Tighter Information-Theoretic Generalization Bounds from Supersamples

Ziqiao Wang [1] **Yongyi Mao** [1]

Abstract

In this work, we present a variety of novel information-theoretic generalization bounds for learning algorithms, from the supersample setting of Steinke & Zakynthinou (2020)—the setting of the "conditional mutual information" framework. Our development exploits projecting the loss pair (obtained from a training instance and a testing instance) down to a single number and correlating loss values with a Rademacher sequence (and its shifted variants). The presented bounds include square-root bounds, fast-rate bounds, including those based on variance and sharpness, and bounds for interpolating algorithms etc. We show theoretically or empirically that these bounds are tighter than all information-theoretic bounds known to date on the same supersample setting.

1. Introduction

Using information-theoretic bounds to analyze the generalization properties of a learning algorithm has attracted increasing attention since the seminal works of (Russo & Zou, 2016; 2019; Xu & Raginsky, 2017). One major advantage of such bounds is that the information-theoretic quantities, e.g., the mutual information (MI) between the training sample and the trained parameter weights, are both distribution-dependent and algorithm-dependent. This makes them an ideal tool to characterize the generalization properties of a learning algorithm, particularly when the traditional algorithm-independent learning-theoretic tools (e.g., VC-dimension (Vapnik, 1998) and Rademacher complexity (Bartlett & Mendelson, 2002)) appear inadequate. For example, Zhang et al. (2017; 2021) show that the high-capacity deep neural networks can still generalize well, contradicting the traditional wisdom in statistical learning theory that suggests complex models tend to overfit the training data and perform poorly on unseen data (Vapnik, 1998). In contrast, the information-theoretic bounds have experimentally demonstrated that they are capable of tracking the generalization behaviour of modern neural networks (Negrea et al., 2019; Wang et al., 2021; Harutyunyan et al., 2021; Wang & Mao, 2022a;b; Hellström & Durisi, 2022a).

The original information-theoretic bound of Xu & Raginsky (2017) has been extended or improved in many different ways, such as the chaining method (Asadi et al., 2018; Hafez-Kolahi et al., 2020; Zhou et al., 2022b; Clerico et al., 2022), the random subset or individual technique (Negrea et al., 2019; Bu et al., 2019; Haghifam et al., 2020; Rodríguez-Gálvez et al., 2021; Zhou et al., 2022a) and so on. Remarkably, Steinke & Zakynthinou (2020) has developed generalization bounds based on a conditional mutual information (CMI) measure obtained for a "supersample" setting. Specifically, the supersample is an $n \times 2$ matrix of data instances. In each row, one instance is selected at random for training and the other is masked out for testing. The authors then show that the CMI of the mask variables and the learned weights conditioned on the supersample can be used to upper-bound the generalization error. Although better behaving than the unconditional weight-based MI bounds (e.g., having boundedness guaranty), the CMI bounds can be difficult to measure for high-dimensional weights, which limits their application. To overcome such difficulty, functional CMI (f-CMI) bounds are proposed by Harutyunyan et al. (2021), where the weight variable in CMI is replaced by the predictions for the supersample. In this case, each prediction pair is a two-dimensional discrete random variable, making the CMI easier to measure and also a tighter bound. More recently, Hellström & Durisi (2022a) uses loss pairs to replace the predictions in f-CMI and obtain even tighter CMI bounds, known as evaluated CMI (e-CMI) bounds. In fact, the earliest version of e-CMI bound appeared in Steinke & Zakynthinou (2020). The notion was also exploited in later works (Haghifam et al., 2021; 2022; 2023). Note that e-CMI still measures the dependence between an one-dimensional variable (mask) and a two-dimensional variable (loss pair). In this work, we show that it is possible to further tighten the CMI bounds, using MI terms involving only two one-dimensional variables.

Our development is restricted to the supersample setting

[1]Department of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Canada. Correspondence to: Ziqiao Wang <zwang286@uottawa.ca>, Yongyi Mao <ymao@uottawa.ca>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

of Steinke & Zakynthinou (2020), on which we establish novel CMI/MI bounds which are all easy to measure and tighter than the existing bounds in the same setting. Specifically, **1)** we first show that the loss pair used in e-CMI can be replaced by the loss difference, giving rise to a disintegrated CMI bound (Theorem 3.1) and an unconditional MI bound (Theorem 3.2). Both are tighter than the previous square-root CMI bounds, all within the context of the same supersample construction. In particular, the obtained unconditional MI term can be interpreted as the *achievable rate* over a memoryless channel in communications. We then show that in the interpolating regime (i.e., training error being zero) and under zero-one loss, the generalization error of the learning algorithm can be precisely expressed by the averaged communication rate (Theorem 3.3). In other words, we obtain the "tightest bound" of generalization error in this setting. We also establish a novel chained MI bound (Theorem 3.4) that is particularly advantageous for continuous and unbounded losses. **2)** Following a symmetric argument for Rademacher process, similar to Zhivotovskiy & Hanneke (2018), we explicitly exploit the symmetric structure of expected generalization error by correlating losses with a Rademacher sequence and obtain a novel MI bound involving single losses (Theorem 4.1). Using the communication perspective, we show that the MI quantities in the bound are upper-bounded by the entropy function evaluated at half of the testing error (Theorem 4.2). **3)** By correlating losses with a shifted Rademacher sequence, we give novel fast-rate MI bounds of the weighted generalization error (Theorem 4.3). **4)** In order to enhance the fast-rate bound in the non-zero training error regime, we extend our analysis by deriving two additional bounds: a variance-based MI bound (Theorem 4.4) and a sharpness-based MI bound (Theorem 4.5). These novel bounds also incorporate symmetric arguments, as shown in Lemma 4.4 and Lemma 4.6, respectively. **5)** Experimental results show that our bounds nicely track the generalization dynamics of both linear models and non-linear neural networks, and our fast-rate bounds are tighter than the binary KL bound proposed in Hellström & Durisi (2022a), the tightest information-theoretic bound known to date for small, non-zero training error. **6)** As a by-product, we also develop a novel Wasserstein distance based bound (Theorem 3.5).

Proofs, additional analysis and experimental results are included in Appendix.

2. Preliminaries

2.1. Probability and Information Theory Notation

Unless otherwise noted, a random variable will be denoted by a capitalized letter, and its realization by the corresponding lower-case letter. Let P_X denote the distribution of a random variable X and let $P_{X|Y}$ be the conditional distribution of X conditioned on Y, which, upon conditioning on a specific realization, is denoted by $P_{X|Y=y}$ or simply $P_{X|y}$. Similarly, $\mathbb{E}_X$ is the expectation taken over $X \sim P_X$ and $\mathbb{E}_{X|Y=y}$ (or $\mathbb{E}_{X|y}$)is the expectation taken over $X \sim P_{X|Y=y}$. Let $H(\cdot)$ be the entropy and let $\mathrm{D}_{\mathrm{KL}}(P||Q)$ denote the KL divergence of P with respect to Q. Let $I(X;Y)$ be the mutual information (MI) between X and Y, and $I(X;Y|Z)$ the conditional mutual information between X and Y conditioned on Z. Following (Negrea et al., 2019), we refer to $I^z(X;Y) \triangleq \mathrm{D}_{\mathrm{KL}}(P_{X,Y|Z=z}||P_{X|Z=z}P_{Y|Z=z})$ as the disintegrated mutual information, and note that $I(X;Y|Z) = \mathbb{E}_Z\left[I^Z(X;Y)\right]$. Also, we use $\mathbb{W}(\cdot,\cdot)$ to denote the Wasserstein distance (formal definition is given in Appendix). Throughout the paper, logarithm takes base e, making the unit of mutual information *nat*.

2.2. Generalization Error

We consider the supervised learning setting. Let $\mathcal{Z} = \mathcal{X}\times\mathcal{Y}$ be the domain of the instances, where $\mathcal{X}$ and $\mathcal{Y}$ are input and label spaces respectively. A model of interest prescribes a family $\mathcal{F}$ of predictors, $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$, where each $f \in \mathcal{F}$ is parameterized by a vector w in some space $\mathcal{W}$. We may write f as f_w as needed. Let μ be the distribution of the instance and let $S = \{Z_i\}_{i=1}^n \overset{i.i.d}{\sim} \mu^n$ be the training sample. There is a learning algorithm $\mathcal{A} : \mathcal{Z}^n \to \mathcal{W}$, which takes the training sample S as the input and outputs a hypothesis $W \in \mathcal{W}$, giving rise to a predictor $f_W \in \mathcal{F}$ that predicts label Y for input X. Note that the algorithm $\mathcal{A}$ is characterized by a conditional distribution $P_{W|S}$. Suppose that the quality of the output hypothesis W is evaluated by a loss function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}_0^+$. Then for a given w, we define the population risk $L_\mu(w) \triangleq \mathbb{E}_{Z'}\left[\ell(w,Z')\right]$, where $Z' \sim \mu$ is a testing instance. The quantity $L_\mu = \mathbb{E}_W\left[L_\mu(W)\right]$ is then the expected population risk. In practice, we cannot access the data distribution μ, so we usually use the empirical risk as a proxy of the population risk, which is defined as $L_S(w) \triangleq \frac{1}{n}\sum_{i=1}^n \ell(w,Z_i)$ for a fixed w. Similarly, $L_n = \mathbb{E}_{W,S}\left[L_S(W)\right]$ is the expected empirical risk, where the expectation is taken over $P_{W,S} = \mu^n \otimes P_{W|S}$. Thus, $\mathrm{Err} \triangleq L_\mu - L_n$ is the expected generalization error.

