Bounding Excess Minimum Risk via Rényi's Divergence

Ananya Omanwar, Fady Alajaji, Tamás Linder

Department of Mathematics and Statistics
Queen's University
Kingston, ON, Canada
{a.omanwar,fa,tamas.linder}@queensu.ca

***Abstract*—Given finite dimensional random vectors Y, X and Z that form a Markov chain in that order ($Y \to X \to Z$), we derive Rényi divergence based upper bounds for excess minimum risk, where Y is a (target) vector that is to be estimated from an observed (feature) vector X or its (stochastically) degraded version Z. We define the excess minimum risk as the difference between the minimum expected loss in estimating Y from X and the minimum expected loss in estimating Y from Z. We obtain a family of bounds which generalize the bounds developed by Györfi *et al.* (2023) expressed in terms of Shannon's mutual information. Our bounds are similar to the bounds by Modak *et al.* (2021) obtained in the context of the generalization error of learning algorithms, but unlike the latter they do not involve fixed sub-Gaussian parameters and therefore hold for more general joint distributions of Y, X, and Z. We also provide an example with Bernoulli random variables where Rényi's divergence based upper bound are tighter than mutual information bounds.**

I. Introduction

The excess minimum risk in statistical inference quantifies the difference between the minimum expected loss attained by estimating a (target) hidden random vector from a feature (observed) random vector and the minimum expected loss incurred by estimating the hidden vector from a stochastically degraded version of the feature vector. The aim of this work is to derive upper bounds on the excess minimum risk in terms of the Rényi divergence measure [1].

Recently, several bounds of this nature expressed in terms of information theoretic measures have appeared in the literature, including among others [2]–[13]. Most of these works have focused on the (expected) generalization error of learning algorithms. In [2], Xu and Raginsky establish bounds on the generalization error in terms of Shannon's mutual information between the (input) training data set and the (output) hypothesis; these bounds are tightened in [3] using the mutual information between individual data samples (instead of the entire data set) and the hypothesis. In [7], Modak *et al.* extend the later works, obtaining upper bounds on the generalization error in terms of the Rényi divergence by employing the variational characterization of the Rényi divergence [14]–[16]. The authors also derive bounds on the probability of generalization error via Rényi's divergence, which recover the bounds of Esposito *et al.* in [5] (see also [4], [6] for bounds expressed in terms of the f-divergence [17]). More recently, Aminian *et al*. [13] obtained a family of bounds on the generalization error that are applicable to supervised learning settings using a so-called "auxiliary distribution method." In particular, they obtain new bounds based on the α-Jensen-Shannon and the α-Rényi mutual informations. Here both mutual information measures are defined via a divergence between a joint distribution and a product of its marginals: the former using the Jensen-Shannon divergence of weight α [18, Eq. (4.1)] (which is always finite), while the later using the Rényi divergence of order α. Other work on the analysis of the generalization error include [8], [19] for deep learning generative adversarial networks [20] and [12] for the Gibbs algorithm (see also the extensive lists of references therein).

In this work, we focus on the excess minimum risk in statistical inference. Our motivation is to generalize the work of Györfi *et al*. [10], where mutual information based upper bounds on the minimum excess risk were derived that hold for a larger class of loss functions satisfying some (standard) sub-Gaussianity conditions. Prior related (but different) work on information-theoretic bounds on excess risk include [9] and [11], where the bounds are developed in a Bayesian learning framework (involving training data). Using Rényi's divergence, we herein extend the bound in [10] by providing a family of bounds (parameterized by the Rényi order). We adopt a similar approach to [7] by leveraging the variational representation of the Rényi divergence. However unlike the bounds in [7] (and other generalization error bounds in the literature including [13]) where the sub-Gaussian parameter is a constant, our bounds are expressed in terms of a sub-Gaussian parameter that is dependent on the (target) estimated vector; as a result, our bounds are valid for more general joint distributions of the involved random vectors.

This paper is organized as follows. In Section II, we provide preliminary definitions and introduce the statistical inference problem. In Section III, we prove a family of upper bounds on the excess minimum risk in terms of the Rényi divergence of order $\alpha \in (0, 1)$. We also present an example involving two concatenated binary symmetric channels (BSCs) which numerically shows that the derived Rényi divergence based bounds perform better than the mutual information bound for a certain range of α values. In Section IV, we provide concluding remarks and point out directions for future work.

