail: {l.toni, m.rodrigues}@ucl.ac.uk), and S. Masiha is with the École Polytechnique Fédérale de Lausanne, Lausanne 1015, Switzerland. (e-mail: mohammadsaeed.masiha@epfl.ch).

to mutual information-based bounds, our bounds—such as the α-Rényi information for $0 < \alpha < 1$—are finite for some deterministic supervised learning algorithms; (3) We also apply ADM to provide an upper bound on population risk of supervised learning algorithms under a learning algorithm.

In summary, our main contributions are as follows:

1) We suggest a novel method, i.e., ADM, that uses auxiliary distributions over the parameter and data sample spaces to obtain upper bounds on the expected generalization error.
2) Using ADM, we derive new expected generalization error bounds expressed via α-JS divergence, which is known to be finite.
3) Using ADM, we offer an upper bound based on α-Rényi divergence for $0 < \alpha < 1$ with the same convergence rate as the mutual information-based upper bound. Furthermore, in contrast to the mutual information-based bounds, the α-Rényi divergence bounds for $0 < \alpha < 1$ can be finite when the hypothesis (output of the learning algorithm) is a deterministic function of at least one data sample.
4) Using our upper bounds on expected generalization error, we also provide upper bounds on excess risk of some learning algorithms as solutions to regularized empirical risk minimization by α-Rényi or α-Jensen-Shannon divergences.
5) Using ADM, we also provide generalization error upper bound under training and test data distribution mismatch. It turns out that training and test distribution mismatch is captured in our upper bounds via α-Jensen-Shannon or α-Rényi divergences.

It is noteworthy to add that, although the α-JS measure does not appear to have been used to characterize the generalization ability of learning algorithms, these information-theoretic quantities as well as α-Rényi measure for $0 < \alpha < 1$, have been employed to study some machine learning problems, including the use of

- α-JS as a loss function under label noise scenario [25], and Jensen-Shannon divergence (α-JS divergence for $\alpha = 1/2$) in adversarial learning [26] and active learning [27].
- α-Rényi divergence in feature extraction [28] and image segmentation based on clustering [29].

II. Problem Formulation

A. Notations

In this work, we adopt the following notation in the sequel. Calligraphic letters denote spaces (e.g. $\mathcal{Z}$), Uppercase letters denote random variables (e.g., Z), and lowercase 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 α-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 $P^{(\alpha)}_{Z_1,Z_2}$. The set of distributions (measures) over a space $\mathcal{X}$ with 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 $\mathcal{N}(a, B)$ denotes the Gaussian distribution over $\mathbb{R}^d$ with mean $a \in \mathbb{R}^d$ 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 $\mathcal{Z}$ per some unknown distribution μ. We also let $S = \{Z_i\}_{i=1}^n$ be a training set consisting of n input-output data points drawn i.i.d. from $\mathcal{Z}$ according to μ.

Our goal is to learn a parameterized function, $f_W : \mathcal{X} \to \mathcal{Y}$, 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 $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}^+$ 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:

$$L_\mu(w) := \int_{\mathcal{Z}} \ell(w, z) d\mu(z), \tag{1}$$

$$L_E(w, s) := \frac{1}{n} \sum_{i=1}^{n} \ell(w, z_i), \tag{2}$$

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

$$\overline{\text{gen}}(P_{W|S}, \mu) := \mathbb{E}_{P_{W,S}}[\text{gen}(W, S, \mu)], \tag{3}$$

where $\text{gen}(w, s, \mu) := L_\mu(w) - L_E(w, s)$. This (expected) generalization error quantifies by how much the population risk deviates from the empirical risk. This quantity cannot be computed directly because μ 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 $P_{W|S}$,

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

Note that the excess risk can be decomposed as follows,

$$\begin{aligned}&\mathcal{E}_r(P_{W|S}, \mu)\\ &= \overline{\text{gen}}(P_{W|S}, \mu) + \mathbb{E}_{P_{W,S}}[L_E(W, S)] - \inf_{w\in\mathcal{W}} L_\mu(w),\end{aligned}$$

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 (μ and μ', respectively). In particular, we define the population risk based on test distribution μ' as,

$$L_P(w,\mu') \triangleq \int_{\mathcal{Z}} \ell(w,z) d\mu'(z). \tag{5}$$

We define the mismatched(expected) generalization error as

$$\overline{\text{gen}}(P_{W|S},\mu,\mu') \triangleq \mathbb{E}_{P_{W,S}}[\text{gen}(W,S,\mu,\mu')], \tag{6}$$

where $\text{gen}(w,s,\mu,\mu') \triangleq L_P(w,\mu') - L_E(w,s)$.

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 $f:\mathcal{X}\to\mathbb{R}$ 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:

$$|\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)]| \le F(T(P\|Q)), \tag{7}$$

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):

$$\begin{aligned}&|\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)]| \le \\ &|\mathbb{E}_P[f(X)] - \mathbb{E}_R[f(X)]| + |\mathbb{E}_Q[f(X)] - \mathbb{E}_R[f(X)]| \\ &\le F(T(P\|R)) + F(T(Q\|R))\end{aligned} \tag{8}$$

From concavity of F, we have

$$\begin{aligned}&F(T(P\|R)) + F(T(Q\|R)) \le \\ &2F\Big(T(P\|R)/2 + T(Q\|R)/2\Big)\end{aligned} \tag{9}$$

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

$$\min_{R\in\mathcal{P}(X)} T(P\|R) + T(Q\|R) \le T(P\|Q). \tag{10}$$

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

$$\begin{aligned}&|\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)]| \le \\ &2F\Big(T(P\|R^*)/2 + T(Q\|R^*)/2\Big).\end{aligned} \tag{11}$$

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

$$\begin{aligned}&|\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)]| \le \\ &|\mathbb{E}_R[f(X)] - \mathbb{E}_P[f(X)]| + |\mathbb{E}_R[f(X)] - \mathbb{E}_Q[f(X)]| \\ &\le F(T(R\|P)) + F(T(R\|Q)).\end{aligned} \tag{12}$$

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

$$\begin{aligned}&|\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)]| \le \\ &2F\Big(T(\tilde{R}\|P)/2 + T(\tilde{R}\|Q)/2\Big).\end{aligned} \tag{13}$$

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:

$$\alpha\text{KL}(P_{W,Z_i}\|\widehat{P}_{W,Z_i}) + (1-\alpha)\text{KL}(P_W\otimes\mu\|\widehat{P}_{W,Z_i}), \tag{14}$$

$$\alpha\text{KL}(\widehat{P}_{W,Z_i}\|P_{W,Z_i}) + (1-\alpha)\text{KL}(\widehat{P}_{W,Z_i}\|P_W\otimes\mu), \tag{15}$$

where $\widehat{P}_{W,Z_i}$, P_{W,Z_i} 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_i and the product of marginal distributions of random variables W and Z_i, 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, $\widehat{P}_{W,Z_i}$, 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 α-JS divergence[1], α-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

$$\Lambda_X(\lambda) := \log\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}]. \tag{16}$$

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

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

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

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 $\psi^\star(x)$ of $\psi(x)$ defined above is a

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

TABLE I
DIVERGENCE MEASURES DEFINITIONS

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

TABLE II
INFORMATION MEASURES DEFINITIONS

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

non-negative convex and non-decreasing function on $[0,\infty)$ with $\psi^\star(0)=0$. Moreover, its inverse function $\psi^{\star-1}(y) = \inf\{x \geq 0 : \psi^\star(x) \geq y\}$ is concave, and can be written as

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

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

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

The tail behaviour of sub-Exponential and sub-Gamma random variables are introduced in [20].

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. α-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, $\mathrm{KL}(P_W \otimes \mu\|\widehat{P}_{W,Z_i})$ and $\mathrm{KL}(P_{W,Z_i}\|\widehat{P}_{W,Z_i})$. All the proof details are deferred to Appendix A.

Theorem 1: Assume that under an auxiliary joint distribution $\widehat{P}_{W,Z_i} \in \mathcal{P}(\mathcal{W}\times\mathcal{Z}) - \Lambda_{\ell(W,Z_i)}(\lambda)$ 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 $\lambda \in (b_-,0]$, $\forall i = 1,\cdots,n$. 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:

$$\overline{\mathrm{gen}}(P_{W|S},\mu) \leq \frac{1}{n}\sum_{i=1}^n \left(\psi_+^{\star-1}(A_i) + \psi_-^{\star-1}(B_i)\right), \tag{20}$$

$$-\overline{\mathrm{gen}}(P_{W|S},\mu) \leq \frac{1}{n}\sum_{i=1}^n \left(\psi_-^{\star-1}(A_i) + \psi_+^{\star-1}(B_i)\right), \tag{21}$$

where $A_i = \mathrm{KL}(P_W\otimes\mu\|\widehat{P}_{W,Z_i})$, $B_i = \mathrm{KL}(P_{W,Z_i}\|\widehat{P}_{W,Z_i})$, $\psi_-^{\star-1}(x) = \inf_{\lambda\in[0,-b_-)} \frac{x+\psi_-(\lambda)}{\lambda}$ and $\psi_+^{\star-1}(x) = \inf_{\lambda\in[0,b_+)} \frac{x+\psi_+(\lambda)}{\lambda}$.

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 $\widehat{P}_{W,Z_i} = P_W \otimes \mu$ for $i = 1,\cdots,n$. It follows immediately from Theorem 1 that:

$$\overline{\mathrm{gen}}(P_{W|S},\mu) \leq \frac{1}{n}\sum_{i=1}^n \psi_-^{\star-1}(I(W;Z_i)), \tag{22}$$

$$-\overline{\text{gen}}(P_{W|S},\mu) \leq \frac{1}{n}\sum_{i=1}^{n}\psi_{+}^{\star-1}(I(W;Z_i)). \tag{23}$$

Example 2: Choose $\widehat{P}_{W,Z_i} = P_{W,Z_i}$ for $i = 1,\cdots,n$. It also follows immediately from Theorem 1 that:

$$\overline{\text{gen}}(P_{W|S},\mu) \leq \frac{1}{n}\sum_{i=1}^{n}\psi_{+}^{\star-1}(L(W;Z_i)), \tag{24}$$

$$-\overline{\text{gen}}(P_{W|S},\mu) \leq \frac{1}{n}\sum_{i=1}^{n}\psi_{-}^{\star-1}(L(W;Z_i)). \tag{25}$$

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.,

$$\alpha\text{KL}(P_W\otimes\mu\|\widehat{P}_{W,Z_i}) + (1-\alpha)\text{KL}(P_{W,Z_i}\|\widehat{P}_{W,Z_i}),$$

that relies on a certain σ-sub-Gaussian tail assumption.

Proposition 1: Assume that the loss function is $\hat{\sigma}$-sub-Gaussian– under the distribution $\widehat{P}_{W,Z_i}$ $\forall i = 1,\cdots,n$– Then, it holds $\forall\alpha\in(0,1)$ that:

$$|\overline{\text{gen}}(P_{W|S},\mu)| \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2\hat{\sigma}^2\frac{(\alpha A_i + (1-\alpha)B_i)}{\alpha(1-\alpha)}}, \tag{26}$$

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

We propose a Lemma connecting certain KL divergences to the α-JS information.

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

$$\alpha\text{KL}(P_W\otimes\mu\|\widehat{P}_{W,Z_i}) + (1-\alpha)\text{KL}(P_{W,Z_i}\|\widehat{P}_{W,Z_i}) = I_{\text{JS}}^{\alpha}(W;Z_i) + \text{KL}(P_{W,Z_i}^{(\alpha)}\|\widehat{P}_{W,Z_i}).$$

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 $P_{W,Z_i}^{(\alpha)}$ $\forall i = 1,\cdots,n$– Then, it holds $\forall\alpha\in(0,1)$ that:

$$|\overline{\text{gen}}(P_{W|S},\mu)| \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2\sigma_{(\alpha)}^2\frac{I_{\text{JS}}^{\alpha}(W;Z_i)}{\alpha(1-\alpha)}}, \quad \forall\alpha\in(0,1).$$

The bound in Theorem 2 results from minimizing the term $\alpha\text{KL}(P_W\otimes\mu\|\widehat{P}_{W,Z_i}) + (1-\alpha)\text{KL}(P_{W,Z_i}\|\widehat{P}_{W,Z_i})$, in the upper bound (26), presented in Proposition 1, over the joint auxiliary distribution $\widehat{P}_{W,Z_i}$. Such an optimal joint auxiliary distribution is $P_{W,Z_i}^{(\alpha)} := \alpha P_W P_{Z_i} + (1-\alpha)P_{W,Z_i}$. Note that, the parameter of sub-Gaussianity, denoted as $\hat{\sigma}$ in Proposition 1, relies on $\widehat{P}_{W,Z_i}$. 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 α.