2.3. Supersample Setting

The CMI framework for bounding generalization errors is first introduced in Steinke & Zakynthinou (2020). Let $\widetilde{Z} \in \mathcal{Z}^{n\times 2}$ be an $n \times 2$ matrix, serving as "supersample", where every entry is drawn i.i.d. from μ. For notational convenience, we assume that the columns of $\widetilde{Z}$ are indexed by $\{0,1\}$ instead of by $\{1,2\}$. We further denote the ith row of $\widetilde{Z}$ as $\widetilde{Z}_i$ with entries $(\widetilde{Z}_{i,0},\widetilde{Z}_{i,1})$. Let $U = (U_1,U_2,\ldots,U_n)^T \sim \mathrm{Unif}(\{0,1\}^n)$, independent of $\widetilde{Z}$, be used to select a training set S from $\widetilde{Z}$: $U_i = 0$ dictates that $\widetilde{Z}_{i,0}$ in $\widetilde{Z}$ be included in the training set S, and $\widetilde{Z}_{i,1}$ be used for testing; U_i=1 dictates the opposite. Then, the constructed training sample S is equivalent to $\widetilde{Z}_U = \{\widetilde{Z}_{i,U_i}\}_{i=1}^n$. Let $\overline{U}_i = 1 - U_i$, then the testing sample is $\widetilde{Z}_{\overline{U}} = \{\widetilde{Z}_{i,\overline{U}_i}\}_{i=1}^n$. In addition, let $L_{i,0} \triangleq \mathbb{E}_{W,\widetilde{Z}_{i,0}}\left[\ell(\mathcal{A}(\widetilde{Z}_U), \widetilde{Z}_{i,0})\right]$ and $L_{i,1}$ defined similarly. We use $L_i = (L_{i,0}, L_{i,1})$ to denote the loss pair in the ith row and $\Delta L_i = L_{i,1} - L_{i,0}$ be the difference in the pair. To avoid clutter, we might use the superscripts $+$ and $-$ to respectively replace the subscripts 0 and 1 in our notations, e.g., $\widetilde{Z}_i^+ = \widetilde{Z}_{i,0}$, $\widetilde{Z}_i^- = \widetilde{Z}_{i,1}$, $L_i^+ = L_{i,0}$ and $L_i^- = L_{i,1}$.

3. Generalization Bounds via Loss Difference

3.1. Loss-Difference CMI Bound

Using the loss difference, we first present the following square-root CMI bound.

Theorem 3.1. *Assume that the loss is bounded between* $[0, 1]$*, we have*

$$|\mathrm{Err}| \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\widetilde{Z}}\sqrt{2I^{\widetilde{Z}}(\Delta L_i; U_i)} \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2I(\Delta L_i; U_i|\widetilde{Z})}.$$

Noting the Markov chain $U - W - f_W(\widetilde{Z}_i) - L_i - \Delta L_i$ (conditioned on $\widetilde{Z}$) and due to the data-processing inequality (DPI), this "loss-difference CMI" (or "ld-CMI") bound in Theorem 3.1 (the second bound) is tighter than the bound in the previous works (Steinke & Zakynthinou, 2020; Haghifam et al., 2020; Harutyunyan et al., 2021; Hellström & Durisi, 2022a), namely, $\underbrace{I(\Delta L_i; U_i|\widetilde{Z})}_{\text{ld}-\text{CMI}} \leq \underbrace{I(L_i; U_i|\widetilde{Z})}_{\text{e}-\text{CMI}} \leq \underbrace{I(f_W(\widetilde{Z}_i); U_i|\widetilde{Z})}_{f-\text{CMI}} \leq \underbrace{I(W; U_i|\widetilde{Z})}_{\text{CMI}}$. It is remarkable that the ld-CMI bound can be significantly tighter. To see this, we re-express the loss function ℓ as a function on $\mathcal{Y}^2 = \mathcal{Y} \times \mathcal{Y}$, where $l(y, y')$ is the loss value of the predicted label y with respect to true label y'. We say that two elements (y_1, y_1') and (y_2, y_2') in $\mathcal{Y}^2$ are *loss-equivalent* and write $(y_1, y_1') \equiv_\ell (y_2, y_2')$ if $\ell(y_1, y_1') = \ell(y_2, y_2')$. It is straightforward to verify that $\equiv_\ell$ is an equivalence relation on $\mathcal{Y}^2$. Let $\mathcal{L}$ denote the image of $\mathcal{Y}^2$ under ℓ. The quotient space $\mathcal{Y}^2/\equiv_\ell$, or the set of equivalence classes modulo $\equiv_\ell$, has a one-to-one correspondence with $\mathcal{L}$, under which we may identify $\mathcal{Y}^2/\equiv_\ell$ with $\mathcal{L}$. Furthermore, we say that two loss pairs (ℓ_A, ℓ_A') and (ℓ_B, ℓ_B') in $\mathcal{L}^2 = \mathcal{L} \times \mathcal{L}$ are *loss-difference-equivalent* and write $(\ell_A, \ell_A') \equiv_\Delta (\ell_B, \ell_B')$ if $\ell_A - \ell_A' = \ell_B - \ell_B'$. Then $\equiv_\Delta$ is likewise an equivalence relation on $\mathcal{L}^2$, which induces the quotient space $\mathcal{L}^2/\equiv_\Delta$. Note that $f_W(\widetilde{Z}_i)$ is a random variable on $\mathcal{Y}^4 = \mathcal{Y}^2 \times \mathcal{Y}^2$ whereas ΔL_i is a essentially a random variable on $\mathcal{L}^2/\equiv_\Delta$, which can be identified with $\left(\mathcal{Y}^2/\equiv_\ell\right)^2/\equiv_\Delta$ under the aforementioned one-to-one correspondence. There can be a significant reduction of space size from $\mathcal{Y}^4$ to $\left(\mathcal{Y}^2/\equiv_\ell\right)^2/\equiv_\Delta$ when $\mathcal{Y}$ or $\mathcal{L}$ is large (assuming they are finite, to fix ideas). Thus, ΔL_i reveals much less information about U_i than $f_w(\widetilde{Z}_i)$ does, making the term $I(\Delta L_i; U_i|\widetilde{Z})$ significantly smaller than $I(f_W(\widetilde{Z}_i); U_i|\widetilde{Z})$ and suggesting that the ld-CMI bound can be much tighter than the f-CMI bound. A similar argument can be made comparing the ld-CMI and the e-CMI bounds.

It is noteworthy that the loss-difference CMI bound is easier to compute than the f-CMI and e-CMI bounds, since ΔL_i is a scalar. Interestingly, when regarding ΔL_i as a (scaled) one-dimensional projection of L_i on a particular direction, the term $I(\Delta L_i; U_i|\widetilde{Z})$ shares some similarity with the notion of *Sliced Mutual Information* (SMI) recently proposed in (Goldfeld & Greenewald, 2021; Goldfeld et al., 2022); the difference is that SMI requires averaging over a random direction of projection.

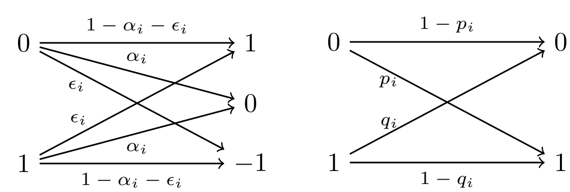


Figure 1. Left: channel from U_i to ΔL_i. Right: channel from U_i to L_i^+. Zero-one loss assumed.

3.2. Loss-Difference MI Bound

Under the setting of supersample as above, we can also obtain a generalization bound based on the loss-difference MI without conditioning on the supersample.

Theorem 3.2. *Assume that* $\ell(\cdot,\cdot) \in [0, 1]$*, then*

$$|\mathrm{Err}| \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2I(\Delta L_i; U_i)}.$$

By the independence of U_i and $\widetilde{Z}$, $I(\Delta L_i; U_i) \leq I(\Delta L_i; U_i) + I(U_i; \widetilde{Z}|\Delta L_i) = I(\Delta L_i; U_i|\widetilde{Z})$. Then the bound in Theorem 3.2 is tighter than ld-CMI bound in Theorem 3.1, although not directly comparable to the first bound in Theorem 3.1.

It is interesting to relate the MI $I(\Delta L_i; U_i)$ to a communication setting where $P_{\Delta L_i|U_i}$ specifies a memoryless channel with input U_i and output ΔL_i. Then $I(\Delta L_i; U_i)$ is the *rate of reliable communication* over this channel achievable with the input distribution P_{U_i} (which is $\mathrm{Bern}(\frac{1}{2})$ by the construction of U) (Shannon, 1948). Consider the special case where $\ell(\cdot,\cdot)$ is the *zero-one loss*, i.e., $\ell(w, z) = \mathbb{1}_{f_w(x)\neq y}$. In this case, $\Delta L_i \in \{-1, 0, 1\}$, and the channel is shown in Figure 1 (left), in which ϵ_i and α_i are transition probabilities as shown on the respective transition edges. In particular, recalling $\Delta L_i = L_i^- - L_i^+$, we see that α_i is the probability that in $\widetilde{Z}_i$ the instance selected from training has the same loss value as that selected for testing, and that ϵ_i is the probability that the training instance in $\widetilde{Z}_i$ has a higher loss value than the testing instance. It follows that any *interpolating algorithm*, namely, one that achieves zero training error must have $\epsilon_i = 0$ for each i. The following theorem can then be proved.