II. Problem Setup

Consider a random vector $Y \in \mathbb{R}^p$, $p \geq 1$, that is to be estimated (predicted) from a random observation vector X taking values in $\mathbb{R}^q$, $q \geq 1$. Given a measurable estimator (predictor) $f : \mathbb{R}^q \to \mathbb{R}^p$ and a loss function $l : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^+$, the loss (risk) realized in estimating Y by $f(X)$ is given by $l(Y, f(X))$. The minimum expected risk in predicting Y from X is defined by

$$L_l^*(Y|X) = \inf_{f:\mathbb{R}^q \to \mathbb{R}^p} \mathbb{E}[l(Y, f(X))] \tag{1}$$

where the infimum is over all measurable f.

We also consider another random observation vector Z that is a random transformation or stochastically degraded version of X, obtained for example by observing X through a noisy channel. Here Z takes values in $\mathbb{R}^r$, $r \geq 1$, and Y, X and Z form a Markov chain in this order, which we denote as $Y \to X \to Z$. We similarly define the minimum expected risk in predicting Y from Z as

$$L_l^*(Y|Z) = \inf_{g:\mathbb{R}^r \to \mathbb{R}^p} \mathbb{E}[l(Y, g(Z))], \tag{2}$$

where the infimum is over all measurable predictors g. With above two predictions, we define the excess minimum risk as the difference $L_l^*(Y|Z) - L_l^*(Y|X)$, which is always non-negative due to the Markov chain condition $Y \to X \to Z$ (e.g., see the data processing inequality for expected risk in [9, Lemma 1]). Our objective is to establish an upper bound to this difference using the Rényi's divergence of order α, hence an upper parameterized by α.

In [10], the random vector Z is taken as $T(X)$, a transformation of random vector X, where $T : \mathbb{R}^p \to \mathbb{R}^r$ is measurable. The authors derive bounds on the excess minimum risk using Shannon's mutual information. We herein generalize these bounds via Rényi's divergence of order $\alpha \in (0, 1)$, which recovers the mutual information as $\alpha \to 1$. Furthermore, we use an arbitrary random vector Z, as the degraded version of the observation X instead of $T(X)$. We also provide an example where the Rényi divergence based bounds perform better than the mutual information based bounds of [10] for a certain range α.

We close this section with some definitions that we will invoke when deriving our results.

Definition 1: The Rényi's divergence of order $\alpha \in (0, \infty)$ (with $\alpha \neq 1$) between two probability measures P and Q on a measurable space $(\Omega, \mathcal{M})$, is denoted $D_\alpha(P\|Q)$ and defined as follows [1], [21]. Given a sigma-finite positive measure ν, let P and Q be absolutely continuous with respect to ν, written as $P, Q \ll \nu$, with Radon–Nikodym derivatives $\frac{dP}{d\nu} = p$ and $\frac{dQ}{d\nu} = q$, respectively. Then

$$D_\alpha(P\|Q) = \begin{cases} \frac{1}{\alpha-1} \log \left[\int p^\alpha q^{1-\alpha} d\nu\right] & \text{if } 0 < \alpha < 1 \text{ or } \alpha > 1 \\ & \text{and } P \ll Q \\ +\infty & \text{if } \alpha > 1 \text{ and } P \not\ll Q. \end{cases}$$

For finite sample space Ω of size n, the Rényi divergence of order $\alpha > 0$, $\alpha \neq 1$, between distributions $P = (p_1, p_2, ..., p_n)$ and $Q = (q_1, q_2, ..., q_n)$ is given by

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \log \sum_{i=1}^{n} p_i^\alpha q_i^{1-\alpha}, \tag{3}$$

where, for $\alpha > 1$ and $p_i = 0$, we use the convention that $p_i^\alpha q_i^{1-\alpha}$ equals 0 (resp., ∞) when $p_i = 0$ (resp., $P_i > 0$).

Definition 2: The conditional Rényi divergence of order α between the conditional distributions $P_{V|U}$ and $Q_{V|U}$ given P_U is denoted by $D_\alpha(P_{V|U}\|Q_{V|U}|P_U)$ and given by

$$D_\alpha(P_{V|U}\|Q_{V|U}|P_U) = \mathbb{E}_{P_U}\left[D_\alpha(P_{V|U}(\cdot\,|U)\|Q_{V|U}(\cdot\,|U))\right], \tag{4}$$

where $\mathbb{E}_{P_U}[\,\cdot\,]$ denotes expectation with respect to the distribution of U.