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 $\widehat{P}_{W,Z_i} = P_W\otimes\mu$ in Lemma 2, leading up to

$$(1-\alpha)I(W;Z_i) = I_{\text{JS}}^{\alpha}(W;Z_i) + \text{KL}(P_{W,Z_i}^{(\alpha)}\|P_W\otimes\mu). \tag{27}$$

and in turn to the following inequality

$$I_{\text{JS}}^{\alpha}(W;Z_i) \leq (1-\alpha)I(W;Z_i), \quad \forall\alpha\in(0,1). \tag{28}$$

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 $\mathcal{O}(\frac{1}{\sqrt{n}})$.

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:

$$|\overline{\text{gen}}(P_{W|S},\mu)| \leq \sigma_{(\alpha)}\sqrt{2\frac{h(\alpha)}{\alpha(1-\alpha)}}, \quad \forall\alpha\in(0,1), \tag{29}$$

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 α-JS information generalization bound is always finite. We can also optimize the bound in (29) with respect to α, where the minimum is achieved at $\alpha = 1/2$.

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

$$|\overline{\text{gen}}(P_{W|S},\mu)| \leq 2\sigma_{(1/2)}\sqrt{2\log(2)}. \tag{30}$$

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 follows $|\overline{\text{gen}}(P_{W|S},\mu)| \leq (b-a)$ 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, Theorem. 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:

$$\mathrm{JSD}(P_X,P_{X'})=\int \mathrm{d}P_X f\left(\frac{\mathrm{d}P_{X'}}{\mathrm{d}P_X}\right), \tag{31}$$

with $f(t)=t\log(t)-(1+t)\log(\frac{1+t}{2})$. However, [11, Theorem. 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 $f(t)=t\log(t)-(1+t)\log(\frac{1+t}{2})$ associated with the Jensen-Shannon divergence.

B. α-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, $\mathrm{KL}(\widehat{P}_{W,Z_i}\|P_W\otimes\mu)$ and $\mathrm{KL}(\widehat{P}_{W,Z_i}\|P_{W,Z_i})$. All the proof details are deferred to Appendix B.

Proposition 4: Suppose that $\Lambda_{\ell(W,Z)}(\lambda)\leq\gamma_+(\lambda)$ and $\Lambda_{\ell(W,Z_i)}(\lambda)\leq\phi_+(\lambda),\quad i=1,\cdots,n$ 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_i}, resp. We also have $\Lambda_{\ell(W,Z)}(\lambda)\leq\gamma_-(-\lambda)$ and $\Lambda_{\ell(\widetilde{W},\widetilde{Z}_i)}(\lambda)\leq\phi_-(-\lambda),\quad i=1,\cdots,n$ 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_i}, resp. Assume that $\gamma_+(\lambda)$, $\phi_+(\lambda)$, $\gamma_-(\lambda)$ and $\phi_-(\lambda)$ are convex functions, $\gamma_-(0)=\gamma_+(0)=\gamma'_+(0)=\gamma'_-(0)=0$ and $\phi_-(0)=\phi_+(0)=\phi'_+(0)=\phi'_-(0)=0$. Then, the following upper bounds hold,

$$\overline{\mathrm{gen}}(P_{W|S},\mu)\leq\frac{1}{n}\sum_{i=1}^{n}\left(\gamma_-^{\star-1}(D_i)+\phi_+^{\star-1}(C_i)\right), \tag{32}$$

$$-\overline{\mathrm{gen}}(P_{W|S},\mu)\leq\frac{1}{n}\sum_{i=1}^{n}\left(\phi_-^{\star-1}(C_i)+\gamma_+^{\star-1}(D_i)\right), \tag{33}$$

where $D_i=\mathrm{KL}(\widehat{P}_{W,Z_i}\|P_W\otimes\mu)$, $C_i=\mathrm{KL}(\widehat{P}_{W,Z_i}\|P_{W,Z_i})$, $\gamma_-^{\star-1}(x)=\inf_{\lambda\in[0,-a_-)}\frac{x+\gamma_-(\lambda)}{\lambda}$, $\gamma_+^{\star-1}(x)=\inf_{\lambda\in[0,a_+)}\frac{x+\gamma_+(\lambda)}{\lambda}$, $\phi_-^{\star-1}(x)=\inf_{\lambda\in[0,-c_-)}\frac{x+\phi_-(\lambda)}{\lambda}$ and $\phi_+^{\star-1}(x)=\inf_{\lambda\in[0,c_+)}\frac{x+\phi_+(\lambda)}{\lambda}$.

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.,

$$\alpha\mathrm{KL}(\widehat{P}_{W,Z_i}\|P_{W,Z_i})+(1-\alpha)\mathrm{KL}(\widehat{P}_{W,Z_i}\|P_W\otimes\mu),$$

and assuming σ-sub-Gaussian tail distribution.

Proposition 5 (Upper bound with Sub-Gaussian assumption): Assume that the loss function is σ-sub-Gaussian under distribution $P_W\otimes\mu$ and γ-sub-Gaussian under P_{W,Z_i} $\forall i=1,\cdots,n$– Then, it holds for $\forall\alpha\in(0,1)$ that,

$$|\overline{\mathrm{gen}}(P_{W|S},\mu)|\leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2(\alpha\sigma^2+(1-\alpha)\gamma^2)\frac{(\alpha C_i+(1-\alpha)D_i)}{\alpha(1-\alpha)}}, \tag{34}$$

where $C_i=\mathrm{KL}(\widehat{P}_{W,Z_i}\|P_W\otimes\mu)$ and $D_i=\mathrm{KL}(\widehat{P}_{W,Z_i}\|P_{W,Z_i})$.

Akin to Proposition 1, the result in Proposition 5 paves the way to offer new tighter expected generalization error upper bound by ADM. We next offer a Lemma connecting certain KL divergences to the α-Rényi information [35, Theorem 30].

Lemma 3: Consider an arbitrary distribution $\widehat{P}_{W,Z_i}$. Then, the following equality holds for $\forall\alpha\in(0,1)$,

$$\begin{aligned}\alpha\mathrm{KL}(\widehat{P}_{W,Z_i}\|P_W\otimes\mu)+(1-\alpha)\mathrm{KL}(\widehat{P}_{W,Z_i}\|P_{W,Z_i})=\ &(1-\alpha)I_{\mathrm{R}}^{\alpha}(W;Z_i)\\ &+\mathrm{KL}\left(\widehat{P}_{W,Z_i}\|\frac{(P_{Z_i}\otimes P_W)^{\alpha}(P_{W,Z_i})^{(1-\alpha)}}{\int_{\mathcal{W}\times\mathcal{Z}}(\mathrm{d}P_{Z_i}\otimes\mathrm{d}P_W)^{\alpha}(\mathrm{d}P_{W,Z_i})^{(1-\alpha)}}\right).\end{aligned} \tag{35}$$

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 α-Rényi information): Consider the same assumptions in Proposition 5. The following upper bound for $\forall\alpha\in(0,1)$ holds,

$$|\overline{\mathrm{gen}}(P_{W|S},\mu)|\leq\frac{1}{n}\sum_{i=1}^{n}\sqrt{2(\alpha\sigma^2+(1-\alpha)\gamma^2)\frac{I_{\mathrm{R}}^{\alpha}(W;Z_i)}{\alpha}}. \tag{36}$$

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

$$\widehat{P}_{W,Z_i}=\frac{(P_{Z_i}\otimes P_W)^{\alpha}(P_{W,Z_i})^{(1-\alpha)}}{\int_{\mathcal{W}\times\mathcal{Z}}(\mathrm{d}P_{Z_i}\otimes\mathrm{d}P_W)^{\alpha}(\mathrm{d}P_{W,Z_i})^{(1-\alpha)}}.$$

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_i} is not *absolutely continuous*[2] with respect to $P_WP_{Z_i}$. However, by considering the α-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,

$$|\overline{\mathrm{gen}}(P_{W|S},\mu)|\leq\frac{2}{n}\sum_{i=1}^{n}\sqrt{(\sigma^2+\gamma^2)D_B(P_{W,Z_i}\|P_W\otimes\mu)},$$

Remark 5 (Recovering the upper bound based on mutual information and lautum information): We can recover the

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

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 $\widehat{P}_{W,Z_i} = P_{W,Z_i}$, we have,

$$\alpha I(W;Z_i) = (1-\alpha) I_{\mathrm{R}}^{\alpha}(W;Z_i) + \mathrm{KL}\left(P_{W,Z_i} \| \frac{(P_{Z_i} \otimes P_W)^{\alpha} (P_{W,Z_i})^{(1-\alpha)}}{\int_{\mathcal{W}\times\mathcal{Z}} (\mathrm{d}P_{Z_i} \otimes \mathrm{d}P_W)^{\alpha} (\mathrm{d}P_{W,Z_i})^{(1-\alpha)}} \right). \tag{37}$$

Since that KL divergence is non-negative, based on Lemma 3 and the monotonicity of R_α with respect to α, we have,

$$I_{\mathrm{R}}^{\alpha}(W;Z_i) \leq \min\left\{1, \frac{\alpha}{1-\alpha}\right\} I(W;Z_i). \tag{38}$$

The result in (38) implies that our expected generalization error bound based on α-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 α-Rényi information): Assume the hypothesis space is finite and the data samples are i.i.d. Then, the upper bounds based on α-Rényi information in Theorem 3 have a convergence rate of $\mathcal{O}(\frac{1}{\sqrt{n}})$.

We can also provide an upper bound based on Sibson's α-mutual information.

Theorem 4 (Upper bound based on Sibson's α mutual information): Assume that the loss function is σ-sub-Gaussian under distribution μ for all $w \in \mathcal{W}$ and γ-sub-Gaussian under P_{W,Z_i}, $\forall i = 1, \cdots, n$. Then, it holds that:

$$|\overline{\mathrm{gen}}(P_{W|S}, \mu)| \leq \frac{1}{n} \sum_{i=1}^{n} \sqrt{2(\alpha\sigma^2 + (1-\alpha)\gamma^2) \frac{I_{\mathrm{S}}^{\alpha}(W;Z_i)}{\alpha}}.$$

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

$$I_{\mathrm{S}}^{\alpha}(W;Z_i) = \min_{Q_W \in \mathcal{P}(\mathcal{W})} \mathrm{R}_{\alpha}(P_{W,Z_i} \| Q_W \otimes \mu) \tag{39}$$

$$\leq \mathrm{R}_{\alpha}(P_{W,Z_i} \| P_W \otimes \mu) = I_{\mathrm{R}}^{\alpha}(W;Z_i), \tag{40}$$

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 σ-sub-Gaussian under $P_W \otimes \mu$ distribution in Theorem 3. However, in Theorem 4, we consider the loss function is σ-sub-Gaussian under μ distribution for all $w \in \mathcal{W}$.

We can also apply generalized Pinsker's inequality [35] to bounded loss functions for bounding the expected generalization error using the α-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

$$|\overline{\mathrm{gen}}(P_{W|S}, \mu)| \leq \frac{1}{n} \sum_{i=1}^{n} \sqrt{\frac{2b^2}{\alpha} I_{\mathrm{R}}^{\alpha}(W;Z_i)}, \quad \forall \alpha \in (0,1]. \tag{41}$$

Considering the bounded loss function can help to provide an upper bound based on α-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 σ-sub-Gaussian assumptions is provided in Table III.