Theorem 3.3. *Under zero-one loss and for any interpolating algorithm $\mathcal{A}$, $I(\Delta L_i; U_i) = (1-\alpha_i)\ln 2$ nats for each i, and $|\mathrm{Err}| = L_\mu = \sum_{i=1}^n \frac{I(\Delta L_i;U_i)}{n\ln 2}$.*

In this case, the generalization error is exactly determined by the communication rate over the channel in Figure 1 (left) averaged over all such channels, making Theorem 3.3 the obviously the "tightest bound" of generalization error in the "interpolating regime". It is of course also tighter than the interpolating bound in Hellström & Durisi (2022a), which may be alternatively seen from $I(\Delta L_i; U_i) \leq I(L_i; U_i|\widetilde{Z})$. Note that Haghifam et al. (2022) also gives a MI quantity that can determine the generalization error in the interpolating case, although their leave-one-out MI is between an $n+1$-dimensional random variable and an one-dimensional random variable, and its corresponding bound is established without exploiting the communication perspective.

Furthermore, it is possible to establish further tightened loss-difference MI bounds for more general loss functions than those required in Theorem 3.2. Specifically, the loss function can be unbounded and continuous, as presented in next theorem, where we apply the chaining technique (Asadi et al., 2018; Hafez-Kolahi et al., 2020; Zhou et al., 2022b; Clerico et al., 2022) and the obtained bound consists of MI terms between U_i and the successively quantized versions of ΔL_i. To that end, let $\mathrm{Err}^i(\Delta\ell_i) \triangleq (-1)^{U_i}\Delta\ell_i$ and let $\Gamma \subseteq \mathbb{R}$ be the range of $\Delta\ell$. Then $\{\mathrm{Err}^i(\Delta\ell_i)\}_{\Delta\ell_i\in\Gamma}$ is a random process[1], applying the *stochastic chaining* method (Zhou et al., 2022b) gives the following chained MI bound.

Theorem 3.4. *For each $i \in [n]$, we assume $\{\Delta L_{i,k}\}_{k=k_0}^\infty$ is a stochastic chain[1] of $(\{\mathrm{Err}^i(\Delta\ell_i)\}_{\Delta\ell_i\in\Gamma}, \Delta L_i)$, then*

$$\mathrm{Err} \leq \frac{1}{n}\sum_{i=1}^n \sum_{k=k_0}^\infty \sqrt{2\mathbb{E}\left[|\Delta L_{i,k} - \Delta L_{i,k-1}|^2\right] I(\Delta L_{i,k}; U_i)},$$

where $\Delta L_{i,k}$ is the kth level of quantization of ΔL_i, the RHS expectation is taken over $(\Delta L_{i,k}, \Delta L_{i,k-1})$.

Notice that the bound is expressed as MI terms each involving U_i and $\Delta L_{i,k}$, both being discrete random variables. This has not arose in the previous chained weight-based MI bounds where they either contain the continuous random variable S (Asadi et al., 2018; Zhou et al., 2022b; Clerico et al., 2022) or are conditioned on the continuous random variable $\widetilde{Z}$ (Hafez-Kolahi et al., 2020). Additionally, by the master definition of MI (Cover & Thomas, 2006, Eq.(8.54)), we know that $I(\Delta L_i; U_i) = \sup_k I(\Delta L_{i,k}; U_i)$, and $I(\Delta L_{i,k}; U_i) \to I(\Delta L_i; U_i)$ when $k \to \infty$.

For bounded loss, the diameter $\mathrm{diam}(\Gamma)$ is finite, we can use hierarchical partitions as in Asadi et al. (2018) to construct a deterministic sequence of $\{\Delta L_{i,k}\}_{k=k_0}^\infty$. This is deferred to Corollary B.1 in Appendix.

3.3. Loss-Difference Bound Beyond CMI and MI

It is possible to develop generalization bounds based on the loss differences in the supersample using distances or divergences beyond the information-theoretic measures. Here we present such a bound based on Wasserstein distance. As investigated in the previous literature (Rodríguez Gálvez et al., 2021), Wasserstein distance usually gives a tighter bounds than the mutual information.

Theorem 3.5. *Assume that $\ell(\cdot,\cdot) \in [0,1]$, then*

$$|\mathrm{Err}| \leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{U_i}\left[\mathbb{W}(P_{\Delta L_i|U_i}, P_{\Delta L_i})\right].$$

Unlike the results in Rodríguez Gálvez et al. (2021), here we do not require the loss to be Lipschitz continuous.

4. Generalization Bounds via Correlating with Rademacher Sequence

We have so far obtained tighter square-root MI bounds based on the information measures (and their variants) between the loss difference ΔL_i and the mask variable U_i. However, the loss difference may not be used to obtain the fast-rate generalization bound where the square root function is removed (Grunwald et al., 2021; Hellström & Durisi, 2021; 2022a). This is because deriving the fast-rate bound usually relies on a weighted generalization error, for which one loses the center-symmetric structure of the standard generalization error. Specifically, knowing ΔL_i and U_i is sufficient to determine the generalization error at ith position by $(-1)^{U_i}\Delta L_i$. However, for the weighted generalization error at ith row defined by $E^i_{C_1} = L_{i,\overline{U}_i} - C_1 L_{i,U_i}$ (for some constant $C_1 > 0$), having U_i and a weighted loss difference $\Delta_{C_1} L_i = L_i^- - C_1 L_i^+$, does not allow its recovering from $(-1)^{U_i}\Delta_{C_1} L_i$ since $L_i^+ - C_1 L_i^- \neq C_1 L_i^+ - L_i^-$ in general. Indeed, knowing both $L_i^- - C_1 L_i^+$ and $L_i^+ - C_1 L_i^-$ requires knowing L_i. Then in order to obtain fast-rate bounds, we need to give up the loss difference and return to the original e-CMI as in Hellström & Durisi (2022a).

Therefore, if we still want to use a MI between two one-dimensional random variables to bound the error, we need to find another trick. This motivates us to use a Rademacher

[1] Some prerequisite definitions of the chaining technique (such as *stochastic chain*, *separable process* and *sub-Gaussian process*) are give in the Appendix A.

viewpoint to derive the CMI bounds. Before we handle the fast-rate CMI bound for the weighted generalization error, we again consider the standard generalization error.

4.1. Single-Loss MI Bounds

Although the CMI setting, particularly its construction of the "ghost sample", is conceptually related to the Rademacher complexity (Bartlett & Mendelson, 2002), the information-theoretic generalization bounds in previous literature do not explicitly exploit this connection. Fortunately, both information-theoretic bounds (Negrea et al., 2019; Hellström & Durisi, 2020; 2021; 2022a) and the Rademacher viewpoint (Kakade et al., 2008; Yang et al., 2019) are shown connected to the PAC-Bayes bounds, we thus derive a variant of e-CMI bound by invoking a similar symmetric argument with (Zhivotovskiy & Hanneke, 2018; Yang et al., 2019).

We first note the following lemma.

Lemma 4.1. *The expected generalization error* $\mathrm{Err} = \frac{2}{n}\sum_{i=1}^n \mathbb{E}_{L_i^+,\varepsilon_i}\left[\varepsilon_i L_i^+\right]$, *where* $\varepsilon_i = (-1)^{\overline{U}_i}$.

Note that $\varepsilon_{1:n} = \{\varepsilon_i\}_{i=1}^n$ is a sequence of Rademacher random variables, and the lemma suggests that $\mathrm{Err} = 2\mathbb{E}_{\widetilde{Z}^+_{1:n}}\mathbb{E}_{\varepsilon_{1:n}}\mathbb{E}_{L^+_{1:n}|\varepsilon_{1:n},\widetilde{Z}^+_{1:n}}\left[\frac{1}{n}\sum_{i=1}^n \varepsilon_i L_i^+\right]$, where $\widetilde{Z}^+_{1:n} = \{\widetilde{Z}^+_i\}_{i=1}^n$ and $L^+_{1:n} = \{L_i^+\}_{i=1}^n$. Then, recall that the Rademacher complexity is defined as $\mathfrak{R}_n(\mathcal{W}) \triangleq \mathbb{E}_S\mathbb{E}_{\varepsilon_{1:n}}\left[\sup_{w\in\mathcal{W}}\frac{1}{n}\sum_{i=1}^n \varepsilon_i \ell(w,Z_i)\right]$ (Bartlett & Mendelson, 2002). Notably, the expected generalization error can be viewed, up to a scale factor 2, as an "average" version of the Rademacher complexity. While Err considers the average correlation between the loss sequence and the Rademacher sequence, the Rademacher complexity measures the worst such correlation. Thus, $\mathrm{Err} \leq 2\mathfrak{R}_n(\mathcal{W})$.

Based on Lemma 4.1, we have the following bound.

Theorem 4.1. *Assume* $\ell(\cdot) \in [0,1]$, *we have*

$$|\mathrm{Err}| \leq \frac{2}{n}\sum_{i=1}^n\sqrt{2I(L_i^+;U_i)} \leq \frac{2}{n}\sum_{i=1}^n\sqrt{2I(f_W(X_i^+);U_i|\widetilde{Z})}.$$

The variable U_i in the above MI/CMI terms can obviously be replaced by ε_i. Thus the theorem can be interpreted as using a different notion of "average correlation", namely mutual information, between losses (or predictions) and Rademacher noises to bound the original notion of average correlation (as stated in Lemma 4.1 and discussed earlier).