Note that the above definition of conditional Rényi divergence *differs* from the standard one, which is given as $D_\alpha(P_{V|U}P_U\|Q_{V|U}P_U)$, e.g., see [22, Definition 3]. However as $\alpha \to 1$, both notions of conditional Rényi divergence recover the conditional Kullback-Liebler (KL) divergence, which is

$$D_{\text{KL}}(P_{V|U}\|Q_{V|U}|P_U) = D_{\text{KL}}(P_{V|U}P_U\|Q_{V|U}P_U).$$

Definition 3: A real random variable U with finite expectation is said to be σ^2-sub-Gaussian for some $\sigma^2 > 0$ if

$$\log \mathbb{E}[e^{\lambda(U - \mathbb{E}(U))}] \leq \frac{\sigma^2 \lambda^2}{2} \tag{5}$$

for all $\lambda \in \mathbb{R}$.

III. Bounding Excess Minimum Risk

In this section, we derive Rényi's divergence based bounds on excess minimum risk. We first state the variational characterization of the Rényi divergence [16], which generalizes the Donsker-Varadhan variational formula for KL divergence [23].

Lemma 1: [16, Theorem 3.1] Let P and Q be two probability measures on $(\Omega, \mathcal{M})$ and $\alpha \in (0, \infty)$, $\alpha \neq 1$. Let g be a measurable function such that $e^{(\alpha-1)g} \in \mathcal{L}^1(P)$ and $e^{\alpha g} \in \mathcal{L}^1(Q)$, where $\mathcal{L}^1(\mu)$ denotes the collection of all measurable functions with finite $\mathcal{L}^1$-norm. Then,

$$D_\alpha(P\|Q) \geq \frac{\alpha}{\alpha - 1} \log \mathbb{E}_P[e^{(\alpha-1)g(X)}] - \log \mathbb{E}_Q[e^{\alpha g(X))}]. \tag{6}$$

We next provide the following lemma, whose proof is an extension of [7, Lemma 2] and [10, Lemma 1].

Lemma 2: Consider two arbitrary jointly distributed random variables U and V defined on the same probability and taking values in spaces $\mathcal{U}$ and $\mathcal{V}$, respectively. Given a measurable function $h : \mathcal{U} \times \mathcal{V} \to \mathbb{R}$, assume that $h(u, V)$ is $\sigma^2(u)$-sub-Gaussian under P_V and $P_{V|U=u}$ for all $u \in \mathcal{U}$, where $\mathbb{E}[\sigma^2(U)] < \infty$. Then for $\alpha \in (0, 1)$,

$$|\mathbb{E}[h(U, V)] - \mathbb{E}[h(\bar{U}, \bar{V})]| \leq \sqrt{\mathbb{E}[2\sigma^2(U)] \, \frac{D_\alpha(P_{V|U}\|P_V|P_U)}{\alpha}},$$

where $\bar{U}$ and $\bar{V}$ are independent copies of U and V, respectively.

Proof: By the sub-Gaussian property, we have

$$\log \mathbb{E}\left[e^{(\alpha-1)\lambda h(u,V)-\mathbb{E}[(\alpha-1)\lambda h(u,V)|U=u]}|U=u\right] \leq \frac{\lambda^2(\alpha-1)^2\sigma^2(u)}{2} \tag{7}$$

and

$$\log \mathbb{E}[e^{\alpha\lambda h(u,V)-\mathbb{E}[\alpha\lambda h(u,V)]}] \leq \frac{\lambda^2\alpha^2\sigma^2(u)}{2}. \tag{8}$$

Re-arranging the terms gives us

$$-\log \mathbb{E}\left[e^{(\alpha-1)\lambda h(u,V)}|U=u\right] \geq -\frac{\lambda^2(\alpha-1)^2\sigma^2(u)}{2} + \mathbb{E}[(1-\alpha)\lambda h(u,V)|U=u] \tag{9}$$

and

$$-\log \mathbb{E}[e^{\alpha\lambda h(u,V)}] \geq -\frac{\lambda^2\alpha^2\sigma^2(u)}{2} - \mathbb{E}[\alpha\lambda h(u,V)]. \tag{10}$$

Note that by (7) and (8), $e^{(\alpha-1)\lambda h(u,V)} \in \mathcal{L}^1(P_{V|U=u})$ and $e^{\alpha\lambda h(u,V)} \in \mathcal{L}^1(P_V)$. Thus by Rényi's variational formula in (6), we have that

$$D_\alpha(P_{V|U=u}\|P_V) \geq \frac{\alpha}{\alpha-1}\log \mathbb{E}[e^{(\alpha-1)\lambda h(u,V)}|U=u] - \log \mathbb{E}[e^{\alpha\lambda h(u,V)}]. \tag{11}$$