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

$$\sigma = \gamma = \sigma_{(\alpha)} = \frac{(b-a)}{2}. \tag{42}$$

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

Proposition 8 (Comparison of upper bounds based on α'-Jensen-Shannon and α-Rényi information measures): Consider the same assumptions in Theorem 2. Then, it follows that α'-Jensen-Shannon bound given by:

$$|\overline{\mathrm{gen}}(P_{W|S}, \mu)| \leq \frac{1}{n} \sum_{i=1}^{n} \sqrt{2\sigma_{(\alpha')}^2 \frac{I_{\mathrm{JS}}^{\alpha'}(W;Z_i)}{\alpha'(1-\alpha')}}, \quad 0 \leq \alpha' \leq 1 \tag{43}$$

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

$$|\overline{\mathrm{gen}}(P_{W|S}, \mu)| \leq \frac{1}{n} \sum_{i=1}^{n} \sqrt{2(\alpha\sigma^2 + (1-\alpha)\gamma^2) \frac{I_{\mathrm{R}}^{\alpha}(W;Z_i)}{\alpha}}, \tag{44}$$

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

Remark 7: The condition in Proposition 8, i.e. $\frac{\alpha h(\alpha')}{(1-\alpha')\alpha'} \leq I_{\mathrm{R}}^{\alpha}(W;Z_i)$, 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 [8] provided that $4\log(2) \leq I(W;Z_i)$ for all $i = 1, \cdots, n$ 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 α-Rényi divergence or α-JS divergence.

A. α-JS-Regularized ERM

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

$$\min_{P_{W|S}} \mathbb{E}[L_E(W,S)] + \frac{1}{\beta} I_{\mathrm{JS}}^{\alpha}(W;S), \tag{45}$$

TABLE III
EXPECTED GENERALIZATION ERROR UPPER BOUNDS. WE COMPARED OUR BOUNDS WITH MUTUAL INFORMATION AND LAUTUM INFORMATION BOUNDS BASED ON THE FINITENESS AND THE ASSUMPTION NEEDED FOR SUB-GAUSSIANITY.

Upper Bound Measure	sub-Gaussian Assumption	Bound	Is finite?
Mutual information ([8])	$P_W \otimes \mu$	$\frac{1}{n}\sum_{i=1}^{n}\sqrt{2\sigma^2 I(W;Z_i)}$	No
Lautum information ([41])	P_{W,Z_i}, $\forall i=1,\ldots,n$	$\frac{1}{n}\sum_{i=1}^{n}\sqrt{2\gamma^2 L(W;Z_i)}$	No
α-JS information (Proposition 3)	$P^{(\alpha)}_{W,Z_i}$, $\forall i=1,\ldots,n$	$\frac{1}{n}\sum_{i=1}^{n}\sqrt{2\sigma^2_{(\alpha)}\frac{I^{\alpha}_{\mathrm{JS}}(W;Z_i)}{\alpha(1-\alpha)}}$	Yes $(\sigma_{(\alpha)}\sqrt{2\frac{h(\alpha)}{\alpha(1-\alpha)}})$
α-Rényi information ($0\le\alpha<1$) (Theorem 3)	$P_W\otimes\mu$ and P_{W,Z_i}, $\forall i=1,\ldots,n$	$\frac{1}{n}\sum_{i=1}^{n}\sqrt{2(\alpha\sigma^2+(1-\alpha)\gamma^2)\frac{I^{\alpha}_{\mathrm{R}}(W;Z_i)}{\alpha}}$	No

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

$$\min_{P_{W|S}} \mathbb{E}[L_E(W,S)] + \frac{1}{\beta}\mathrm{JS}_\alpha(P_{W|S}\|Q_W|P_S), \tag{46}$$

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

Lemma 4 (Solution existence of α-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 term of $P_{W|S}$ and the second term $\frac{1}{\beta}\mathrm{JS}_\alpha(P_{W|S}\|Q_W|P_S)$ is convex in $P_{W|S}$ for $0<\alpha<1$ due to [45]. Therefore, a solution exists. ∎

Let us define the solution of (45),

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

In the following, we provide an upper bound on excess risk under $P^{\star,\beta,\mathrm{JS}_\alpha}_{W|S}$ as a learning algorithm.

Theorem 5 (Upper bound on excess risk under $P^{\star,\beta,\mathrm{JS}_\alpha}_{W|S}$): Assume the bounded loss function, i.e., $|\ell(w,z)|\le b$ for all $(w,z)\in\mathcal{W}\times\mathcal{Z}$ and $\tilde{L}$-Lipschitz. Then, the following upper bound holds on the excess risk under $P^{\star,\beta,\mathrm{JS}_\alpha}_{W|S}$,

$$\mathcal{E}_r(P^{\star,\beta,\mathrm{JS}_\alpha}_{W|S},\mu) \le \sqrt{\frac{2b^2}{n\alpha(1-\alpha)}\sum_{i=1}^{n} I^{\alpha}_{\mathrm{JS}}(W;Z_i)} + \frac{\tilde{L}\sqrt{d}}{\beta} + \frac{\mathrm{JS}_\alpha(\mathcal{N}(w^\star,\beta^{-1}I_d)\|Q)}{\beta},$$

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

Corollary 2 (Convergence rate of excess risk for under $P^{\star,\beta,\mathrm{JS}_\alpha}_{W|S}$): Under the same assumptions in Theorem 5, assuming that hypothesis space is finite and β is of order $\sqrt{n}$, the following upper bound holds on excess risk of $P^{\star,\beta,\mathrm{JS}_\alpha}_{W|S}$ with convergence rate of $\mathcal{O}(n^{-1/2})$,

$$\mathcal{E}_r(P^{\star,\beta,\mathrm{JS}_\alpha}_{W|S},\mu) \le \sqrt{\frac{2b^2}{n\alpha(1-\alpha)}\sum_{i=1}^{n} I^{\alpha}_{\mathrm{JS}}(W;Z_i)} + \frac{\tilde{L}\sqrt{d}}{\sqrt{n}} + \frac{h(\alpha)}{\sqrt{n}},$$

Remark 9 (Comparison to the Gibbs algorithm): Our convergence rate of the upper bound on the excess risk under $P^{\star,\beta,\mathrm{JS}_\alpha}_{W|S}$ 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 $\mathcal{O}(n^{-1/4})$, [7, Corollary 3] and [46].

B. α-Rényi-Regularized ERM

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

$$\min_{P_{W|S}} \mathbb{E}[L(W,S)] + \frac{1}{\beta}I^{\alpha}_{\mathrm{R}}(W;S), \tag{47}$$

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

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

$$\min_{P_{W|S}} \mathbb{E}[L_E(W,S)] + \frac{1}{\beta}\mathrm{R}_\alpha(P_{W|S}\|Q_W|P_S), \tag{48}$$

where $Q_W\in\mathcal{P}(\mathcal{W})$.

Lemma 5 (Solution existence of α-Rényi-regularized ERM): The optimization problem considered in (48) is a convex optimization problem.

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

Let us define

$$P^{\star,\beta,\mathrm{R}_\alpha}_{W|S} := \arg\min_{P_{W|S}\in\mathcal{P}(\mathcal{W})} \mathbb{E}[L_E(W,S)] + \frac{1}{\beta}\mathrm{R}_\alpha(P_{W|S}\|Q_W|P_S),$$

as the solution of convex optimization problem (48).

Theorem 6 (Upper bound on excess risk under $P^{\star,\beta,\mathrm{R}_\alpha}_{W|S}$): Assume the bounded loss function, i.e., $|\ell(w,z)| \leq b$ for all $(w,z) \in \mathcal{W} \times \mathcal{Z}$ and $\tilde{L}$-Lipschitz. Then, the following upper bound holds on the excess risk under $P^{\star,\beta,\mathrm{R}_\alpha}_{W|S}$,

$$\mathcal{E}_r(P^{\star,\beta,\mathrm{R}_\alpha}_{W|S},\mu) \leq \sqrt{\frac{2b^2}{n\alpha}\sum_{i=1}^{n} I^{\alpha}_{\mathrm{R}}(W;Z_i)} + \frac{\tilde{L}\sqrt{d}}{\beta} + \frac{\mathrm{R}_\alpha(\mathcal{N}(w^\star,\beta^{-1}I_d)\|Q)}{\beta},$$

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

Corollary 3 (Convergence rate of excess risk under $P^{\star,\beta,\mathrm{R}_\alpha}_{W|S}$): Under the same assumptions in Theorem 6, assuming that hypothesis space is finite and β is of order $\sqrt{n}$, the following upper bound holds on the excess risk of $P^{\star,\beta,\mathrm{R}_\alpha}_{W|S}$ with convergence rate of $\mathcal{O}(\log(n)/\sqrt{n})$,

$$\mathcal{E}_r(P^{\star,\beta,\mathrm{R}_\alpha}_{W|S},\mu) \leq \sqrt{\frac{2b^2}{n\alpha}\sum_{i=1}^{n} I^{\alpha}_{\mathrm{R}}(W;Z_i)} + \frac{\tilde{L}\sqrt{d}}{\sqrt{n}} + \frac{1}{2\sqrt{n}}\|w^\star\|_2^2 + \frac{d}{4\sqrt{n}}\log(n) + \frac{d}{2\sqrt{n}(1-\alpha)}\log(\alpha).$$

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.

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

$$|\overline{\mathrm{gen}}(P_{W|S},\mu,\mu')| \leq \sqrt{2\sigma^2_{(\alpha)}\frac{\mathrm{JS}_\alpha(\mu'\|\mu)}{\alpha(1-\alpha)}} + \frac{1}{n}\sum_{i=1}^{n}\sqrt{2\sigma^2_{(\alpha)}\frac{I^{\alpha}_{\mathrm{JS}}(W;Z_i)}{\alpha(1-\alpha)}}, \quad \forall\alpha\in(0,1). \tag{49}$$

Proposition 10: Assume that the loss function is σ-sub-Gaussian under distributions μ and μ' for all $w \in \mathcal{W}$ and also γ-sub-Gaussian under P_{W,Z_i} $\forall i = 1,\cdots,n$. The following upper bound for $\forall\alpha \in (0,1)$ holds,

$$|\overline{\mathrm{gen}}(P_{W|S},\mu,\mu')| \leq \sqrt{2(\alpha\sigma^2+(1-\alpha)\gamma^2)\frac{\mathrm{R}_\alpha(\mu'\|\mu)}{\alpha}} + \frac{1}{n}\sum_{i=1}^{n}\sqrt{2(\alpha\sigma^2+(1-\alpha)\gamma^2)\frac{I^{\alpha}_{\mathrm{R}}(W;Z_i)}{\alpha}}. \tag{50}$$

The mismatch between the test and training samples distributions is characterised in [21, Theorem 5] as KL divergence between test and training samples distributions, i.e., $\mathrm{KL}(\mu'\|\mu)$. However, assuming that the loss function is $\sigma_{(\alpha)}$-sub-Gaussian under $\alpha\mu + (1-\alpha)\mu'$ for all $w \in \mathcal{W}$, Proposition 9 allows us to explain the distributional mismatch in terms of α-Jensen-Shannon divergence, which is finite.

In Proposition 10, the distributional mismatch is presented in terms of α-Rényi divergence, i.e., $\mathrm{R}_\alpha(\mu'\|\mu)$. If μ' is not absolutely continuous with respect to μ, then we have $\mathrm{KL}(\mu'\|\mu) = \infty$. However, for α-Rényi divergence ($0 < \alpha < 1$), it suffices that the mutual singularity [35], i.e., $\mu' \perp\!\!\!\perp \mu$, does not hold, which is a less restrictive condition about μ' compared to the absolutely continuity condition.

Similar to Remark 6, the sub-Gaussianity assumptions in Propositions 9 and 10 hold for bounded loss functions.