This bound may not be directly comparable to others due to the undesired constant of 2 outside of the square root function in the bound. We will soon see that $I(L_i^+;U_i)$ based bound will be more useful when the square root is removed.

For the zero-one loss, the dependence between U_i and L_i^+ is characterized by the communication channel given in Figure 1 (right). In this case, $U_i = 0$ indicates $\widetilde{Z}_i^+$ is selected for training, then p_i is the error rate on this training instance. Similarly, when $U_i = 1$, $\widetilde{Z}_i^+$ is used for testing, then $1 - q_i$ is the error rate on this testing instance. In practice, we usually have $p_i < 1 - p_i$ since L_i^+ is more likely to be zero when $\widetilde{Z}_i^+$ is a training instance, and we may also have $p_i < 1 - q_i$ since L_i^+ is more likely to be one when $\widetilde{Z}_i^+$ is a testing instance compared with the case when $\widetilde{Z}_i^+$ is used in training. When $p_i = 0$, this channel reduces to a Z-channel (Cover & Thomas, 2006). This corresponds to an interpolating algorithm, for which we have the following theorem.

Theorem 4.2. *For zero-one loss and any interpolating algorithm, we have* $\frac{1}{n}\sum_{i=1}^n I(L_i^+;U_i) \leq H(\frac{L_\mu}{2})$.

When the loss is not discrete, we can again obtain a chained MI bound by quantizing the continuous random variable L_i^+, which is given in Theorem C.1 in Appendix C.4.

4.2. Fast-Rate MI Bound

We are now in a position to discuss the weighted generalization error, $\mathrm{Err}_{C_1} \triangleq L_\mu - (1+C_1)L_n$, where C_1 is a prescribed constant. This notion is important for obtaining the fast-rate PAC-Bayes bounds (Catoni, 2007).

To bound this weighted generalization error, similar to Lemma 4.1, we have the following symmetry argument.

Lemma 4.2. *The weighted generalization error can be rewritten as*

$$\mathrm{Err}_{C_1} = \frac{2+C_1}{n}\sum_{i=1}^n \mathbb{E}_{L_i^+,\tilde{\varepsilon}_i}\left[\tilde{\varepsilon}_i L_i^+\right],$$

where $\tilde{\varepsilon}_i = (-1)^{\overline{U}_i} - \frac{C_1}{C_1+2}$ *is a* shifted *Rademacher variable with mean* $-\frac{C_1}{C_1+2}$.

The relationship between Err and Rademacher complexity also likewise extends to that between Err_{C_1} and "shifted Rademacher complexity" defined as $\widetilde{\mathfrak{R}}_n(\mathcal{W}) \triangleq \mathbb{E}_S\mathbb{E}_{\tilde{\varepsilon}_{1:n}}\left[\sup_{w\in\mathcal{W}}\frac{1}{n}\sum_{i=1}^n \tilde{\varepsilon}_i \ell(w,Z_i)\right]$, namely $\mathrm{Err}_{C_1} \leq 2\widetilde{\mathfrak{R}}_n(\mathcal{W})$.

Then, we are ready to present the following bounds.

Theorem 4.3. *Let* $\ell(\cdot,\cdot) \in [0,1]$. *There exist* $C_1, C_2 > 0$ *such that*

$$L_\mu \leq (1+C_1)L_n + \sum_{i=1}^n \frac{I(L_i^+;U_i)}{C_2 n}, \tag{1}$$

$$L_\mu \leq L_n + \sum_{i=1}^n \frac{4I(L_i^+;U_i)}{n} + 4\sqrt{\sum_{i=1}^n \frac{L_n I(L_i^+;U_i)}{n}}. \tag{2}$$

Furthermore, if $\mathcal{A}$ is an interpolating algorithm, we have

$$L_\mu \leq \sum_{i=1}^{n} \frac{2I(L_i^+;U_i)}{n\ln 2}. \tag{3}$$

Notice that Eq. (2) does not depend on C_1, C_2, as it is obtained via minimizing the bound in Eq (1) over a region of (C_1, C_2) in which Eq (1) hold.

Comparing Eq. (3) with the interpolating bound in Hellström & Durisi (2022a, Eq. (12)), the main difference is that their bounds[2] are based on $I(L_i^+, L_i^-; U_i)$, instead of $2I(L_i^+;U_i)$. This difference could be characterized by the *interaction information* (Yeung, 1991), namely $I(L_i^+;U_i;L_i^-) = I(L_i^+;U_i) - I(L_i^+;U_i|L_i^-) = 2I(L_i^+;U_i) - I(L_i;U_i)$ (where the second equality is by the chain rule of MI), and the value $I(L_i^+;L_i^-;U_i)$ could be positive, negative and zero. Hence, the interpolating bound could be further improved as below

$$L_\mu \leq \sum_{i=1}^{n} \frac{\min\{2I(L_i^+;U_i), I(L_i;U_i)\}}{n\ln 2}. \tag{4}$$

This bound is strictly non-vacuous since the RHS of Eq. (2) is upper-bounded by $\frac{\sum_{i=1}^n H(U_i)}{n\ln 2} = 1$. Note that the "tightest bound" of the interpolating algorithm is already obtained in Theorem 3.3.

Previous works (Steinke & Zakynthinou, 2020; Hellström & Durisi, 2022a) suggest that the fast-rate bounds for the weighted generalization error are typically useful when the empirical risk is small or even zero, which may restrict their applications. In the sequel, we introduce two new types of MI bound that can further extend Eq. (1) in Theorem 4.3.

4.3. Variance Based MI Bound

Inspired by the above Rademacher perspective, we first present a new bound that depends on the MI term and a notion of loss variance, defined below.

Definition 4.1 (γ-Variance)**.** *For any $\gamma \in (0,1)$, γ-variance for a learning algorithm is defined as*

$$V(\gamma) \triangleq \mathbb{E}_{W,S}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\ell(W,Z_i) - (1+\gamma)L_S(W)\right)^2\right].$$

By definition, γ-variance also depends on the data distribution. In the zero-one loss case, it can be characterized by the following lemma.

Lemma 4.3. *Under the zero-one loss assumption, we have* $V(\gamma) = L_n - (1-\gamma^2)\mathbb{E}_{W,S}\left[L_S^2(W)\right]$.

[2]Note that Hellström & Durisi (2022a) uses $I(L_i^+, L_i^-; U_i|\widetilde{Z})$ but this CMI term can be strengthened to the unconditional MI by using the same development in this paper.

Loss variances, of any kind, have not appeared in the information-theoretic bounds developed to date. Such a notion however does arise in the PAC-Bayes literature, where such an idea traces back to (Seldin et al., 2012; Tolstikhin & Seldin, 2013). Different from these works, here we utilize an expected empirical variance, and the distribution of W in this case is generated by the learning algorithm rather than the posterior distribution used for prediction in PAC-Bayes.

The gap between Err and $V(\gamma)$ also has a "symmetry lemma" (similar to Lemma 4.2) correlating to the shifted Rademacher sequence.

Lemma 4.4. *For any $C_1 > 0$, we have*

$$\mathrm{Err} - C_1 V(\gamma) \leq \frac{2+C_1\gamma^2}{n}\sum_{i=1}^{n}\mathbb{E}_{L_i^+,\tilde{\varepsilon}_i}\left[\tilde{\varepsilon}_i L_i^+\right],$$

where $\tilde{\varepsilon}_i = \varepsilon_i - \frac{C_1\gamma^2}{C_1\gamma^2+2}$ is the shifted Rademacher variable with mean $-\frac{C_1\gamma^2}{C_1\gamma^2+2}$.

Theorem 4.4. *Assume $\ell(\cdot,\cdot) \in \{0,1\}$, $\gamma \in (0,1)$. Then, there exist $C_1, C_2 > 0$ such that*

$$\mathrm{Err} \leq C_1 V(\gamma) + \sum_{i=1}^{n}\frac{I(L_i^+;U_i)}{nC_2}. \tag{5}$$

Notably, the interpolating setting is a sufficient but not necessary condition for the zero γ-variance, that is, $L_n = 0$ makes $V(\gamma) = 0$, but $V(\gamma) = 0$ does not indicate that $L_n = 0$. In addition, by Lemma 4.3, Eq. (5) can be rewritten as $L_\mu \leq (1+C_1)L_n - C_1(1-\gamma^2)\mathbb{E}_{W,S}\left[L_S^2(W)\right] + \sum_{i=1}^{n}\frac{I(L_i^+;U_i)}{C_2 n}$ so for the fixed C_1 and C_2, the bound of Eq. (5) is tighter than the bound of Eq. (1) with the gap being at least $C_1(1-\gamma^2)\mathbb{E}_{W,S}\left[L_S^2(W)\right]$.

4.4. Sharpness Based MI Bound

The nice generalization property of deep neural networks is often credited to the "flat minima" (Jastrzębski et al., 2017) of loss landscapes. Recently, Neu et al. (2021) and Wang & Mao (2022a) have proved that the generalization error can be upper-bounded by a MI based term plus a sharpness (or flatness) related term. Following the similar development in the previous section, we are able to obtain a bound that also depends on a MI term and a sharpness term, where we use a completely different analysis with (Neu et al., 2021; Wang & Mao, 2022a).