Substituting (9) and (10) in (11) yields

$$\begin{aligned}
&D_\alpha(P_{V|U=u}\|P_V) \geq \\
&\frac{\alpha}{1-\alpha}\left(-\frac{\lambda^2(\alpha-1)^2\sigma^2(u)}{2} + \mathbb{E}[(1-\alpha)\lambda h(u,V)|U=u]\right) \\
&\quad - \frac{\lambda^2\alpha^2\sigma^2(u)}{2} - \mathbb{E}[\alpha\lambda h(u,V)] \\
&= \alpha\lambda(\mathbb{E}[h(u,V)|U=u] - \mathbb{E}[h(u,V)]) - \frac{\lambda^2\alpha(1-\alpha)\sigma^2(u)}{2} \\
&\quad - \frac{\lambda^2\alpha^2\sigma^2(u)}{2} \\
&= \alpha\lambda(\mathbb{E}[h(u,V)|U=u] - \mathbb{E}[h(u,V)]) - \frac{\lambda^2\alpha\sigma^2(u)}{2}.
\end{aligned}$$

The left-hand side of the resulting inequality

$$\frac{\lambda^2\alpha\sigma^2(u)}{2} - \alpha\lambda(\mathbb{E}[h(u,V)|U=u] - \mathbb{E}[h(u,V)]) + D_\alpha(P_{V|U=u}\|P_V) \geq 0$$

is a non-negative quadratic polynomial in λ. Thus the discriminant is non-positive and we have

$$(\alpha\lambda(\mathbb{E}[h(u,V)|U=u] - \mathbb{E}[h(u,V)]))^2 \leq 4\left(\frac{\alpha\sigma^2(u)}{2}\right) D_\alpha(P_{V|U=u}\|P_V).$$

Therefore,

$$|\mathbb{E}[h(u,V)|U=u] - \mathbb{E}[h(u,V)]| \leq \sqrt{\frac{2\sigma^2(u)D_\alpha(P_{V|U=u}\|P_V)}{\alpha}}. \tag{12}$$

Since, $\bar{U}$ and $\bar{V}$ are independent, we have that

$$\mathbb{E}[h(u,V)] = \mathbb{E}[h(\bar{U},\bar{V})|\bar{U}=u].$$

Therefore, we have

$$\begin{aligned}
&|\mathbb{E}[h(U,V)] - \mathbb{E}[h(\bar{U},\bar{V})]| \\
&= \left|\int (\mathbb{E}[h(U,V)|U=u] - \mathbb{E}[h(\bar{U},\bar{V})|\bar{U}=u])P_U(du)\right| \\
&= \left|\int (\mathbb{E}[h(u,V)|U=u] - \mathbb{E}[h(u,V)])P_U(du)\right| \\
&\leq \int \left|\int (\mathbb{E}[h(u,V)|U=u] - \mathbb{E}[h(u,V)]\right| P_U(du) && (13)\\
&\leq \int \sqrt{\frac{2\sigma^2(u)D_\alpha(P_{V|U=u}\|P_V)}{\alpha}} P_U(du) && (14)\\
&\leq \sqrt{\int 2\sigma^2(u)P_U(du)}\sqrt{\int \frac{D_\alpha(P_{V|U=u}\|P_V)}{\alpha}P_U(du)} && (15)\\
&= \sqrt{\mathbb{E}[2\sigma^2(U)]\,\frac{D_\alpha(P_{V|U}\|P_V|P_U)}{\alpha}}, && (16)
\end{aligned}$$

where (13) follows from Jensen's inequality, (14) follows from (12), (15) follows from the Cauchy-Schwarz inequality and the definition of conditional Rényi divergence in (4) with $D_\alpha(P_{V|U}\|P_V|P_U) = \mathbb{E}_U[D_\alpha(P_{V|U}(\cdot|U)\|P_V)]$. ∎

We next use Lemma 2 to derive our main theorem; its proof is a generalization of [10, Theorem 3].