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 α-JS and α-Rényi information only. Our example setting involves the estimation of 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 hypothesis (estimate) given by $W = tZ_1 + (1-t)Z_2$ for $0 < t < 1$. 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, α-JS information and α-Rényi information $\forall\alpha \in (0,1)$ as follows:

$$\overline{\mathrm{gen}}(P_{W|Z_1,Z_2},P_Z) \leq \frac{c^2}{4}\left(\sqrt{2I(W;Z_1)} + \sqrt{2I(W;Z_2)}\right), \tag{51}$$

$$\overline{\mathrm{gen}}(P_{W|Z_1,Z_2},P_Z) \leq \frac{c^2}{4}\left(\sqrt{2\frac{I^{\alpha}_{\mathrm{JS}}(W;Z_1)}{\alpha(1-\alpha)}} + \sqrt{2\frac{I^{\alpha}_{\mathrm{JS}}(W;Z_2)}{\alpha(1-\alpha)}}\right), \tag{52}$$

$$\overline{\mathrm{gen}}(P_{W|Z_1,Z_2},P_Z) \leq \frac{c^2}{4}\left(\sqrt{2\frac{I^{\alpha}_{R}(W;Z_1)}{\alpha}} + \sqrt{2\frac{I^{\alpha}_{R}(W;Z_2)}{\alpha}}\right). \tag{53}$$

It can be immediately 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 $\rho_2 = \frac{(1-t)}{\sqrt{t^2+(1-t)^2}}$. Therefore, it can be shown that the mutual information appearing above is given by $I(W;Z_1) = -\frac{1}{2}\log(1-\rho_1^2)$ and $I(W;Z_2) = -\frac{1}{2}\log(1-\rho_2^2)$. In contrast, the α-JS information appearing above can be computed via an extension of entropic-based formulation of the Jensen-Shannon measure as follows [34]:

$$I_{\mathrm{JS}}(W;Z_i) = h\left(P^{(\alpha)}_{W,Z_i}\right) - (\alpha h(P_W) + \alpha h(P_{Z_i}) + (1-\alpha)h(P_{Z_i,W})), \tag{54}$$

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

$$h(P_{Z_i}) = \frac{1}{2}\log(2\pi\sigma^2),$$
$$h(P_W) = \frac{1}{2}\log(2\pi\sigma^2(t^2+(1-t)^2)),$$
$$h(P_{W,Z_i}) = \log(2\pi\sigma^2(t^2+(1-t)^2)(1-\rho_i^2)),$$

whereas $h\left(P^{(\alpha)}_{W,Z_i}\right)$ can be computed numerically.

Fig.1 depicts the true generalization error, the mutual information based bound in (51), and the α-JS information based bound for $\alpha = 0.25, 0.5, 0.75$ in (52) for values of $t \in (0, 0.5]$, considering $\sigma^2 = 1, 10$, $\mu = 1$, $c = \frac{\sigma}{4}$.

It can be seen that for $\alpha = 0.75$ the α-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,

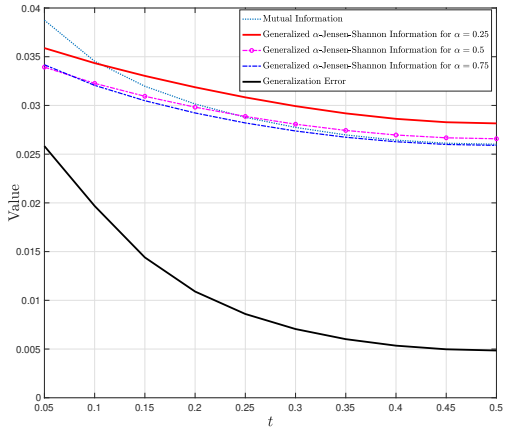

Fig. 1. True generalization error, α-JS based bound for $\alpha = 0.25, 0.5, 0.75$, and Mutual Information based bound.

the mutual information based bound in (51), and the α-Rényi information based bound for $\alpha = 0.25, 0.5, 0.75$ in (53). It can be seen that the α-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 $I(W; Z_2) \to \infty$ (or $I(W; Z_1) \to \infty$). However, the α-Rényi based bound is finite.

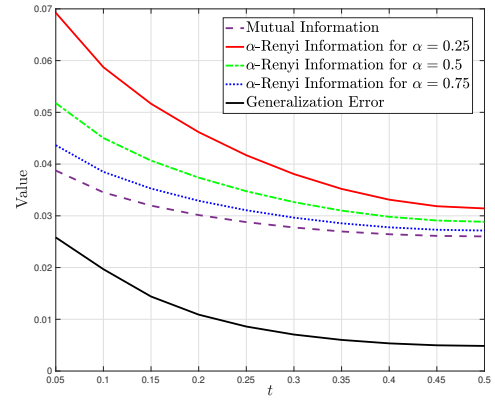

Fig. 2. True generalization error, α-Rényi based bound for $\alpha = 0.25, 0.5, 0.75$, and Mutual Information based bound.

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 α-JS and α-Rényi information measures. Notably, our upper bounds, which are rooted in the α-JS information measure, are finite, in contrast to mutual information-based bounds. Moreover, our upper bound based on α-Rényi information, for $\alpha \in (0, 1)$, remains finite when considering a deterministic learning process. An intriguing observation is that our newly introduced α-JS information measure can, in certain regimes, yield tighter bounds compared to existing approaches. We also discuss the existence of algorithms under α-JS-regularized and α-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 α-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 α-JS and α-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.

Acknowledgment

Gholamali Aminian is supported in part by the Royal Society Newton International Fellowship, grant no. NIF\R1 \192656, the UKRI Prosperity Partnership Scheme (FAIR) under the EPSRC Grant EP/V056883/1, and the Alan Turing Institute. Saeed Masiha worked on this project at Sharif university of Technology before joining EPFL University.

Appendix A
Proof of Section III-A

Proof of Theorem 1: The proofs of the bounds to $\overline{\text{gen}}(P_{W|S}, \mu)$ and $-\overline{\text{gen}}(P_{W|S}, \mu)$ are similar. Therefore, we focus on the latter.

Let us consider the Donsker–Varadhan variational representation of KL divergence between two probability distributions α and β on a common space Ψ given by [47]:

$$\text{KL}(\alpha \| \beta) = \sup_f \int_\Psi f d\alpha - \log \int_\Psi e^f d\beta, \tag{55}$$

where $f \in \mathcal{F} = \{f : \Psi \to \mathbb{R} \text{ s.t. } \mathbb{E}_\beta[e^f] < \infty\}$.

Using the Donsker-Varadhan representation to bound $\text{KL}(P_{W,Z_i} \| \widehat{P}_{W,Z_i})$ for $\lambda \in (b_-, 0]$ as follows:

$$\text{KL}(P_{W,Z_i} \| \widehat{P}_{W,Z_i}) \geq \tag{56}$$
$$\mathbb{E}_{P_{W,Z_i}}[\lambda \ell(W, Z_i)] - \log \mathbb{E}_{\widehat{P}_{W,Z_i}}[e^{\lambda \ell(W,Z_i)}] \geq$$
$$\lambda(\mathbb{E}_{P_{W,Z_i}}[\ell(W, Z_i)] - \mathbb{E}_{\widehat{P}_{W,Z_i}}[\ell(W, Z_i)]) - \psi_-(-\lambda), \tag{57}$$

where the last inequality is due to:

$$\Lambda_{\ell(W,Z_i)}(\lambda) = \log\left(\mathbb{E}_{\widehat{P}_{W,Z_i}}[e^{\lambda(\ell(W,Z_i)-\mathbb{E}_{\widehat{P}_{W,Z_i}}[\ell(W,Z_i)])}]\right) \leq \psi_-(-\lambda). \tag{58}$$

It can then be shown from (57) that the following holds for $\lambda \in (b_-, 0]$:

$$\mathbb{E}_{\widehat{P}_{W,Z_i}}[\ell(W,Z_i)] - \mathbb{E}_{P_{W,Z_i}}[\ell(W,Z_i)] \leq \inf_{\lambda\in[0,-b_-)} \frac{\mathrm{KL}(P_{W,Z_i}\|\widehat{P}_{W,Z_i}) + \psi_-(\lambda)}{\lambda} = \tag{59}$$

$$\psi_-^{\star-1}(\mathrm{KL}(P_{W,Z_i}\|\widehat{P}_{W,Z_i})). \tag{60}$$

It can likewise also be shown by adopting similar steps that the following holds for $\lambda \in [0, b_+)$:

$$\mathbb{E}_{P_{W,Z_i}}[\ell(W,Z_i)] - \mathbb{E}_{\widehat{P}_{W,Z_i}}[\ell(W,Z_i)] \leq \inf_{\lambda\in[0,b_+)} \frac{\mathrm{KL}(P_{W,Z_i}\|\widehat{P}_{W,Z_i}) + \psi(\lambda)}{\lambda} = \tag{61}$$

$$\psi_+^{\star-1}(\mathrm{KL}(P_{W,Z_i}\|\widehat{P}_{W,Z_i})). \tag{62}$$

We can similarly show using an identical procedure that:

$$\mathbb{E}_{P_W\otimes\mu}[\ell(W,Z_i)] - \mathbb{E}_{\widehat{P}_{W,Z_i}}[\ell(W,\widehat{Z}_i)] \leq \psi_+^{\star-1}(\mathrm{KL}(P_W\otimes\mu\|\widehat{P}_{W,Z_i})) \tag{63}$$

$$\mathbb{E}_{\widehat{P}_{W,Z_i}}[\ell(W,\widehat{Z}_i)] - \mathbb{E}_{P_W\otimes\mu}[\ell(W,Z_i)] \leq \psi_-^{\star-1}(\mathrm{KL}(P_W\otimes\mu\|\widehat{P}_{W,Z_i})). \tag{64}$$

Finally, we can immediately bound the expected generalization error by leveraging (63) and (59) as follows:

$$\begin{aligned}
\overline{\mathrm{gen}}(P_{W|S},\mu) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{P_W\otimes\mu}[\ell(W,Z_i)] - \mathbb{E}_{P_{W,Z_i}}[\ell(W,Z_i)] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{P_W\otimes\mu}[\ell(W,Z_i)] - \mathbb{E}_{\widehat{P}_{W,Z_i}}[\ell(W,Z_i)] + \\
&\mathbb{E}_{\widehat{P}_{W,Z_i}}[\ell(W,Z_i)] - \mathbb{E}_{P_{W,Z_i}}[\ell(W,Z_i)] \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left(\psi_+^{\star-1}(A_i) + \psi_-^{\star-1}(B_i)\right),
\end{aligned}$$

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

■

Proof of Proposition 1:

The assumption that the loss function is σ-sub-Gaussian under the distribution $\widehat{P}_{W,Z_i}$ implies that $\psi_-^{\star-1}(y) = \psi_+^{\star-1}(y) = \sqrt{2\sigma^2 y}$, [8].

Consider arbitrary auxiliary distributions $\{\widehat{P}_{W,Z_i}\}_{i=1}^n$ defined on $\mathcal{W}\times\mathcal{Z}$.

$$\begin{aligned}
\overline{\mathrm{gen}}(\mu,P_{W|S}) &= \mathbb{E}_{P_WP_S}[L_E(W,S)] - \mathbb{E}_{P_{W,S}}[L_E(W,S)] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{P_WP_{Z_i}}[\ell(W,Z_i)] - \mathbb{E}_{P_WZ_i}[\ell(W,Z_i)] \qquad (65)\\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{E}_{P_WP_{Z_i}}[\ell(W,Z_i)] - \mathbb{E}_{P_WZ_i}[\ell(W,Z_i)]\right| \qquad (66)
\end{aligned}$$

Using the assumption that the loss function $\ell(w,z_i)$ is $\hat{\sigma}^2$-sub-Gaussian under distribution $\widehat{P}_{W,Z_i}$ and Donsker-Varadhan representation for $\mathrm{KL}(P_{WZ_i}\|\widehat{P}_{W,Z_i})$, we have:

$$\lambda\left(\mathbb{E}_{P_{W,Z_i}}[\ell(W,Z_i)] - \mathbb{E}_{\widehat{P}_{W,Z_i}}[\ell(W,Z_i)]\right) \leq \mathrm{KL}(P_{WZ_i}\|\widehat{P}_{W,Z_i}) + \frac{\lambda^2\hat{\sigma}^2}{2}. \quad \forall\lambda\in\mathbb{R} \tag{67}$$

Using the assumption loss that the function $\ell(w,z_i)$ is $\hat{\sigma}^2$-sub-Gaussian under distribution $\widehat{P}_{W,Z_i}$ and Donsker-Varadhan representation for $\mathrm{KL}(\widehat{P}_{W,Z_i}\|P_WP_{Z_i})$, we have:

$$\lambda'\left(\mathbb{E}_{P_WP_{Z_i}}[\ell(W,Z_i)] - \mathbb{E}_{\widehat{P}_{W,Z_i}}[\ell(W,Z_i)]\right) \leq \mathrm{KL}(P_WP_{Z_i}\|\widehat{P}_{W,Z_i}) + \frac{\lambda'^2\hat{\sigma}^2}{2}. \quad \forall\lambda'\in\mathbb{R} \tag{68}$$