We first define a notion of sharpness.

Definition 4.2 (λ-Sharpness)**.** *For any $\lambda \in (0,1)$, the "λ-sharpness" at position i of the training set is defined as*

$$F_i(\lambda) \triangleq \mathbb{E}_{W,Z_i}\left[\ell(W,Z_i) - (1+\lambda)\mathbb{E}_{W|Z_i}\ell(W,Z_i)\right]^2.$$

This λ-sharpness can be regarded as an expected version of the "flatness" used in Yang et al. (2019) with $W \sim P_{W|Z_i}$ instead of some posterior distribution of W.

Lemma 4.5. *Assume $\ell(\cdot) \in \{0,1\}$, we have $F_i(\lambda) = \mathbb{E}_{W,Z_i}\left[\ell(W,Z_i)\right] - (1-\lambda^2)\mathbb{E}_{Z_i}\left[\mathbb{E}^2_{W|Z_i}\ell(W,Z_i)\right]$.*

Let $F(\lambda) = \frac{1}{n}\sum_{i=1}^n F_i(\lambda)$. Similar to Lemma 4.1, Lemma 4.2 and Lemma 4.4, we have the following symmetric argument.

Lemma 4.6. *For any $C_1 > 0$, we have*

$$\mathrm{Err}-C_1F(\lambda) = \frac{C_1+2}{n}\sum_{i=1}^n \mathbb{E}_{L_i^+,U_i}\left[\tilde{\varepsilon}_i L_i^+ - \frac{C_1(1-\lambda^2)}{C_1+2}\hat{\varepsilon}_i h(U_i)\right],$$

where $\tilde{\varepsilon}_i = \varepsilon_i - \frac{C_1}{C_1+2}$ and $\hat{\varepsilon}_i = \varepsilon_i - 1$ are the shifted Rademacher variables, and $h(U_i) = \mathbb{E}_{\widetilde{Z}_i^+|U_i}\left[\mathbb{E}^2_{L_i^+|\widetilde{Z}_i^+,U_i} L_i^+\right]$.

We are then ready to present the following bound.

Theorem 4.5. *Assume $\ell(\cdot,\cdot) \in \{0,1\}$, $\lambda \in (0,1)$. Then, there exist $C_1, C_2 > 0$ such that*

$$\mathrm{Err} \leq C_1F(\lambda) + \sum_{i=1}^n \frac{I(L_i^+;U_i)}{C_2 n}. \tag{6}$$

Similar to the variance based bound, zero λ-sharpness is a weaker condition than the interpolating assumption. In particular, Eq. (6) could be tighter than Eq. (1) in Theorem 4.3 when the empirical risk is non-zero. Specifically, Eq. (6) can be rewritten as $L_\mu \leq (1+C_1)L_n - C_1(1-\lambda^2)L_n^2 + \sum_{i=1}^n \frac{I(L_i^+;U_i)}{C_2 n}$ by Lemma 4.5 and Jensen's inequality. If C_1, C_2 are fixed, then the sharpness based bound is always tighter than Eq. (1) and the gap is at least $C_1(1-\lambda^2)L_n^2$.

To conclude this section, we give a bound that combines the variance and the sharpness.

Corollary 4.1. *Assume $\ell(\cdot,\cdot) \in \{0,1\}$ and $\gamma, \lambda \in (0,1)$, then there exist $C_1, C_2 > 0$ such that*

$$\mathrm{Err} \leq C_1 \min\{V(\gamma), F(\lambda)\} + \sum_{i=1}^n \frac{I(L_i^+;U_i)}{nC_2}.$$

We remark that if $\mathcal{A}$ satisfies any of the following: (i) $L_n \to 0$; (ii)$V(\gamma) \to 0$ for some $\gamma \in (0,1)$; (iii)$F(\lambda) \to 0$ for some $\lambda \in (0,1)$, we all have $\mathrm{Err} \leq \sum_{i=1}^n \frac{2I(L_i^+;U_i)}{n \ln 2}$.

5. Numerical Results

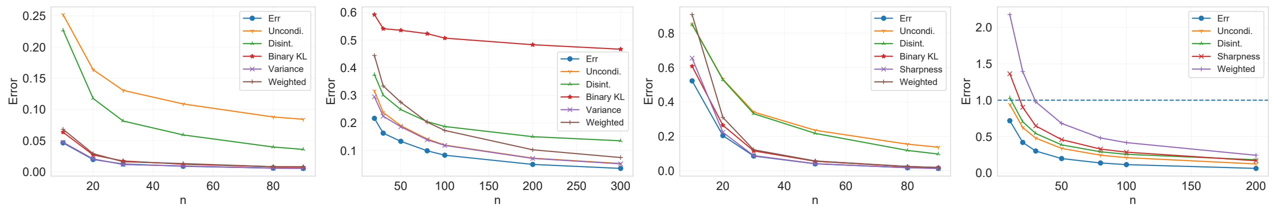
In this section, we empirically compare some CMI and MI bounds discussed in our paper. Our first experiment is based on a synthetic Gaussian dataset, where a simple linear classifier (with a softmax output layer) will be trained. The second experiment follows the same deep learning scenario setting with (Harutyunyan et al., 2021; Hellström & Durisi, 2022a), where we will train a 4-layer CNN on MNIST (LeCun et al., 2010) and fine-tune a ResNet-50 (He et al., 2016) (pretrained on ImageNet (Deng et al., 2009)) on CIFAR10 (Krizhevsky, 2009). In all of these experiments, we let the loss be the zero-one loss, namely $\ell(w,z) = \mathbb{1}_{f_w(x)\neq y}$, and we apply the empirical risk minimization (ERM) to find the hypothesis, namely $w = \arg\min_{w\in\mathcal{W}} L_S(w)$. Since such loss is not differentiable, to enable the gradient based optimization methods such as SGD, we hereby use the cross-entropy loss as a surrogate classification loss during training. Notice that Err is still defined with respect to the zero-one loss in our experiments.

Under these settings, we will mainly compare the disintegrated ld-CMI bound in the first inequality of Theorem 3.1 (*Disint.*), the unconditional MI bound in Theorem 3.2 (*Uncondi.*), the weighted generalization error bound in Eq. (2) of Theorem 4.3 (*Weighted*), the variance bound in Theorem 4.4 (*Variance*) and the sharpness bound in Theorem 4.5 (*Sharpness*). Besides, we will include the binary KL bound in Hellström & Durisi (2022a) as a baseline, which is, to the best of our knowledge, the tightest fast-rate CMI bound in the literature when L_n is close (but not equal) to zero. In addition, we note that the difference between the variance bound and the sharpness bound is negligible in the current scale of the figures, so for each figure we only report one of them. The comparison between the variance bound and the sharpness bound, and more comparison of other bounds mentioned in this paper (such as interpolating bounds and single-loss based square-root bounds) are given in Appendix E.

5.1. Linear Classifier

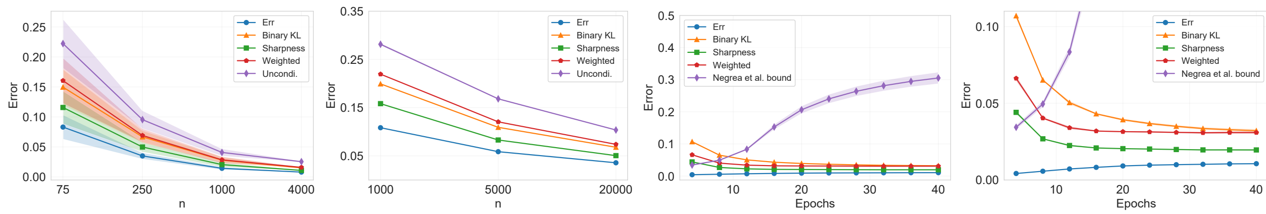
We will first use a simple linear classifier to carry out the Gaussian data classification task (see Appendix E.1 for more details of data generation and training). There are at least two major benefits of using such a synthetic dataset. On the one hand, the ground-truth distribution μ is known so we can draw unlimited supersamples, allowing repeatition of experiments so as to obtain an accurate estimate of the desired quantity (e.g., for each n, we repeat the experiment 5000 times, each with a random $(\widetilde{Z}, U)$ and report the average). On the other hand, the separability of different classes is adjustable, allowing for a control of the task difficulty. Specifically, we will consider both the zero training loss case (i.e. a separable μ) and the high training loss case (i.e. a non-separable μ).

In the binary classification tasks (i.e. $|\mathcal{Y}| = 2$), the evaluations of Err and the bounds are given in Figure 2a and



(a) $|\mathcal{Y}| = 2$ (Realizable) (b) $|\mathcal{Y}| = 2$ (Non-Separable) (c) $|\mathcal{Y}| = 10$ (Realizable) (d) $|\mathcal{Y}| = 10$ (Non-Separable)

Figure 2. Comparison of bounds on the synthetic dataset. (a) Binary classification with a separable μ (i.e. the interpolating setting). Notice that the variance bound nearly coincides with Err. (b) Binary classification with a non-separable μ. (c) Ten-class classification with a separable μ. (d) Ten-class classification with a non-separable μ. The binary KL bound is removed in (d) since it is always ≥ 1.



(a) CNN on MNIST (b) ResNet on CIFAR10 (c) SGLD (MNIST) (d) Zoomed-in of (c)

Figure 3. Comparison of bounds on two real datasets, MNIST (“4 vs 9”) and CIFAR10.