Theorem 1: Let X, Y and Z be random vectors such that $Y \to X \to Z$, as described in Section II. Assume that there exists an optimal estimator f of Y from X such that $l(y, f(X))$ is $\sigma^2(y)$-sub-Gaussian under $P_{X|Z}$ and $P_{X|Z,Y=y}$ for all $y \in \mathbb{R}^p$, i.e.,

$$\log \mathbb{E}[e^{(\lambda(l(y,f(X)))-\mathbb{E}[l(y,f(X))|Z])}|Z] \leq \frac{\sigma^2(y)\lambda^2}{2}$$

and

$$\log \mathbb{E}[e^{(\lambda(l(y,f(X)))-\mathbb{E}[l(y,f(X))|Z,Y=y])}|Z,Y=y] \leq \frac{\sigma^2(y)\lambda^2}{2}$$

for all $\lambda \in \mathbb{R}$ and $y \in \mathbb{R}$, where $\sigma^2 : \mathbb{R} \to \mathbb{R}$, satisfies $\mathbb{E}[\sigma^2(Y)] < \infty$. Then for $\alpha \in (0,1)$, the excess minimum risk satisfies

$$L_l^*(Y|Z) - L_l^*(Y|X) \leq \sqrt{\frac{2\mathbb{E}[\sigma^2(Y)]}{\alpha}\, D_\alpha(P_{X|Y,Z}\|P_{X|Z}|P_{Y,Z})}. \tag{17}$$

Remark 1: If $Y \to X \to Z$ have a joint probability density function f_{YXZ} then in terms of conditional densities, the conditional Rényi's divergence on the right hand side of (17) can be written as

$$\begin{aligned}
&D_\alpha(P_{X|Y,Z}\|P_{X|Z}|P_{Y,Z}) \\
&= \iint \frac{1}{\alpha-1}\log\left(\int (f_{X|Y,Z}(x|y,z))^\alpha \right. \\
&\qquad \left. \times\, (f_{X|Z}(x|z))^{(1-\alpha)}\,dx\right) f_{Y,Z}(y,z)\,dy\,dz.
\end{aligned}$$

Remark 2: One setup (of the many possible) where our sub-Gaussian conditions allow for a more general class of distribution for X, Y and Z is the regression problem with squared loss $l(y,y') = (y-y')^2$, $y,y' \in \mathbb{R}$. Let $Y = m(X) + N$, where X and N are independent real random variables with $\mathbb{E}[N] = 0$, $\mathbb{E}[N^4] < \infty$, and m is a bounded (regression) function, such that $|m(x)| \leq K$. Then the optimal predictor of Y from X is $f(x) = m(x)$ and $l(y, f(X)) \leq (|y| + K)^2$ for all y. Thus $l(y, f(X))$ is $\sigma^2(y) = (|y| + K)^4/4$-sub-Gaussian under $P_{X|Z}$ and $P_{X|Z,Y=y}$, and $\mathbb{E}\big[\sigma^2(Y)\big] < \infty$; see [10, Section 4.2] (note that Z can be arbitrary as long as the Markov-chain condition $Y \to X \to Z$ holds). Thus the conditions of the theorem are satisfied for this setup, while the sub-Gaussian conditions in [7] or [13] fail in this case.

Proof of Theorem 1: Let $\bar{X}$, $\bar{Y}$ and $\bar{Z}$ be random variables such that $P_{\bar{Y}|\bar{Z}} = P_{Y|Z}$, $P_{\bar{X}|\bar{Z}} = P_{X|Z}$, $P_{\bar{Z}} = P_Z$ and $\bar{Y}$ and $\bar{X}$ are conditionally independent given $\bar{Z}$, i.e., $P_{\bar{Y},\bar{X},\bar{Z}} = P_{Y|Z}P_{X|Z}P_Z$.

We apply Lemma 2 by setting $U = Y$, $V = X$ and $h(u,v) = l(y, f(x))$. Consider $\mathbb{E}[l(Y, f(X))|Z = z])$ and $\mathbb{E}[l(\bar{Y}, f(\bar{X}))|Z = z]$ as regular expectations taken with respect to $P_{Y,X|Z=z}$ and $P_{\bar{Y},\bar{X}|\bar{Z}=z}$. Since, $\bar{Y}$ and $\bar{X}$ are conditionally independent given $\bar{Z} = z$ and $P_{\bar{Z}} = P_Z$, we have that

$$\begin{aligned}&|\mathbb{E}[l(Y, f'(X))|Z = z]) - \mathbb{E}[l(\bar{Y}, f(\bar{X}))|Z = z])| \\ &\leq \sqrt{\frac{2\mathbb{E}[\sigma^2(Y)|Z=z]}{\alpha} D_\alpha(P_{X|Y,Z=z}\|P_{X|Z=z}|P_{Y|Z=z})}.\end{aligned} \tag{18}$$