Now if we consider $\lambda < 0$, then we can choose $\lambda' = \frac{\alpha}{\alpha-1}\lambda$. Hence we have:

$$\mathbb{E}_{\widehat{P}_{W,Z_i}}[\ell(W,Z_i)] - \mathbb{E}_{P_{W,Z_i}}[\ell(W,Z_i)] \leq \frac{\mathrm{KL}(P_{W,Z_i}\|\widehat{P}_{W,Z_i})}{|\lambda|} + \frac{|\lambda|\hat{\sigma}^2}{2}. \quad \forall\lambda\in\mathbb{R}^- \tag{69}$$

and,

$$\mathbb{E}_{P_WP_{Z_i}}[\ell(W,Z_i)] - \mathbb{E}_{\widehat{P}_{W,Z_i}}[\ell(W,Z_i)] \leq \frac{\mathrm{KL}(P_WP_{Z_i}\|\widehat{P}_{W,Z_i})}{\lambda'} + \frac{\lambda'\hat{\sigma}^2}{2}. \quad \forall\lambda'\in\mathbb{R}^+ \tag{70}$$

Now sum up two Inequalities (69) and (70).

$$\mathbb{E}_{P_WP_{Z_i}}[\ell(W,Z_i)] - \mathbb{E}_{P_WZ_i}[\ell(W,Z_i)] \leq \frac{\alpha\mathrm{KL}(P_{W,Z_i}\|\widehat{P}_{W,Z_i}) + (1-\alpha)\mathrm{KL}(P_WP_{Z_i}\|\widehat{P}_{W,Z_i})}{\alpha|\lambda|} + \frac{|\lambda|\hat{\sigma}^2}{2} + \frac{|\lambda|\frac{\alpha}{1-\alpha}\hat{\sigma}^2}{2}, \quad \forall\lambda\in\mathbb{R}^-. \tag{71}$$

Similarly, using an identical approach, we also obtain:

$$-\left(\mathbb{E}_{P_WP_{Z_i}}[\ell(W,Z_i)] - \mathbb{E}_{P_WZ_i}[\ell(W,Z_i)]\right) \leq \frac{\alpha\mathrm{KL}(P_{W,Z_i}\|\widehat{P}_{W,Z_i}) + (1-\alpha)\mathrm{KL}(P_WP_{Z_i}\|\widehat{P}_{W,Z_i})}{\alpha\lambda} + \frac{\lambda\hat{\sigma}^2}{2} + \frac{\lambda\frac{\alpha}{1-\alpha}\hat{\sigma}^2}{2}, \quad \forall\lambda\in\mathbb{R}^+. \tag{72}$$

Considering (71) and (72), we have a nonnegative parabola in λ, whose discriminant must be nonpositive, and we have $\forall\alpha\in(0,1)$:

$$\left|\mathbb{E}_{P_WP_{Z_i}}[\ell(W,Z_i)] - \mathbb{E}_{P_WZ_i}[\ell(W,Z_i)]\right|^2 \leq 2\hat{\sigma}^2\frac{\left(\alpha\mathrm{KL}(P_{W,Z_i}\|\widehat{P}_{W,Z_i}) + (1-\alpha)\mathrm{KL}(P_WP_{Z_i}\|\widehat{P}_{W,Z_i})\right)}{\alpha(1-\alpha)}. \tag{73}$$

Now using (65), we prove the claim.

■

Proof of Lemma 2:

$$\begin{aligned}
&\alpha\mathrm{KL}(P_W\otimes P_{Z_i}\|\widehat{P}_{W,Z_i}) + (1-\alpha)\mathrm{KL}(P_{W,Z_i}\|\widehat{P}_{W,Z_i}) \qquad (74)\\
&= \int_{\mathcal{W}\times\mathcal{Z}}\alpha(\mathrm{d}P_W\otimes\mathrm{d}P_{Z_i})\log(\mathrm{d}P_W\otimes\mathrm{d}P_{Z_i}) \qquad (75)
\end{aligned}$$

$$
\begin{aligned}
&+\int_{\mathcal{W}\times\mathcal{Z}}(1-\alpha)\mathrm{d}P_{W,Z_i}\log(\mathrm{d}P_{W,Z_i})\\
&-\int_{\mathcal{W}\times\mathcal{Z}}((\alpha(\mathrm{d}P_W\otimes\mathrm{d}P_{Z_i})+(1-\alpha)\mathrm{d}P_{W,Z_i})\log(d\widehat{P}_{W,Z_i}))\\
&=\int_{\mathcal{W}\times\mathcal{Z}}\alpha(\mathrm{d}P_W\otimes\mathrm{d}P_{Z_i})\log(\mathrm{d}P_W\otimes\mathrm{d}P_{Z_i}) \qquad (76)\\
&+\int_{\mathcal{W}\times\mathcal{Z}}(1-\alpha)\mathrm{d}P_{W,Z_i}\log(\mathrm{d}P_{W,Z_i})-\mathrm{d}P^{(\alpha)}_{W,Z_i}\log(d\widehat{P}_{W,Z_i})\\
&+\int_{\mathcal{W}\times\mathcal{Z}}\mathrm{d}P^{(\alpha)}_{W,Z_i}\log(\mathrm{d}P^{(\alpha)}_{W,Z_i})-\mathrm{d}P^{(\alpha)}_{W,Z_i}\log(\mathrm{d}P^{(\alpha)}_{W,Z_i})\\
&=I^{\alpha}_{\mathrm{JS}}(W;Z_i)+\mathrm{KL}(P^{(\alpha)}_{W,Z_i}\|\widehat{P}_{W,Z_i}). \qquad (77)
\end{aligned}
$$

■

Proof of Theorem 2: As shown in [48], and by considering the Lemma 2 we have

$$
\begin{aligned}
&\min_{\widehat{P}_{W,Z_i}}\alpha\mathrm{KL}(P_W\otimes\mu\|\widehat{P}_{W,Z_i})+(1-\alpha)\mathrm{KL}(P_{W,Z_i}\|\widehat{P}_{W,Z_i})= \qquad (78)\\
&\min_{\widehat{P}_{W,Z_i}} I^{\alpha}_{\mathrm{JS}}(W;Z_i)+\mathrm{KL}(P^{(\alpha)}_{W,Z_i}\|\widehat{P}_{W,Z_i}).
\end{aligned}
$$

As we have $0\leq\mathrm{KL}(P^{(\alpha)}_{W,Z_i}\|\widehat{P}_{W,Z_i})$, therefore, the minimum of (26) is achieved with $\widehat{P}_{W,Z_i}=P^{(\alpha)}_{W,Z_i}$. Now, considering $\widehat{P}_{W,Z_i}=P^{(\alpha)}_{W,Z_i}$ in Proposition 1, completes the proof. ■

Proof of Proposition 2:

Using (27),

$$I^{\alpha}_{\mathrm{JS}}(W;Z_i)\leq(1-\alpha)I(W;Z_i), \qquad (79)$$

we have:

$$
\begin{aligned}
|\overline{\mathrm{gen}}(P_{W|S},\mu)| &\leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2\sigma^2_{(\alpha)}\frac{I^{\alpha}_{\mathrm{JS}}(W;Z_i)}{\alpha(1-\alpha)}} && (80)\\
&\leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2\sigma^2_{(\alpha)}\frac{I(W;Z_i)}{\alpha}} && (81)\\
&\leq \sqrt{2\sigma^2_{(\alpha)}\frac{\sum_{i=1}^{n}I(W;Z_i)}{\alpha n}} && (82)\\
&\leq \sqrt{2\sigma^2_{(\alpha)}\frac{I(W;S)}{\alpha n}} && (83)\\
&\leq \sqrt{2\sigma^2_{(\alpha)}\frac{H(W)}{\alpha n}}, && (84)
\end{aligned}
$$

where the final result would follow from the finite hypothesis space. ■

Proof of Proposition 3: This proposition follows from the fact that $I^{\alpha}_{\mathrm{JS}}(W,Z_i)\leq h(\alpha)$ for $i=1,\cdots,n$.

We prove that $I^{\alpha}_{\mathrm{JS}}(W,Z_i)\leq h(\alpha)$.

$$
\begin{aligned}
&I^{\alpha}_{\mathrm{JS}}(W,Z_i)= \qquad (85)\\
&\alpha\mathrm{KL}(P_W\otimes P_{Z_i}\|P^{(\alpha)}_{W,Z_i})+(1-\alpha)\mathrm{KL}(P_{W,Z_i}\|P^{(\alpha)}_{W,Z_i})\\
&=\alpha\int_{\mathcal{W}\times\mathcal{Z}}\mathrm{d}P_W\otimes\mathrm{d}P_{Z_i}\log\left(\frac{\mathrm{d}P_W\otimes\mathrm{d}P_{Z_i}}{\mathrm{d}P^{(\alpha)}_{W,Z_i}}\right) \qquad (86)\\
&+(1-\alpha)\int_{\mathcal{W}\times\mathcal{Z}}\mathrm{d}P_{W,Z_i}\log\left(\frac{\mathrm{d}P_{W,Z_i}}{\mathrm{d}P^{(\alpha)}_{W,Z_i}}\right)\\
&\leq\alpha\int_{\mathcal{W}\times\mathcal{Z}}\mathrm{d}P_W\otimes\mathrm{d}P_{Z_i}\log\left(\frac{\mathrm{d}P_W\otimes\mathrm{d}P_{Z_i}}{\alpha(\mathrm{d}P_W\otimes\mathrm{d}P_{Z_i})}\right) \qquad (87)\\
&+(1-\alpha)\int_{\mathcal{W}\times\mathcal{Z}}\mathrm{d}P_{W,Z_i}\log\left(\frac{\mathrm{d}P_{W,Z_i}}{(1-\alpha)\mathrm{d}P_{W,Z_i}}\right)\\
&=-\alpha\log(\alpha)-(1-\alpha)\log(1-\alpha) \qquad (88)\\
&=h(\alpha). \qquad (89)
\end{aligned}
$$

■

Proof of Corollary 1: We first compute the derivative of $\frac{h(\alpha)}{\alpha(1-\alpha)}$ with respect to $\alpha\in(0,1)$

$$\frac{d\frac{h(\alpha)}{\alpha(1-\alpha)}}{d\alpha}=\frac{\log(1-\alpha)}{\alpha^2}-\frac{\log(\alpha)}{(1-\alpha)^2}. \qquad (90)$$

Now for $\alpha=\frac{1}{2}$, we have $\frac{d\frac{h(\alpha)}{\alpha(1-\alpha)}}{d\alpha}=0$. ■

Appendix B
Proofs of Section III-B

Proof of Proposition 5: Consider arbitrary auxiliary distributions $\{\widehat{P}_{W,Z_i}\}_{i=1}^n$ defined on $\mathcal{W}\times\mathcal{Z}$. Then,

$$
\begin{aligned}
&\overline{\mathrm{gen}}(P_{W|S},\mu)=\mathbb{E}_{P_WP_S}[L_E(W,S)]-\mathbb{E}_{P_W,S}[L_E(W,S)]\\
&=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{P_WP_{Z_i}}[\ell(W,Z_i)]-\mathbb{E}_{P_WZ_i}[\ell(W,Z_i)] && (91)\\
&\leq\frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{E}_{P_WP_{Z_i}}[\ell(W,Z_i)]-\mathbb{E}_{P_WZ_i}[\ell(W,Z_i)]\right| && (92)\\
&\leq\frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{E}_{\widehat{P}_{W,Z_i}}[\ell(W,Z_i)]-\mathbb{E}_{P_WZ_i}[\ell(W,Z_i)]\right| && (93)\\
&\quad+\left|\mathbb{E}_{\widehat{P}_{W,Z_i}}[\ell(W,Z_i)]-\mathbb{E}_{P_WP_{Z_i}}[\ell(W,Z_i)]\right|.
\end{aligned}
$$

Using the assumption that loss function $\ell(w,z_i)$ is γ^2-sub-Gaussian under distribution P_{W,Z_i} and Donsker-Varadhan representation we have:

$$
\begin{aligned}
&\lambda\left(\mathbb{E}_{\widehat{P}_{W,Z_i}}[\ell(W,Z_i)]-\mathbb{E}_{P_WZ_i}[\ell(W,Z_i)]\right)\leq \qquad (94)\\
&\mathrm{KL}(\widehat{P}_{W,Z_i}\|P_WZ_i)+\frac{\lambda^2\gamma^2}{2},\quad\forall\lambda\in\mathbb{R}.
\end{aligned}
$$

Using the assumption that $\ell(w,Z)$ is σ^2-sub-Gaussian under $P_W\otimes P_{Z_i}$, and again Donsker-Varadhan representation we have:

$$
\begin{aligned}
&\lambda'\left(\mathbb{E}_{\widehat{P}_{W,Z_i}}[\ell(W,Z_i)]-\mathbb{E}_{P_WP_{Z_i}}[\ell(W,Z_i)]\right)\leq \qquad (95)\\
&\mathrm{KL}(\widehat{P}_{W,Z_i}\|P_WP_{Z_i})+\frac{\lambda'^2\sigma^2}{2}.\quad\forall\lambda'\in\mathbb{R}
\end{aligned}
$$

Note that $\mathbb{E}_{P_WP_{Z_i}}[\ell(W,Z_i)-\mathbb{E}_{P_{Z_i}}[\ell(W,Z_i)]]=0$.