Figure 2b. When μ is separable (Figure 2a), the algorithm can always interpolate the training sample. In this case, the fast-rate bounds are tighter than the square-root bounds, and the variance bound (or the sharpness bound) is the tightest. Moreover, notice that here the the disintegrated CMI bound is tighter than the unconditional MI bound. For a more challenging classification task (Figure 2b), L_n is no longer zero, the square-root bounds become tighter than the binary KL bound. Indeed, Hellström & Durisi (2022a) shows that when the empirical risk is large, the square-root bound will be tighter than their fast-rate bounds. In contrast, our variance bound is even slightly tighter than the square-root bound of Theorem 3.2 in Figure 2b. Additionally, notice that unlike the realizable case (the one with separable μ), the unconditional MI bound is now tighter than the disintegrated CMI bound.

We also conduct experiments in the ten-class classification task (i.e. $|\mathcal{Y}| = 10$), and the results are shown in Figure 2c and Figure 2d. In the realizable case (Figure 2c), the results are similar to binary classification except that the binary KL bound is tighter than all the other bounds when $n = 10$, which is the only case we observe where the binary KL bound outperforms Theorem 4.4 and Theorem 4.5. In addition, it is worth mentioning that Eq. (2) decays much faster than the square-root bounds in Figure 2c (and also in Figure 2b). For the non-separable case in Figure 2d, only the unconditional MI bound in Theorem 3.2 is non-vacuous for all the values of n. While the binary KL bound is removed in this case since it is always vacuous, our sharpness bound is competitive to the square-root bound when $n \geq 50$.

5.2. Neural Networks

To compare information-theoretic generalization bounds of modern deep neural networks, we follow the same experiment settings in (Harutyunyan et al., 2021; Hellström & Durisi, 2022a). Specifically, we train a 4-layer CNN model on a binary MNIST dataset (“4 vs 9”) and also fine-tune a pretrained ResNet-50 on CIFAR10. Unlike the previous synthetic dataset case, here we can only repeatedly run experiments (with different $\widetilde{Z}$ and U) for limited times due to the high computation complexity. Thus, we report the both average numerical values and their standard deviations. Notice that our code is primarily the same as the code provided by Hellström & Durisi (2022a), which is originally based on the code in https://github.com/hrayrhar/f-CMI. More experiment details can be found in Appendix E.1.

Observations in the binary MNIST experiment (Figure 3a) are close to the realizable binary classification case in Figure 2a (both near the interpolating regime). For example, the fast-rate bounds are tighter than the square-root bound. Notably, our sharpness bound (or variance bound) significantly improve the the binary KL bound in both the MNIST experiment (Figure 3a) and the CIFAR10 experiment (Figure 3b), while Eq. (2) is slightly weaker than the binary KL bound. Furthermore, we also compare the bounds when the CNN model is trained by a SGLD algorithm (Raginsky et al., 2017), a variant of SGD, on the binary MNIST dataset.

In this case, we add the weight-based MI bound of SGLD in Negrea et al. (2019, Eq. 6) as a baseline. Figure 3c suggests that both of our sharpness bound and Eq. (2) improve the binary KL bound. Notably, Harutyunyan et al. (2021) and Hellström & Durisi (2022a) observe that f-CMI bound and e-CMI bound are worse than the SGLD bound in Negrea et al. (2019) at the beginning of training. As shown in Figure 3d, although our sharpness bound is still looser than the SGLD bound in Negrea et al. (2019) before the fifth epoch, our sharpness bound significantly shrinks the gap with the SGLD bound during early training.

6. Limitations and Other Related Literature

Limitations More recently, the limitations of information-theoretic bounds in explaining the generalization properties of stochastic convex optimization (SCO) problems have been investigated by Haghifam et al. (2023). In their study, the authors demonstrate that almost all existing information-theoretic bounds, except for the chained MI/CMI bounds, fail to vanish in at least one of their counterexamples, despite the true generalization error vanishing. Unfortunately, neither our loss-difference MI/CMI bounds nor our single-loss MI bounds are capable of overcoming such limitations revealed in their constructed counterexample presented in (Haghifam et al., 2023, Theorem 17). These limitations shed light on certain inherent properties of mutual information measures, which may not be easily overcome solely by introducing new information measures.

In Hellström & Durisi (2022a), the authors provide an e-CMI generalization bound for a generic convex function of the training loss and test loss. Although our analysis, using either loss-difference CMI/MI bounds or single-loss MI bounds, may not be directly applicable to general convex comparison functions between the training loss and testing loss, one potential alternative is to consider the square of the loss difference, for which similar techniques can be employed to derive generalization bounds.

Furthermore, it is important to note that all our new information-theoretic generalization bounds are derived under the assumption of independent and identically distributed (i.i.d.) training instances. Exploring the possibility of relaxing this assumption represents a promising avenue for future research.

Other Related Work Information-theoretic generalization bounds have been explored for some specific algorithms. For example, the weight based information-theoretic bounds have been successfully applied to characterize the generalization properties of SGLD (Pensia et al., 2018; Bu et al., 2019; Negrea et al., 2019; Haghifam et al., 2020; Rodríguez-Gálvez et al., 2021; Wang et al., 2021), and more recently, these bounds are also used to analyze either the vanilla SGD (Neu et al., 2021; Wang & Mao, 2022a) or the stochastic differential equations (SDEs) approximation of SGD (Wang & Mao, 2022b). To apply the weight based MI or CMI bounds for SGD and its variants, unlike the bounds in our paper and (Harutyunyan et al., 2021; Hellström & Durisi, 2022a) that treat the learning algorithm as a black-box, these weight based bounds are usually further upper bounded by some quantities along the trajectories of the algorithms (e.g., gradient incoherence (Negrea et al., 2019)). This then points to a future direction: Can the losses-based or predictions-based information-theoretic bounds be exploited the similar trajectory information of the gradient based algorithms?

It is also noteworthy that recently Haghifam et al. (2022) and Rammal et al. (2022) concurrently propose a variant of the initial CMI framework (Steinke & Zakynthinou, 2020), the "leave-one-out" (LOO) CMI setting, where their supersample is a $n+1$ vector instead of a $n \times 2$ matrix. While our development in this paper is restricted to the latter, it is curious and tempting to compare—or connect—the two.

In addition to the supervised learning setting, information-theoretic bounds have found applicability in various other learning scenarios, showcasing their versatility. These scenarios include meta-learning (Hellström & Durisi, 2022b), semi-supervised learning (He et al., 2022), transfer learning (Wu et al., 2020), domain adaptation (Wang & Mao, 2023), and so on. It is reasonable to expect that our findings can be effectively utilized in these diverse learning settings as well.

7. Concluding Remarks

In this work, we obtain some novel and easy-to-measure information-theoretic generalization bounds. These bounds are demonstrated to be tighter than the previous results in the same supersample setting of Steinke & Zakynthinou (2020), either theoretically or empirically. In our development, we also discuss some other viewpoints of generalization in the current supersample construction including explaining generalization via the rate of reliable communication over the memoryless channel, and via correlating with the Rademacher sequence. These new insights may help to design new learning algorithms or discover novel algorithm-dependent complexity measures.

Acknowledgements

This work is supported partly by an NSERC Discovery grant and a National Research Council of Canada (NRC) Collaborative R&D grant (AI4D-CORE-07). Ziqiao Wang is also supported in part by the NSERC CREATE program through the Interdisciplinary Math and Artificial Intelligence (INTER-MATH-AI) project. The authors would like to thank the anonymous reviewers for their careful reading and valuable suggestions.

References

Asadi, A., Abbe, E., and Verdú, S. Chaining mutual information and tightening generalization bounds. *Advances in Neural Information Processing Systems*, 31, 2018.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Bu, Y., Zou, S., and Veeravalli, V. V. Tightening mutual information based bounds on generalization error. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 587–591. IEEE, 2019.

Catoni, O. Pac-bayesian supervised classification: the thermodynamics of statistical learning. *Vol. 56. Lecture Notes - Monograph Series. Institute of Mathematical Statistics*, 2007.

Clerico, E., Shidani, A., Deligiannidis, G., and Doucet, A. Chained generalisation bounds. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 4212–4257. PMLR, 02–05 Jul 2022.

Cover, T. M. and Thomas, J. A. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.

Cédric, V. *Optimal Transport: Old and New (Grundlehren der mathematischen Wissenschaften, 338)*. Springer, 2008.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, 2009.

Goldfeld, Z. and Greenewald, K. Sliced mutual information: A scalable measure of statistical dependence. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021.

Goldfeld, Z., Greenewald, K., Nuradha, T., and Reeves, G. \$k\$-sliced mutual information: A quantitative study of scalability with dimension. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022.

Grunwald, P., Steinke, T., and Zakynthinou, L. Pac-bayes, mac-bayes and conditional mutual information: Fast rate bounds that handle general vc classes. In *Conference on Learning Theory*. PMLR, 2021.

Hafez-Kolahi, H., Golgooni, Z., Kasaei, S., and Soleymani, M. Conditioning and processing: Techniques to improve information-theoretic generalization bounds. *Advances in Neural Information Processing Systems*, 33:16457–16467, 2020.

Haghifam, M., Negrea, J., Khisti, A., Roy, D. M., and Dziugaite, G. K. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. *Advances in Neural Information Processing Systems*, 2020.