Now,

$$\begin{aligned}&\big|\mathbb{E}[l(Y, f(X))] - \mathbb{E}[l(\bar{Y}, f(\bar{X}))]\big| \leq \\ &\int \big|\mathbb{E}[l(Y, f(X))|Z = z]) - \mathbb{E}[l(\bar{Y}, f(\bar{X}))|Z = z])\big| \, P_Z(dz) \\ &\leq \int \left(\sqrt{\frac{2\mathbb{E}[\sigma^2(Y)|Z=z]}{\alpha}} \right. \\ &\qquad \left. \times \sqrt{D_\alpha(P_{X|Y,Z=z}\|P_{X|Z=z}|P_{Y|Z=z})}\right) P_Z(dz) \\ &\leq \sqrt{2\int \mathbb{E}[\sigma^2(Y)|Z=z]P_Z(dz)} \\ &\qquad \times \sqrt{\int \frac{D_\alpha(P_{X|Y,Z=z}\|P_{X|Z=z}|P_{Y|Z=z})}{\alpha} P_Z(dz)} \\ &= \sqrt{\frac{2\mathbb{E}[\sigma^2(Y)]}{\alpha} D_\alpha(P_{X|Y,Z}\|P_{X|Z}|P_{Y,Z})},\end{aligned} \tag{19}$$

where the first inequality follows from Jensen's inequality and since $P_{\bar{Z}} = P_Z$, the second inequality follows from (18), the third from the Cauchy-Schwarz inequality, and the equality follows from (4). Since, $\bar{Y}$ and $\bar{X}$ are conditionally independent given $\bar{Z}$, we get the Markov chain $\bar{Y} \to \bar{Z} \to \bar{X}$. Then we have

$$\begin{aligned}\mathbb{E}[l(\bar{Y}, f(\bar{X}))]) &\geq L_l^*(\bar{Y}|\bar{X}) \\ &\geq L_l^*(\bar{Y}|\bar{Z}) \\ &= L_l^*(Y|Z),\end{aligned} \tag{20}$$

where the first inequality follows since $\bar{Y} \to \bar{X} \to f(\bar{X})$, the second inequality holds since $\bar{Y} \to \bar{Z} \to \bar{X}$ by construction, and the equality follows since $(\bar{Y}, \bar{Z})$ and (Y, Z) have the same distribution by construction. Since, f is an optimal estimator of Y from X, we also have

$$\mathbb{E}[l(Y, f(X))]) = L_l^*(Y|X). \tag{21}$$

Therefore using (20) and (21) in (19) combined with the fact that $L_l^*(Y|Z) \geq L_l^*(Y|X)$, we arrive at the desired inequality:

$$\begin{aligned}&L_l^*(Y|Z) - L_l^*(Y|X) \\ &\qquad \leq \sqrt{\frac{2\mathbb{E}[\sigma^2(Y)]}{\alpha} D_\alpha(P_{X|Y,Z}\|P_{X|Z}|P_{Y,Z})}.\end{aligned}$$

■

Remark 3: Taking the limit as $\alpha \to 1$ of the right-hand side of (17) in Theorem 1, we have that

$$\begin{aligned}&L_l^*(Y|Z) - L_l^*(Y|X) \\ &\qquad \leq \sqrt{2\mathbb{E}[\sigma^2(Y)] \, D_{\mathrm{KL}}(P_{X|Y,Z}\|P_{X|Z}|P_{Y,Z})} \\ &\qquad = \sqrt{2\mathbb{E}[\sigma^2(Y)] \, (I(X;Y) - I(Z;Y))},\end{aligned} \tag{22}$$

recovering the bound in [10, Theorem 3].

We next give a corollary for bounded loss functions.

Corollary 1: Suppose the loss function l is bounded, i.e., $\|l\|_\infty = \sup_{y,y'} l(y,y') < \infty$. Then for random vectors X, Y and Z such that $Y \to X \to Z$ as described in Section II, we have the following inequality for $\alpha \in (0,1)$ on the excess minimum risk:

$$L_l^*(Y|Z) - L_l^*(Y|X) \leq \frac{\|l\|_\infty}{\sqrt{2}} \sqrt{\frac{D_\alpha(P_{X|Y,Z}\|P_{X|Z}|P_{Y,Z})}{\alpha}}. \tag{23}$$