Now if we consider $\lambda>0$, then we choose $\lambda'=\frac{\alpha}{\alpha-1}\lambda$. Hence we have

$$
\begin{aligned}
&\mathbb{E}_{\widehat{P}_{W,Z_i}}[\ell(W,Z_i)-\mathbb{E}_{P_{W,Z_i}}[\ell(W,Z_i)]]\leq \qquad (96)\\
&\frac{\mathrm{KL}(\widehat{P}_{W,Z_i}\|P_WZ_i)}{\lambda}+\frac{\lambda\gamma^2}{2},\quad\forall\lambda\in\mathbb{R}^+.
\end{aligned}
$$

Using the assumption that $\ell(w,Z)$ is σ^2-sub-Gaussian and again Donsker-Varadhan representation,

$$-\mathbb{E}_{\widehat{P}_{W,Z_i}}[\ell(W,Z_i)-\mathbb{E}_{P_W P_{Z_i}}[\ell(W,Z_i)]] \leq \frac{\mathrm{KL}(\widehat{P}_{W,Z_i}\|P_W P_{Z_i})}{|\lambda'|}+\frac{|\lambda'|\sigma^2}{2}, \quad \forall \lambda' \in \mathbb{R}^-. \tag{97}$$

Now sum up two Inequalities (96) and (97), to obtain

$$\mathbb{E}_{P_W P_{Z_i}}[\ell(W,Z_i)]-\mathbb{E}_{P_{W Z_i}}[\ell(W,Z_i)] \leq \frac{\alpha \mathrm{KL}(\widehat{P}_{W,Z_i}\|P_{W,Z_i})+(1-\alpha)\mathrm{KL}(\widehat{P}_{W,Z_i}\|Q_W P_{Z_i})}{\alpha\lambda}+\frac{\lambda\gamma^2}{2}+\frac{\lambda\frac{\alpha}{1-\alpha}\sigma^2}{2}, \quad \forall \lambda \in \mathbb{R}^+. \tag{98}$$

Considering (98), we have a nonnegative parabola in λ, whose discriminant must be nonpositive, and we have:

$$\left|\mathbb{E}_{P_W P_{Z_i}}[\ell(W,Z_i)]-\mathbb{E}_{P_{W Z_i}}[\ell(W,Z_i)]\right| \leq \sqrt{2(\alpha\sigma^2+(1-\alpha)\gamma^2)\frac{(\alpha C_i+(1-\alpha)D_i)}{\alpha(1-\alpha)}}, \tag{99}$$

where $C_i = \mathrm{KL}(\widehat{P}_{W,Z_i}\|P_W \otimes \mu)$ and $D_i = \mathrm{KL}(\widehat{P}_{W,Z_i}\|P_{W,Z_i})$. Finally, we prove the claim using (91). ∎

Proof of Theorem 3: Using Lemma 3, we have:

$$\min_{\widehat{P}_{W,Z_i}} \alpha\mathrm{KL}(\widehat{P}_{W,Z_i}\|P_W\otimes\mu)+(1-\alpha)\mathrm{KL}(\widehat{P}_{W,Z_i}\|P_{W,Z_i}) = (1-\alpha)I^{\alpha}_{\mathrm{R}}(W;Z_i) + \min_{\widehat{P}_{W,Z_i}} \mathrm{KL}\left(\widehat{P}_{W,Z_i}\|\frac{(P_{Z_i}\otimes P_W)^{\alpha}(P_{W,Z_i})^{(1-\alpha)}}{\int_{\mathcal{W}\times\mathcal{Z}}(\mathrm{d}P_{Z_i}\otimes\mathrm{d}P_W)^{\alpha}(\mathrm{d}P_{W,Z_i})^{(1-\alpha)}}\right).$$

Now by considering the $d\widehat{P}_{W,Z_i} = \frac{(\mathrm{d}P_{Z_i}\otimes\mathrm{d}P_W)^{\alpha}(\mathrm{d}P_{W,Z_i})^{(1-\alpha)}}{\int_{\mathcal{W}\times\mathcal{Z}}(\mathrm{d}P_{Z_i}\otimes\mathrm{d}P_W)^{\alpha}(\mathrm{d}P_{W,Z_i})^{(1-\alpha)}}$, the KL term would be equal to zero. The final result holds by using Proposition 5. ∎

Proof of Lemma 3:

Our proof is based on [35, Theorem 30]. For $0 \leq \alpha \leq 1$, we have:

$$\begin{aligned}
&\alpha\mathrm{KL}(\widehat{P}_{W,Z_i}\|P_W\otimes\mu)+(1-\alpha)\mathrm{KL}(\widehat{P}_{W,Z_i}\|P_{W,Z_i}) && (100)\\
&= \int_{\mathcal{W}\times\mathcal{Z}} d\widehat{P}_{W,Z_i}\log(d\widehat{P}_{W,Z_i}) && (101)\\
&\quad - \int_{\mathcal{W}\times\mathcal{Z}} \widehat{P}_{W,Z_i}\log((dP_W\otimes\mathrm{d}P_{Z_i})^{\alpha}(\mathrm{d}P_{W,Z_i})^{(1-\alpha)})\\
&= \int_{\mathcal{W}\times\mathcal{Z}} d\widehat{P}_{W,Z_i}\log(d\widehat{P}_{W,Z_i}) && (102)\\
&\quad - \int_{\mathcal{W}\times\mathcal{Z}} d\widehat{P}_{W,Z_i}\log\left((\mathrm{d}P_W\otimes\mathrm{d}P_{Z_i})^{\alpha}(\mathrm{d}P_{W,Z_i})^{(1-\alpha)}\right)\\
&\quad + \log\left(\int_{\mathcal{W}\times\mathcal{Z}}(\mathrm{d}P_W\otimes\mathrm{d}P_{Z_i})^{\alpha}(\mathrm{d}P_{W,Z_i})^{(1-\alpha)}\right)\\
&\quad - \log\left(\int_{\mathcal{W}\times\mathcal{Z}}(\mathrm{d}P_W\otimes\mathrm{d}P_{Z_i})^{\alpha}(\mathrm{d}P_{W,Z_i})^{(1-\alpha)}\right)\\
&= -\log\left(\int_{\mathcal{W}\times\mathcal{Z}}(P_W\otimes\mathrm{d}P_{Z_i})^{\alpha}(\mathrm{d}P_{W,Z_i})^{(1-\alpha)}\right) && (103)\\
&\quad + \int_{\mathcal{W}\times\mathcal{Z}} d\widehat{P}_{W,Z_i}\log(d\widehat{P}_{W,Z_i})\\
&\quad - \int_{\mathcal{W}\times\mathcal{Z}} d\widehat{P}_{W,Z_i}\log\left(\frac{(\mathrm{d}P_W\otimes\mathrm{d}P_{Z_i})^{\alpha}(\mathrm{d}P_{W,Z_i})^{(1-\alpha)}}{\int_{\mathcal{W}\times\mathcal{Z}}(\mathrm{d}P_W\otimes\mathrm{d}P_{Z_i})^{\alpha}(\mathrm{d}P_{W,Z_i})^{(1-\alpha)}}\right)\\
&= (1-\alpha)I^{\alpha}_{\mathrm{R}}(W;Z_i) && (104)\\
&\quad + \mathrm{KL}\left(\widehat{P}_{W,Z_i}\|\frac{(P_{Z_i}\otimes P_W)^{\alpha}(P_{W,Z_i})^{(1-\alpha)}}{\int_{\mathcal{W}\times\mathcal{Z}}(\mathrm{d}P_{Z_i}\otimes\mathrm{d}P_W)^{\alpha}(\mathrm{d}P_{W,Z_i})^{(1-\alpha)}}\right).
\end{aligned}$$

∎

Proof of Proposition 6: Using (38),

$$I^{\alpha}_{\mathrm{R}}(W;Z_i) \leq \frac{\alpha}{1-\alpha}I(W;Z_i), \tag{105}$$

and considering the hypothesis space is finite and the upper bound in Theorem 3, we have:

$$\begin{aligned}
|\overline{\mathrm{gen}}(P_{W|S},\mu)| &\leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2(\alpha\sigma^2+(1-\alpha)\gamma^2)\frac{I^{\alpha}_{\mathrm{R}}(W;Z_i)}{\alpha}} && (106)\\
&\leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2(\alpha\sigma^2+(1-\alpha)\gamma^2)\min\left\{\frac{1}{\alpha},\frac{1}{1-\alpha}\right\}I(W;Z_i)} && (107)\\
&\leq \sqrt{2(\alpha\sigma^2+(1-\alpha)\gamma^2)\min\left\{\frac{1}{\alpha},\frac{1}{1-\alpha}\right\}\frac{\sum_{i=1}^{n}I(W;Z_i)}{n}} && (108)\\
&\leq \sqrt{2(\alpha\sigma^2+(1-\alpha)\gamma^2)\min\left\{\frac{1}{\alpha},\frac{1}{1-\alpha}\right\}\frac{I(W;S)}{n}} && (109)\\
&\leq \sqrt{2(\alpha\sigma^2+(1-\alpha)\gamma^2)\min\left\{\frac{1}{\alpha},\frac{1}{1-\alpha}\right\}\frac{H(W)}{n}} && (110)\\
&\leq \sqrt{2(\alpha\sigma^2+(1-\alpha)\gamma^2)\min\left\{\frac{1}{\alpha},\frac{1}{1-\alpha}\right\}\frac{\log(k)}{n}},
\end{aligned}$$

where (108) follows from Jensen inequality and (109) follows from i.i.d assumption for Z_i's. ∎

Proof of Theorem 4: Consider arbitrary auxiliary distributions $\{\tilde{P}_{W,Z_i}\}_{i=1}^n$ defined on $\mathcal{W}\times\mathcal{Z}$.

$$\begin{aligned}
\overline{\mathrm{gen}}(P_{W|S},\mu) &= \mathbb{E}_{P_W P_S}[L_E(W,S)]-\mathbb{E}_{P_{W,S}}[L_E(W,S)]\\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{P_W P_{Z_i}}[\ell(W,Z_i)]-\mathbb{E}_{P_{W Z_i}}[\ell(W,Z_i)] && (111)\\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{E}_{P_W P_{Z_i}}[\ell(W,Z_i)]-\mathbb{E}_{P_{W Z_i}}[\ell(W,Z_i)]\right|. && (112)
\end{aligned}$$

Using the assumption centered loss function $\ell(w,z_i)-\mathbb{E}_{P_{Z_i}}[\ell(w,Z_i)]$ is γ^2-sub-Gaussian under distribution P_{W,Z_i} and Donsker-Varadhan representation by considering function $\ell(w,z_i)-\mathbb{E}_{P_{Z_i}}[\ell(w,Z_i)]$ we have:

$$\begin{aligned}
&\lambda\Big(\mathbb{E}_{\tilde{P}_{W,Z_i}}[\ell(W,Z_i)-\mathbb{E}_{P_{Z_i}}[\ell(W,Z_i)]] && (113)\\
&\qquad -\mathbb{E}_{P_{W Z_i}}[\ell(W,Z_i)-\mathbb{E}_{P_{Z_i}}[\ell(W,Z_i)]]\Big)\\
&\leq \mathrm{KL}(\tilde{P}_{W,Z_i}\|P_{W Z_i})+\frac{\lambda^2\gamma^2}{2}. \quad \forall\lambda\in\mathbb{R}
\end{aligned}$$

Note that $\mathbb{E}_{P_{W Z_i}}[\ell(W,Z_i)-\mathbb{E}_{P_{Z_i}}[\ell(W,Z_i)]] = \mathbb{E}_{P_{W Z_i}}[\ell(W,Z_i)]-\mathbb{E}_{P_W P_{Z_i}}[\ell(W,Z_i)]$.