Haghifam, M., Dziugaite, G. K., Moran, S., and Roy, D. Towards a unified information-theoretic framework for generalization. *Advances in Neural Information Processing Systems*, 34:26370–26381, 2021.

Haghifam, M., Moran, S., Roy, D. M., and Dziugiate, G. K. Understanding generalization via leave-one-out conditional mutual information. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pp. 2487–2492. IEEE, 2022.

Haghifam, M., Rodríguez-Gálvez, B., Thobaben, R., Skoglund, M., Roy, D. M., and Dziugaite, G. K. Limitations of information-theoretic generalization bounds for gradient descent methods in stochastic convex optimization. In *International Conference on Algorithmic Learning Theory*, pp. 663–706. PMLR, 2023.

Harutyunyan, H., Raginsky, M., Steeg, G. V., and Galstyan, A. Information-theoretic generalization bounds for black-box learning algorithms. In *Advances in Neural Information Processing Systems*, 2021.

He, H., Yan, H., and Tan, V. Y. Information-theoretic characterization of the generalization error for iterative semi-supervised learning. *Journal of Machine Learning Research*, 23:1–52, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hellström, F. and Durisi, G. Generalization bounds via information density and conditional information density. *IEEE Journal on Selected Areas in Information Theory*, 1 (3):824–839, 2020.

Hellström, F. and Durisi, G. Fast-rate loss bounds via conditional information measures with applications to neural networks. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pp. 952–957. IEEE, 2021.

Hellström, F. and Durisi, G. A new family of generalization bounds using samplewise evaluated CMI. In *Advances in Neural Information Processing Systems*, 2022a.

Hellström, F. and Durisi, G. Evaluated CMI bounds for meta learning: Tightness and expressiveness. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022b.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Jastrzębski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

Kakade, S. M., Sridharan, K., and Tewari, A. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. *Advances in neural information processing systems*, 21, 2008.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

LeCun, Y., Cortes, C., and Burges, C. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Negrea, J., Haghifam, M., Dziugaite, G. K., Khisti, A., and Roy, D. M. Information-theoretic generalization bounds for sgld via data-dependent estimates. *Advances in Neural Information Processing Systems*, 2019.

Neu, G., Dziugaite, G. K., Haghifam, M., and Roy, D. M. Information-theoretic generalization bounds for stochastic gradient descent. In *Conference on Learning Theory*. PMLR, 2021.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Pensia, A., Jog, V., and Loh, P.-L. Generalization error bounds for noisy, iterative algorithms. In *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018.

Polyanskiy, Y. and Wu, Y. Lecture notes on information theory. *Lecture Notes for 6.441 (MIT), ECE 563 (UIUC), STAT 364 (Yale), 2019.*, 2019.

Raginsky, M., Rakhlin, A., and Telgarsky, M. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pp. 1674–1703. PMLR, 2017.

Rammal, M. R., Achille, A., Golatkar, A., Diggavi, S., and Soatto, S. On leave-one-out conditional mutual information for generalization. In *Advances in Neural Information Processing Systems*, 2022.

Rodríguez-Gálvez, B., Bassi, G., Thobaben, R., and Skoglund, M. On random subset generalization error bounds and the stochastic gradient langevin dynamics algorithm. In *2020 IEEE Information Theory Workshop (ITW)*, pp. 1–5. IEEE, 2021.

Rodríguez Gálvez, B., Bassi, G., Thobaben, R., and Skoglund, M. Tighter expected generalization error bounds via wasserstein distance. *Advances in Neural Information Processing Systems*, 34, 2021.

Russo, D. and Zou, J. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*. PMLR, 2016.

Russo, D. and Zou, J. How much does your data exploration overfit? controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1):302–323, 2019.

Seldin, Y., Laviolette, F., Cesa-Bianchi, N., Shawe-Taylor, J., and Auer, P. Pac-bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12):7086–7093, 2012.

Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Steinke, T. and Zakynthinou, L. Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory*. PMLR, 2020.

Tolstikhin, I. O. and Seldin, Y. Pac-bayes-empirical-bernstein inequality. *Advances in Neural Information Processing Systems*, 26, 2013.

Vapnik, V. *Statistical learning theory*. Wiley, 1998. ISBN 978-0-471-03003-4.

Wang, H., Huang, Y., Gao, R., and Calmon, F. Analyzing the generalization capability of sgld using properties of gaussian channels. *Advances in Neural Information Processing Systems*, 34:24222–24234, 2021.

Wang, Z. and Mao, Y. On the generalization of models trained with SGD: Information-theoretic bounds and implications. In *International Conference on Learning Representations*, 2022a.

Wang, Z. and Mao, Y. Two facets of sde under an information-theoretic lens: Generalization of sgd via training trajectories and via terminal states. *arXiv preprint arXiv:2211.10691*, 2022b.

Wang, Z. and Mao, Y. Information-theoretic analysis of unsupervised domain adaptation. In *International Conference on Learning Representations*, 2023.

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

Xu, A. and Raginsky, M. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 2017.

Yang, J., Sun, S., and Roy, D. M. Fast-rate pac-bayes generalization bounds via shifted rademacher processes. *Advances in Neural Information Processing Systems*, 32, 2019.

Yeung, R. W. A new outlook on shannon's information measures. *IEEE transactions on information theory*, 37 (3):466–474, 1991.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Zhivotovskiy, N. and Hanneke, S. Localization of vc classes: Beyond local rademacher complexities. *Theoretical Computer Science*, 742:27–49, 2018.

Zhou, R., Tian, C., and Liu, T. Individually conditional individual mutual information bound on generalization error. *IEEE Transactions on Information Theory*, 68(5): 3304–3316, 2022a.

Zhou, R., Tian, C., and Liu, T. Stochastic chaining and strengthened information-theoretic generalization bounds. In *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2022b.

Appendices

A. Some Useful Definitions and Lemmas

Definition A.1 (Wasserstein Distance)**.** *Let $d(\cdot,\cdot)$ be a metric and let P and Q be probability measures on $\mathcal{X}$. Denote $\Gamma(P,Q)$ as the set of all couplings of P and Q (i.e. the set of all joint distributions on $\mathcal{X}\times\mathcal{X}$ with two marginals being P and Q), then the Wasserstein Distance of order one between P and Q is defined as $\mathbb{W}(P,Q)\triangleq\inf_{\gamma\in\Gamma(P,Q)}\int_{\mathcal{X}\times\mathcal{X}}d(x,x')d\gamma(x,x')$.*

Definition A.2-A.5 are used in the context of chaining methode, e.g., Theorem 3.4, Corollary B.1 and Theorem C.1.

The following is a technique assumption.

Definition A.2 (Separable Process)**.** *The random process $\{X_t\}_{t\in\mathcal{T}}$ is called separable if there is a countable set $T_0\subseteq T$ s.t. $X_t\in\lim_{s\to t,s\in T_0}X_s$ for $\forall t\in T$ a.s., where $x\in\lim_{s\to t,s\in T_0}x_s$ means that there is a sequence (s_n) in T_0 s.t. $s_n\to t$ and $x_{s_n}\to x$.*

Definition A.3 (Sub-Gaussian Process)**.** *The random process $\{X_t\}_{t\in T}$ on the metric space (T,d) is called subgaussian if $\mathbb{E}\left[X_t\right]=0$ for all $t\in T$ and $\mathbb{E}\left[e^{\lambda(X_t-X_s)}\right]\leq e^{\frac{1}{2}\lambda^2d^2(t,s)}$ for all $t,s\in T$, $\lambda\geq 0$.*

Definition A.4 (Stochastic Chain (Zhou et al., 2022b))**.** *Let $(X_\mathcal{T},T)$ be a random process and random variable pair, where T is a random variable in the index set $\mathcal{T}$. A sequence of random variables $\{T_k\}_{k=k_0}^{\infty}$ (with each distributed in $\mathcal{T}$) is called a stochastic chain of the pair $(X_\mathcal{T},T)$, if 1) $\lim_{k\to\infty}\mathbb{E}\left[X_{T_k}\right]=\mathbb{E}\left[X_T\right]$, 2) $\mathbb{E}\left[X_{T_{k_0}}\right]=0$ and 3) $\{X_t\}_{t\in\mathcal{T}}-T-T_k-T_{k-1}$ is a Markov chain for every $k>k_0$.*

Definition A.5 (Increasing Sequence of ϵ-Partition)**.** *A partition $\mathcal{P}=\{A_1,A_2,\dots,A_m\}$ of the set T is called an ϵ-partition of the metric space (T,d) if for all $i=1,2,\dots,m$, A_i can be contained within a ball of radius ϵ. A sequence of partitions $\{\mathcal{P}_k\}_{k=m}^{\infty}$ of a set T is called an increasing sequence if for any $k\geq m$ and each $A\in\mathcal{P}_{k+1}$, there exists $B\in\mathcal{P}_k$ s.t. $A\subseteq B$.*

The following lemmas are foundations of the most proofs in this paper.