Proof: We show that the bounded loss function l satisfies the sub-Gaussian property in Theorem 1. Since l is bounded we have that for any $f : \mathbb{R}^q \to \mathbb{R}^p$, $x \in \mathbb{R}^q$ and $y \in \mathbb{R}^p$, $l(y, f(x)) \in [0, \|l\|_\infty]$. Then by Hoeffding's lemma [24], we can write

$$\log \mathbb{E}[e^{(\lambda(l(y,f(X)))}|Z] \leq \mathbb{E}[l(y, f(X))|Z]) + \frac{\|l\|_\infty^2 \lambda^2}{2}$$

and

$$\begin{aligned}&\log \mathbb{E}[e^{(\lambda(l(y,f(X)))}|Z, Y = y] \\ &\qquad \leq \mathbb{E}[l(y, f(X))|Z, Y = y]) + \frac{\|l\|_\infty^2 \lambda^2}{2}\end{aligned}$$

for all $\lambda \in \mathbb{R}$ and $y \in \mathbb{R}$. Rearranging the above inequalities gives us that $l(y, f(X))$ is $\|l\|_\infty^2$-sub-Gaussian under both $P_{X|Z}$ and $P_{X|Z,Y=y}$ for all $y \in \mathbb{R}^p$. Then by (17), we have

$$L_l^*(Y|Z) - L_l^*(Y|X) \leq \frac{\|l\|_\infty}{\sqrt{2}} \sqrt{\frac{D_\alpha(P_{X|Y,Z}\|P_{X|Z}|P_{Y,Z})}{\alpha}}. \tag{24}$$

■

Remark 4: Taking the limit as $\alpha \to 1$ of (24) in Corollary 1 yields the mutual information based bound:

$$L_l^*(Y|Z) - L_l^*(Y|X) \leq \frac{\|l\|_\infty}{\sqrt{2}} \sqrt{D_{KL}(P_{X|Y,Z} \| P_{X|Z} | P_{Y,Z})} = \frac{\|l\|_\infty}{\sqrt{2}} \sqrt{I(X;Y) - I(Z;Y)}, \quad (25)$$

which recovers the bound in [10, Corrolary 1].

Example 1: We consider a concatenation of two BSCs and set X, Y and Z as scalar binary-valued (Bernoulli) random variables. More specifically, we let Y have distribution $P_Y(0) = p = 1 - P_Y(0)$ and be the input of the first BSC with crossover probability ϵ_1. We let X be the resulting output; its distributions is given by $P_X(0) = p(1-\epsilon_1) + (1-p)\epsilon_1 = 1 - P_X(1)$. We then take X as the input of the second BSC with crossover probability ϵ_2 and set Z as the output. This construction yields the Markov chain, $Y \to X \to Z$. Using a $0-1$ loss function (given by $l(y, y') = 1(y \neq y')$, where $1(\cdot)$ is the indicator function) for Corollary 1, we compute the bound in (23) as a function of $\alpha \in (0, 1)$. Figure 1 compares the Rényi based bound in (23) with the mutual information based bound in (25). We note that the Rényi based bound is tighter for the region of α values between about 0.4 and 1; this improvement is similar to the one obtained in the binary example of [7] regarding generalization error.

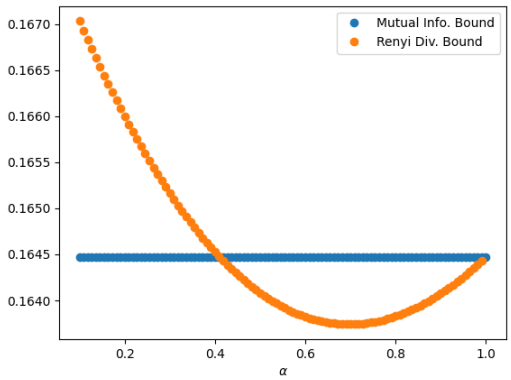

Fig. 1. Comparison of bounds vs α on minimum excess risk for two concatenated BSCs, where p = 0.4, $\epsilon_1 = 0.2$ and $\epsilon_2 = 0.05$.

IV. Conclusion

We derived Rény divergence based bounds (parameterized by the Rényi order $\alpha \in (0, 1)$) on the excess minimum risk using the variational characterization of Rényi's divergence [16] and generalizing the mutual information based bounds recently obtained in [10]. Unlike the related generalization error bounds in [7], [12] where the sub-Gaussian parameter is a fixed constant, the derived bounds involve a sub-Gaussian parameter that can depend on the estimated vector Y and therefore allow for more general joint distributions of the involved random vactors. We also illustrate the upper bounds via an example using a cascade of two BSCs, showing that the Rényi divergence based bounds perform better than the mutual information bounds for certain values of α. Future directions include tightening the Rényi-type bounds as well as identifying other examples where the bounds are sharp.