Using the assumption that $\ell(w,Z)$ is σ^2-sub-Gaussian under P_{Z_i} for all $w\in\mathcal{W}$, and again Donsker-Varadhan representation by considering function $\ell(w,z_i)-\mathbb{E}_{P_{Z_i}}[\ell(w,Z_i)]$ we have:

$$\lambda'\Big(\mathbb{E}_{\tilde{P}_{W,Z_i}}[\ell(W,Z_i)-\mathbb{E}_{P_{Z_i}}[\ell(W,Z_i)]] - \mathbb{E}_{Q_W P_{Z_i}}[\ell(W,Z_i)-\mathbb{E}_{P_{Z_i}}[\ell(W,Z_i)]]\Big) \leq \mathrm{KL}(\tilde{P}_{W,Z_i}\|Q_W P_{Z_i}) + \frac{\lambda'^2\sigma^2}{2}. \quad \forall \lambda'\in\mathbb{R} \tag{114}$$

Note that $\mathbb{E}_{Q_W P_{Z_i}}[\ell(W,Z_i)-\mathbb{E}_{P_{Z_i}}[\ell(W,Z_i)]]=0$.

Now if we consider $\lambda>0$, then we choose $\lambda'=\frac{\alpha}{\alpha-1}\lambda$. Hence we have

$$\mathbb{E}_{\tilde{P}_{W,Z_i}}[\ell(W,Z_i)-\mathbb{E}_{P_{Z_i}}[\ell(W,Z_i)]] - \mathbb{E}_{P_{WZ_i}}[\ell(W,Z_i)-\mathbb{E}_{P_{Z_i}}[\ell(W,Z_i)]] \leq \frac{\mathrm{KL}(\tilde{P}_{W,Z_i}\|P_{WZ_i})}{\lambda} + \frac{\lambda\gamma^2}{2}, \quad \forall\lambda\in\mathbb{R}^+. \tag{115}$$

Using the assumption $\ell(w,Z)$ is σ^2-sub-Gaussian and again Donsker-Varadhan representation,

$$-\mathbb{E}_{\tilde{P}_{W,Z_i}}[\ell(W,Z_i)-\mathbb{E}_{P_{Z_i}}[\ell(W,Z_i)]] \leq \frac{\mathrm{KL}(\tilde{P}_{W,Z_i}\|Q_W P_{Z_i})}{|\lambda'|} + \frac{|\lambda'|\sigma^2}{2}, \quad \forall\lambda'\in\mathbb{R}^-. \tag{116}$$

Now sum up the two Inequalities (115) and (116) to obtain,

$$\mathbb{E}_{P_W P_{Z_i}}[\ell(W,Z_i)]-\mathbb{E}_{P_{WZ_i}}[\ell(W,Z_i)] \leq \frac{\alpha\mathrm{KL}(\tilde{P}_{W,Z_i}\|P_{W,Z_i})+(1-\alpha)\mathrm{KL}(\tilde{P}_{W,Z_i}\|Q_W P_{Z_i})}{\alpha\lambda} + \frac{\lambda\gamma^2}{2}+\frac{\lambda\frac{\alpha}{1-\alpha}\sigma^2}{2}, \quad \forall\lambda\in\mathbb{R}^+. \tag{117}$$

Taking infimum on $\tilde{P}_{W,Z_i}$ and using [35, Theorem 30] that states

$$(1-\alpha)\mathrm{R}_\alpha(P_1\|P_2)=\inf_R\{\alpha\mathrm{KL}(R\|P_1)+(1-\alpha)\mathrm{KL}(R\|P_2)\}$$

Now, we have:

$$(1-\alpha)\mathrm{R}_\alpha(P_{W,Z_i}\|Q_W P_{Z_i}) = \inf_{\tilde{P}_{W,Z_i}}\{\alpha\mathrm{KL}(\tilde{P}_{W,Z_i}\|P_{W,Z_i})+(1-\alpha)\mathrm{KL}(\tilde{P}_{W,Z_i}\|Q_W P_{Z_i})\} \tag{118}$$

and taking infimum on Q_W, we have:

$$\inf_{Q_W}\mathrm{R}_\alpha(P_{W,Z_i}\|Q_W P_{Z_i})=I_{\mathrm{S}}^\alpha(Z_i;W). \tag{119}$$

Using (119) in (117), we get:

$$\mathbb{E}_{P_W P_{Z_i}}[\ell(W,Z_i)]-\mathbb{E}_{P_{WZ_i}}[\ell(W,Z_i)] \leq \frac{(1-\alpha)I_\alpha^s(Z_i;W)}{\lambda\alpha}+\frac{\lambda\gamma^2}{2}+\frac{\lambda\frac{\alpha}{1-\alpha}\sigma^2}{2} \quad \forall\lambda\in\mathbb{R}^+. \tag{120}$$

Using the same approach for $\lambda\in R^-$, we have:

$$\mathbb{E}_{P_{WZ_i}}[\ell(W,Z_i)]-\mathbb{E}_{P_W P_{Z_i}}[\ell(W,Z_i)] \leq \frac{(1-\alpha)I_\alpha^s(Z_i;W)}{|\lambda|\alpha}+\frac{|\lambda|\gamma^2}{2}+\frac{|\lambda|\frac{\alpha}{1-\alpha}\sigma^2}{2}, \quad \forall\lambda\in\mathbb{R}^-. \tag{121}$$

Considering (120) and (121), we have a non-negative parabola in λ, whose discriminant must be non-positive, and we have:

$$\big|\mathbb{E}_{P_W P_{Z_i}}[\ell(W,Z_i)]-\mathbb{E}_{P_{WZ_i}}[\ell(W,Z_i)]\big| \leq \sqrt{2(\alpha\sigma^2+(1-\alpha)\gamma^2)\frac{I_{\mathrm{S}}^\alpha(Z_i;W)}{\alpha}}. \tag{122}$$

We prove the claim using (91). ∎

Proof of Proposition 7: The Generalized Pinsker's inequality is introduced in [35], as follows,

$$\mathbb{TV}(P,Q)^2\leq\frac{2}{\alpha}\mathrm{R}_\alpha(P\|Q), \quad \alpha\in(0,1], \tag{123}$$

where $\mathbb{TV}(P,Q)=\int_{\mathcal{X}}|P(\mathrm{d}x)-Q(\mathrm{d}x)|$. Denote $f:\mathcal{X}\to\mathbb{R}$ a bounded function $|f|\leq L$, then

$$\mathbb{E}_P[f(X)]-\mathbb{E}_Q[f(X)] = \int f(x)(P(dx)-Q(dx)) \leq \sup_x f(x)\cdot\int|P(dx)-Q(dx)| \leq L\sqrt{\frac{2}{\alpha}\mathrm{R}_\alpha(P\|Q)}. \tag{124}$$

Let $P=P_{W,Z}$, $Q=P_W P_Z$ and $f(w,z)=L_\mu(w)-L_E(w,z)$. Then, we have the final result,

$$\overline{\mathrm{gen}}(P_{W|S},\mu)=\frac{1}{n}\sum_{i=1}^n\mathbb{E}[L_\mu(W)-L_E(W,Z_i)] \leq \frac{1}{n}\sum_{i=1}^n\sqrt{\frac{2b^2}{\alpha}\mathrm{R}_\alpha(P_{W,Z_i}\|P_W P_{Z_i})}.$$

∎

Appendix C
Proof of Section III-C

Proof of Proposition 8: It follows from

$$I_{\mathrm{JS}}^\alpha(W;Z_i)\leq\frac{h(\alpha')}{\alpha'(1-\alpha')},$$

that if we have $\frac{\alpha h(\alpha')}{\alpha'(1-\alpha')}\leq I_{\mathrm{R}}^\alpha(W;Z_i)$ for all $i=1,\cdots,n$, then the results holds for $\sigma=\gamma=\sigma_{JS}$. ∎

Appendix D
Proofs of Section IV

Proof of Theorem 5: Let us define $P_{\mathcal{N}}:=\mathcal{N}(w^\star,\beta^{-1}I_d)$ and $w^\star:=\arg\inf_{w\in\mathcal{W}}L_\mu(w)$.

$$\begin{aligned}
&\mathcal{E}_r(P_{W|S}^{\star,\beta,\mathrm{JS}_\alpha},\mu)\\
&\leq\big|\overline{\mathrm{gen}}(P_{W|S}^{\star,\beta,\mathrm{JS}_\alpha},\mu)\big|+\mathbb{E}_{P_S\otimes P_{W|S}^{\star,\beta,\mathrm{JS}_\alpha}}[L_E(W,S)]-L_\mu(w^\star)\\
&\leq\sqrt{\frac{2b^2}{n\alpha(1-\alpha)}\sum_{i=1}^n I_{\mathrm{JS}}^\alpha(W;Z_i)}+\mathbb{E}_{P_S\otimes P_{\mathcal{N}}}[L_E(W,S)]-L_\mu(w^\star)\\
&\quad+\frac{\mathrm{JS}_\alpha(\mathcal{N}(w^\star,\beta^{-1}I_d)\|Q)}{\beta}\\
&\leq\sqrt{\frac{2b^2}{n\alpha(1-\alpha)}\sum_{i=1}^n I_{\mathrm{JS}}^\alpha(W;Z_i)}+\mathbb{E}_{P_{\mathcal{N}}}[L_\mu(W)]-L_\mu(w^\star)
\end{aligned}$$

$$+\frac{\mathrm{JS}_\alpha(\mathcal{N}(w^\star,\beta^{-1}I_d)\|Q)}{\beta}$$
$$\le\sqrt{\frac{2b^2}{n\alpha(1-\alpha)}\sum_{i=1}^n I^\alpha_{\mathrm{JS}}(W;Z_i)}+\mathbb{E}_{P_{\mathcal{N}}}[L_\mu(w^\star)$$
$$+\tilde{L}\|W-w^\star\|_2]-L_\mu(w^\star)+\frac{\mathrm{JS}_\alpha(\mathcal{N}(w^\star,\beta^{-1}I_d)\|Q)}{\beta}$$
$$\le\sqrt{\frac{2b^2}{n\alpha(1-\alpha)}\sum_{i=1}^n I^\alpha_{\mathrm{JS}}(W;Z_i)}+\frac{\tilde{L}\sqrt{d}}{\beta}$$
$$+\frac{\mathrm{JS}_\alpha(\mathcal{N}(w^\star,\beta^{-1}I_d)\|Q)}{\beta},$$

Note that $\tilde{L}\mathbb{E}_{P_{\mathcal{N}}}[\|W-w^\star\|_2]=\frac{\tilde{L}d}{\beta}$. ■

Proof of Theorem 6: The proof is similar to Proof of Theorem 5, by replacing the second inequality with the following inequality,

$$\mathcal{E}_r(P^{\star,\beta,\mathrm{R}_\alpha}_{W|S},\mu)\le\left|\overline{\mathrm{gen}}(P^{\star,\beta,\mathrm{R}_\alpha}_{W|S},\mu)\right|$$
$$+\mathbb{E}_{P_S\otimes P^{\star,\beta,\mathrm{JS}_\alpha}_{W|S}}[L_E(W,S)]-L_\mu(w^\star)$$
$$\le\sqrt{\frac{2b^2}{n\alpha(1-\alpha)}\sum_{i=1}^n I^\alpha_{\mathrm{R}}(W;Z_i)}$$
$$+\mathbb{E}_{P_S\otimes P_{\mathcal{N}}}[L_E(W,S)]-L_\mu(w^\star)+\frac{\mathrm{R}_\alpha(\mathcal{N}(w^\star,\beta^{-1}I_d)\|Q)}{\beta}$$

■

Proof of Corollary 2: Using the boundedness of α-JS divergence in Theorem 5, we have,

$$\mathcal{E}_r(P^{\star,\beta,\mathrm{JS}_\alpha}_{W|S},\mu)\le\sqrt{\frac{2b^2}{n\alpha}\sum_{i=1}^n I^\alpha_{\mathrm{R}}(W;Z_i)}+\frac{\tilde{L}\sqrt{d}}{\beta}+\frac{h(\alpha)}{\beta},$$

where $h(\alpha)=-\alpha\log(\alpha)-(1-\alpha)\log(1-\alpha)$. Therefore, by setting $\beta=n^{1/2}$, the convergence rate of excess risk is $\mathcal{O}(1/\sqrt{n})$. ■

Proof of Corollary 3:

We consider the normal distribution as prior, i.e., $Q=\mathcal{N}(0,I_d)$, in Theorem 6. Then, we can compute the α-Rényi divergence between two multivariate Gaussian distributions [49],

$$\mathrm{R}_\alpha(\mathcal{N}(w^\star,\beta^{-1}I_d)\|\mathcal{N}(0,I_d))=\frac{\alpha}{2}\|w^\star\|_2^2(\alpha+(1-\alpha)\beta^{-1})^{-1}$$
$$+\frac{d}{2(\alpha-1)}\log\Big(\frac{\beta^{\alpha-1}}{\alpha+(1-\alpha)\beta^{-1}}\Big).$$

Then, the following upper bound holds on the excess risk under $P^{\star,\beta,\mathrm{R}_\alpha}_{W|S}$,

$$\mathcal{E}_r(P^{\star,\beta,\mathrm{R}_\alpha}_{W|S},\mu)\le\sqrt{\frac{2b^2}{n\alpha}\sum_{i=1}^n I^\alpha_{\mathrm{R}}(W;Z_i)}+\frac{\tilde{L}\sqrt{d}}{\beta}$$
$$+\frac{\alpha}{2\beta}\|w^\star\|_2^2(\alpha+(1-\alpha)\beta^{-1})^{-1}$$
$$+\frac{d}{2\beta(\alpha-1)}\log\Big(\frac{\beta^{(\alpha-1)}}{\alpha+(1-\alpha)\beta^{-1}}\Big)$$
$$\le\sqrt{\frac{2b^2}{n\alpha}\sum_{i=1}^n I^\alpha_{\mathrm{R}}(W;Z_i)}+\frac{\tilde{L}\sqrt{d}}{\beta}$$
$$+\frac{1}{2\beta}\|w^\star\|_2^2+\frac{d}{2\beta}\log(\beta)+\frac{d}{2\beta(1-\alpha)}\log(\alpha).$$

■

Appendix E
Proofs of Section V

We first propose the following Lemma to provide an upper bound on the expected generalization error under distribution mismatch.

Lemma 6: The following upper bound holds on expected generalization error under distribution mismatch between the test and training distributions:

$$|\overline{\mathrm{gen}}(P_{W|S},\mu,\mu')|\le \tag{125}$$
$$|\overline{\mathrm{gen}}(P_{W|S},\mu)|+|\mathbb{E}_{P_W\otimes\mu'}[\ell(W,Z)]-\mathbb{E}_{P_W\otimes\mu}[\ell(W,Z)]|.$$

Proof: We have:

$$|\overline{\mathrm{gen}}(P_{W|S},\mu,\mu')| \tag{126}$$
$$=|\mathbb{E}_{P_{W,S}}[L_P(W,\mu')-L_P(W,\mu)+L_P(W,\mu)-L_E(E,S)]|$$
$$\le|\mathbb{E}_{P_{W,S}}[L_P(W,\mu')-L_P(W,\mu)]|+|\overline{\mathrm{gen}}(P_{W|S},\mu)|$$
$$=|\mathbb{E}_{P_W\otimes\mu'}[\ell(W,Z)]-\mathbb{E}_{P_W\otimes\mu}[\ell(W,Z)]|+|\overline{\mathrm{gen}}(P_{W|S},\mu)|$$

■

Proof of Proposition 9: In Lemma 6, the generalization error under distribution mismatch can be upper bounded by two terms. Considering Theorem 2, we can provide the upper bound based on α-Jensen-Shannon information over $|\overline{\mathrm{gen}}(P_{W|S},\mu)|$. We can also provide an upper bound on the term $|\mathbb{E}_{P_W\otimes\mu'}[\ell(W,Z)]-\mathbb{E}_{P_W\otimes\mu}[\ell(W,Z)]|$ in Lemma 6 by applying ADM using a similar approach as in Theorem 2 and using the α-Jensen-Shannon divergence as follows:

$$|\mathbb{E}_{P_W\otimes\mu'}[\ell(W,Z)]-\mathbb{E}_{P_W\otimes\mu}[\ell(W,Z)]| \tag{127}$$
$$\le\sqrt{2\sigma^2_{(\alpha)}\frac{\mathrm{JS}_\alpha(P_W\otimes\mu'\|P_W\otimes\mu)}{\alpha(1-\alpha)}} \tag{128}$$
$$=\sqrt{2\sigma^2_{(\alpha)}\frac{\mathrm{JS}_\alpha(\mu'\|\mu)}{\alpha(1-\alpha)}}. \tag{129}$$

■

Proof of Proposition 10: Based on Lemma 6, the generalization error is upper bounded by two terms (See Equation (125)). We can provide the upper bound based on α-Rényi information over $|\overline{\mathrm{gen}}(P_{W|S},\mu)|$ using Theorem 3. We can also provide an upper bound on the term

$|\mathbb{E}_{P_W\otimes\mu'}[\ell(W,Z)] - \mathbb{E}_{P_W\otimes\mu}[\ell(W,Z)]|$ by applying ADM using a similar approach as in Theorem 3 and using α-Rényi divergence as follows:

$$|\mathbb{E}_{P_W\otimes\mu'}[\ell(W,Z)] - \mathbb{E}_{P_W\otimes\mu}[\ell(W,Z)]| \tag{130}$$

$$\leq \sqrt{2(\alpha\sigma^2 + (1-\alpha)\gamma^2)\frac{\mathrm{R}_\alpha(P_W\otimes\mu'\|P_W\otimes\mu)}{\alpha}} \tag{131}$$

$$= \sqrt{2(\alpha\sigma^2 + (1-\alpha)\gamma^2)\frac{\mathrm{R}_\alpha(\mu'\|\mu)}{\alpha}}. \tag{132}$$

■

REFERENCES

[1] G. Aminian, L. Toni, and M. R. D. Rodrigues, "Jensen-shannon information based characterization of the generalization error of learning algorithms," in *2020 IEEE Information Theory Workshop (ITW)*, pp. 1–5, 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*, pp. 2524–2533, 2017.

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

[10] 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.

[11] A. R. Esposito, M. Gastpar, and I. Issa, "Generalization error bounds via rényi-, f-divergences and maximal leakage," *IEEE Transactions on Information Theory*, vol. 67, no. 8, pp. 4986–5004, 2021.

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

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

[14] 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)*, pp. 577–581, IEEE, 2019.

[15] 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.

[16] G. Aminian, Y. Bu, G. Wornell, and M. Rodrigues, "Tighter expected generalization error bounds via convexity of information measures," in *IEEE International Symposium on Information Theory (ISIT)*, 2022.

[17] T. Steinke and L. Zakynthinou, "Reasoning about generalization via conditional mutual information," in *Conference on Learning Theory*, pp. 3437–3452, PMLR, 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] 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.

[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. Mansour, M. Mohri, and A. Rostamizadeh, "Domain adaptation: Learning bounds and algorithms," *arXiv preprint arXiv:0902.3430*, 2009.

[23] Z. Wang, "Theoretical guarantees of transfer learning," 2018.

[24] X. Wu, J. H. Manton, U. Aickelin, and J. Zhu, "Information-theoretic analysis for transfer learning," in *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 2819–2824, IEEE, 2020.

[25] E. Englesson and H. Azizpour, "Generalized jensen-shannon divergence loss for learning with noisy labels," *Advances in Neural Information Processing Systems*, vol. 34, pp. 30284–30297, 2021.

[26] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[27] P. Melville, S. M. Yang, M. Saar-Tsechansky, and R. Mooney, "Active learning for probability estimation using jensen-shannon divergence," in *European conference on machine learning*, pp. 268–279, Springer, 2005.

[28] E. Choi and C. Lee, "Feature extraction based on the bhattacharyya distance," *Pattern Recognition*, vol. 36, no. 8, pp. 1703–1709, 2003.

[29] G. B. Coleman and H. C. Andrews, "Image segmentation by clustering," *Proceedings of the IEEE*, vol. 67, no. 5, pp. 773–785, 1979.

[30] F. Topsøe, "Information theory at the service of science," in *Entropy, Search, Complexity*, pp. 179–207, Springer, 2007.

[31] F. Topsoe, "Some inequalities for information divergence and related measures of discrimination," *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1602–1609, 2000.

[32] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[33] F. Nielsen, "On a generalization of the jensen-shannon divergence and the jensen-shannon centroid," *Entropy*, vol. 22, no. 2, p. 221, 2020.

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

[35] T. Van Erven and P. Harremos, "Rényi divergence and kullback-leibler divergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.

[36] T. Kailath, "The divergence and bhattacharyya distance measures in signal selection," *IEEE transactions on communication technology*, vol. 15, no. 1, pp. 52–60, 1967.

[37] D. P. Palomar and S. Verdú, "Lautum information," *IEEE Transactions on Information Theory*, vol. 54, no. 3, pp. 964–975, 2008.

[38] I. Sason and S. Verdú, "f-divergence inequalities," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 5973–6006, 2016.

[39] S. Verdú, "α-mutual information," in *2015 Information Theory and Applications Workshop (ITA)*, pp. 1–6, IEEE, 2015.

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

[41] M. Gastpar, A. R. Esposito, and I. Issa, "Information measures, learning and generalization," *5th London Symposium on Information Theory*, 2019.

[42] F. Topsoe, "Inequalities for the jensen-shannon divergence," *Draft available at http://www. math. ku. dk/topsoe*, 2002.

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

[44] V. Anantharam, "A variational characterization of rényi divergences," *IEEE Transactions on Information Theory*, vol. 64, no. 11, pp. 6979–6989, 2018.

[45] I. Sason, "On f-divergences: Integral representations, local behavior, and inequalities," *Entropy*, vol. 20, no. 5, p. 383, 2018.

[46] I. Kuzborskij, N. Cesa-Bianchi, and C. Szepesvári, "Distribution-dependent analysis of gibbs-erm principle," in *Conference on Learning Theory*, pp. 2028–2054, PMLR, 2019.

[47] P. Dupuis and R. S. Ellis, *A weak convergence approach to the theory of large deviations*, vol. 902. John Wiley & Sons, 2011.

[48] F. Topsøe, "Jenson-shannon divergence and norm-based measures of discrimination and variation," *preprint*, 2003.

[49] M. Gil, F. Alajaji, and T. Linder, "Rényi divergence measures for commonly used univariate continuous distributions," *Information Sciences*, vol. 249, pp. 124–131, 2013.

Gholamali Aminian (Member, IEEE) received a 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 awarded the Newton International Fellowship by the Royal Society and he was an Honorary Research Fellow at 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.

Saeed Masiha 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 (EPFL). His research interests include the theory of machine learning, nonconvex optimization, reinforcement learning, and information theory.

Laura Toni (Senior Member, IEEE) is an associate professor in the Department of Electronic and Electrical Engineering at University College London (UCL). She received her PhD degree in electrical engineering in 2009 from the University of Bologna, Italy. After her PhD, she was a Post-Doc at the University of California at San Diego (UCSD) from 2011-2012 and at the Swiss Federal Institute of Technology (EPFL), Switzerland from 2012-2016. 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 (co)-authored 30 high-impact journals and over 60 conference publications, and she is co-inventor of 2 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, IEEE best 10% paper award at VCIP 2016, IEEE Best paper award at IEEE ISM 2016, ACM best 10% paper award at MMSP 2013. She is significantly involved in scientific committees (SIGMM), she is also an ELLIS member and an the Alan Turing Fellow. She has served as the Technical Program Chair of ACM Multimedia 2022, associate editor of IEEE Transaction on Image processing, and EURASIP Journal on Advances in Signal Processing Transactions on Multimedia Computing, Communications, and Applications.

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, UCL, and a Turing Fellow with the Alan Turing Institute - the UK National Institute of Data Science and Artificial Intelligence. His research lies in the general areas of information theory, information processing, and machine learning. 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), and the IEEE Communications and Information Theory Societies Joint Paper Award 2011. He is an Associate Editor for the IEEE TRANSACTIONS ON INFORMATION THEORY, and the IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY. He was an Associate Editor for 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 a 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 (SiGDML).