References

[1] A. Rényi, "On measures of information and entropy," *Proc. 4th Berk. Symp. Math., Stats. Probab.*, Wiley, NY, 1960, pp. 547-561.

[2] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," *Adv. Neural Inf. Process Syst.*, pp. 2521–2530, 2017.

[3] Y. Bu, S. Zou and V. V. Veeravalli, "Tightening mutual information based bounds on generalization error," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 121-130, 2020.

[4] A. R. Esposito, M. Gastpar and I. Issa, "Robust generalization via f-mutual information," *IEEE Symp. Inf. Theory (ISIT)*, 2020.

[5] A. R. Esposito, M. Gastpar and I. Issa, "Robust generalization via α mutual information," *Int. Zurich Sem. Inf. Commun. (IZS)*, 2020.

[6] A. R. Esposito, M. Gastpar and I. Issa, "Generalization error bounds via Rényi-, f-divergences and maximal leakage," *IEEE Trans. Inf. Theory*, vol. 67, no. 8, pp. 4986-5004, 2021.

[7] E. Modak, H. Asnani and V. M. Prabhakaran, "Rényi divergence based bounds on generalization error, " *IEEE Inf. Theory Workshop (ITW)*, Kanazawa, Japan, 2021.

[8] K. Ji, Y. Zhou and Y. Liang, "Understanding estimation and generalization error of generative adversarial networks," *IEEE Trans. Inf. Theory*, vol. 67, no. 5, pp. 3114-3129, 2021.

[9] A. Xu and M Raginsky, "Minimum excess risk in Bayesian learning,"*IEEE Trans. Inf. Theory*, vol. 68, no. 12, pp. 7935-7955, 2022.

[10] L. Györfi, T. Linder and H. Walk, "Lossless transformations and excess risk bounds in statistical inference," *Entropy*, vol. 25, no. 10, p. 1394, 2023.

[11] H. Hafez-Kolahi, B. Moniri and S. Kasaei, "Information-theoretic analysis of minimax excess risk," *IEEE Trans. Inf. Theory*, vol. 69, no. 7, pp. 4659-4674, 2023.

[12] G. Aminian, Y. Bu, L. Toni, M. R. D. Rodrigues and G. W. Wornell, "Information-theoretic characterizations of generalization error for the Gibbs algorithm," *IEEE Trans. Inf. Theory*, vol. 70, no. 1, pp. 632-655, 2024.

[13] G. Aminian, S. Masiha, L. Toni and M. R. D. Rodrigues, "Learning algorithm generalization error bounds via auxiliary distributions," *IEEE J. Sel. Areas Inf. Theory*, 2024, doi: 10.1109/JSAIT.2024.3391900.

[14] R. Atar, K. Chowdhary and P. Dupuis, "Robust bounds on risk sensitive functionals via Rényi divergence," *SIAM/ASA J. Uncertain. Quantif.*, vol. 3, no. 1, pp. 18–33, 2015.

[15] V. Anantharam, "A variational characterization of Rényi divergences," *IEEE Trans. Inf. Theory*, vol. 64, no. 11, pp. 6979-6989, 2018.

[16] J. Birrell, P. Dupuis, M. A. Katsoulakis, L. Rey-Bellet and J. Wang, "Variational representations and neural network estimation for Rényi divergences," *SIAM J. Math. Data Sci.*, vol. 3, no. 4, pp. 1093-1116, 2021.

[17] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observation," *Studia Sci. Math. Hungar.*, vol. 2, pp. 229-318, 1967.

[18] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 145-151, 1991.

[19] M. Welfert, G. R. Kurri, K. Otstot, and L. Sankar, "Addressing GAN training instabilities via tunable classification losses," arXiv preprint arXiv:2310.18291, 2023.

[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative adversarial nets," *Adv. Neural Inf. Process Syst.*, vol. 27, 2014.

[21] T. van Erven and P. Harremoes, "Rényi's divergence and Kullback-Leibler divergence," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 3797-3820, 2014.

[22] S. Verdú, "α-mutual information," *Proc. Workshop Inf. Theory Appl.*, San Diego, 2015.

[23] M. D. Donsker and S. R. S. Varadhan, "Asymptotic evaluation of certain Markov process expectations for large time IV," *Commun. Pure Appl. Math.*, vol. 36, no. 2, pp. 183–212, 1983.

[24] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Am. Stat. Assoc.* vol. 58, no. 301, pp. 13-30, 1963.