Lemma A.1 (Donsker-Varadhan (DV) variational representation of KL divergence (Polyanskiy & Wu, 2019, Theorem 3.5))**.** *Let Q, P be probability measures on Θ, for any bounded measurable function $f:\Theta\to\mathbb{R}$, we have*

$$\mathrm{D_{KL}}(Q||P)=\sup_f\mathbb{E}_{\theta\sim Q}\left[f(\theta)\right]-\ln\mathbb{E}_{\theta\sim P}\left[\exp f(\theta)\right].$$

Lemma A.2 (Hoeffding's Lemma (Hoeffding, 1963))**.** *Let $X\in[a,b]$ be a bounded random variable with mean μ. Then, for all $t\in\mathbb{R}$, we have $\mathbb{E}\left[e^{tX}\right]\leq e^{t\mu+\frac{t^2(b-a)^2}{8}}$.*

Lemma A.3 (Kantorovich-Rubinstein (KR) duality of Wasserstein distance (Cédric, 2008))**.** *For any two distributions P and Q, we have*

$$\mathbb{W}(P,Q)=\sup_{f\in 1-\mathrm{Lip}(\rho)}\int_{\mathcal{X}}fdP-\int_{\mathcal{X}}fdQ,$$

where the supremum is taken over all 1-Lipschitz functions in the metric d, i.e. $|f(x)-f(x')|\leq d(x,x')$ for any $x,x'\in\mathcal{X}$.

The following result is known in the previous work (Xu & Raginsky, 2017).

Lemma A.4 (Xu & Raginsky (2017, Lemma 1))**.** *If $g(X',Y')$ is σ-subgaussian[3] under $P_{X',Y'}=P_XP_Y$, then*

$$\left|\mathbb{E}_{X,Y}\left[g(X,Y)\right]-\mathbb{E}_{X',Y'}\left[g(X',Y')\right]\right|\leq\sqrt{2\sigma^2I(X;Y)}.$$

B. Omitted Proofs and Additional Results in Section 3

B.1. Proof of Theorem 3.1

The following proof shows that the proof of e-CMI bound in Hellström & Durisi (2022a) can be adapted to the loss-difference MI bound, where we just replace the distribution $P_{L_i^+,L_i^-}$ by $P_{\Delta L_i}$.

[3]A random variable X is σ-subgaussian if for any t, $\ln\mathbb{E}\exp\left(t\left(X-\mathbb{E}X\right)\right)\leq t^2\sigma^2/2$.

Proof. Let U_i' be the independent copy of U_i s.t. $U_i' \sim \mathrm{Bern}(1/2)$ and $U' \perp\!\!\!\perp \Delta L_i|\widetilde{Z}$. Recall Lemma A.1,

$$
\begin{aligned}
I(\Delta L_i; U_i|\widetilde{Z}=\tilde{z}) =& \mathrm{D_{KL}}\left(P_{\Delta L_i,U_i|\widetilde{Z}=\tilde{z}}||P_{\Delta L_i|\widetilde{Z}=\tilde{z}}P_{U_i'}\right)\\
&\geq \sup_{t\in\mathbb{R}} \mathbb{E}_{\Delta L_i,U_i|\tilde{z}}\left[tg(\Delta L_i,U_i,\tilde{z})\right] - \ln \mathbb{E}_{\Delta L_i,U_i'|\tilde{z}}\left[e^{tg(\Delta L_i,U_i',\tilde{z})}\right].
\end{aligned}
$$

Recall that $w=\mathcal{A}(\tilde{z}_u)$, we now let $g(\Delta l_i,u_i,\tilde{z}) = (-1)^{u_i}\Delta l_i = (-1)^{u_i}\left(\ell(\mathcal{A}(\tilde{z}_u),\tilde{z}_i^-)-\ell(\mathcal{A}(\tilde{z}_u),\tilde{z}_i^+)\right)$.

Then,

$$
\begin{aligned}
I(\Delta L_i; U_i|\widetilde{Z}=\tilde{z}) \geq& \sup_t \mathbb{E}_{\Delta L_i,U_i|\tilde{z}}\left[t(-1)^{U_i}\Delta L_i\right] - \ln \mathbb{E}_{\Delta L_i,U_i'|\tilde{z}}\left[e^{t(-1)^{U_i'}\Delta L_i}\right]\\
=& \sup_t \mathbb{E}_{\Delta L_i,U_i|\tilde{z}}\left[t(-1)^{U_i}\Delta L_i\right] - \ln \mathbb{E}_{\Delta L_i|\tilde{z}}\left[\mathbb{E}_{U'}\left[e^{t(-1)^{U_i'}\Delta L_i}|\Delta L_i=\Delta l_i\right]\right], \qquad (7)
\end{aligned}
$$

where the last equality is by the conditional independence.

Since $\mathbb{E}_{U'}\left[t(-1)^{U_i'}\Delta l_i\right]=0$ and $(-1)^{U_i'}$ is bounded between $[-1,1]$, by Lemma A.2, we have

$$
\mathbb{E}_{U_i'}\left[e^{t(-1)^{U_i'}\Delta l_i}\right] \leq e^{\frac{t^2\Delta l_i^2}{2}} \leq e^{\frac{t^2}{2}},
$$

where the second inequality is by the boundedness condition of the loss function, i.e. $\Delta l_i\in[-1,1]$.

Plugging the inequality above into Eq. (7), we have

$$
I(\Delta L_i; U_i|\widetilde{Z}=\tilde{z}) \geq \sup_t \mathbb{E}_{\Delta L_i,U_i|\tilde{z}}\left[t(-1)^{U_i}\Delta L_i\right] - \frac{t^2}{2}.
$$

Then consider the case of $t>0$ and $t<0$ ($t=0$ is trivial), by AM–GM inequality (i.e. the arithmetic mean is greater than or equal to the geometric mean), the following is straightforward,

$$
\left|\mathbb{E}_{\Delta L_i,U_i|\tilde{z}}\left[(-1)^{U_i}\Delta L_i\right]\right| \leq \sqrt{2I(\Delta L_i; U_i|\widetilde{Z}=\tilde{z})}.
$$

Notice that

$$
\begin{aligned}
|\mathrm{Err}| = |\mathbb{E}_{S,W}\left[L_\mu(W)-L_S(W)\right]| =& \left|\mathbb{E}_{\widetilde{Z},U,W}\left[L_{\widetilde{Z}\setminus\widetilde{Z}_U}(W)-L_{\widetilde{Z}_U}(W)\right]\right| \qquad (8)\\
\leq& \mathbb{E}_{\widetilde{Z}}\left|\mathbb{E}_{U,W|\widetilde{Z}}\left[L_{\widetilde{Z}_{\overline{U}}}(W)-L_{\widetilde{Z}_U}(W)\right]\right| \qquad (9)\\
\leq& \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\widetilde{Z}}\left|\mathbb{E}_{U_i,W|\widetilde{Z}}\left[(-1)^{U_i}\left(\ell(W,\widetilde{Z}_i^-)-\ell(W,\widetilde{Z}_i^+)\right)\right]\right|\\
=& \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\widetilde{Z}}\left|\mathbb{E}_{U_i,\Delta L_i|\widetilde{Z}}\left[(-1)^{U_i}\Delta L_i\right]\right|,
\end{aligned}
$$

wherein the two inequalities are by applying the Jensen's inequality to the absolute function.

Hence, putting everything together we have

$$
|\mathrm{Err}| \leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\widetilde{Z}}\sqrt{2I^{\widetilde{Z}}(\Delta L_i;U_i)} \leq \frac{1}{n}\sum_{i=1}^n \sqrt{2I(\Delta L_i;U_i|\widetilde{Z})},
$$

where the second inequality is by applying the Jensen's inequality to the square root function.

This completes the proof. ∎

B.2. Proof of Theorem 3.2

By revisiting the proof of Theorem 3.1, particularly Eq. (8-9), we notice that if we do not move the expectation over $\widetilde{Z}$ outside of the absolute function, we will have a chance to get ride of the expectation over $\widetilde{Z}$ if we take the expectation over ΔL_i.

Proof. By the definition of the expected generalization error, we have

$$\begin{aligned}
|\mathrm{Err}| = \left|\mathbb{E}_{S,W}\left[L_\mu(W) - L_S(W)\right]\right| &= \left|\mathbb{E}_{\widetilde{Z},U,W}\left[L_{\widetilde{Z}\setminus\widetilde{Z}_U}(W) - L_{\widetilde{Z}_U}(W)\right]\right| \\
&= \left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\widetilde{Z},U_i,W}\left[(-1)^{U_i}\left(\ell(W,\widetilde{Z}_i^-) - \ell(W,\widetilde{Z}_i^+)\right)\right]\right| \quad (10)\\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{E}_{\Delta L_i,U_i}\left[(-1)^{U_i}\Delta L_i\right]\right|. \quad (11)
\end{aligned}$$

We know that $(-1)^{U_i'}\Delta L_i$ is bounded between $[-1,1]$, so it is a 1-subgaussian random variable. Then, recall Lemma A.4 and let $g(X,Y) = (-1)^{U_i}\Delta L_i$, we have

$$\left|\mathbb{E}_{\Delta L_i,U_i}\left[(-1)^{U_i}\Delta L_i\right] - \mathbb{E}_{\Delta L_i,U_i'}\left[(-1)^{U_i'}\Delta L_i\right]\right| \leq \sqrt{2I(\Delta L_i;U_i)}.$$

Since $\mathbb{E}_{\Delta L_i,U_i'}\left[(-1)^{U_i'}\Delta L_i\right] = 0$, plugging the inequality above into Eq. (11), we have

$$|\mathrm{Err}| \leq \frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{E}_{\Delta L_i,U_i}\left[(-1)^{U_i}\Delta L_i\right]\right| \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{2I(\Delta L_i;U_i)}.$$

This concludes the proof. □

B.3. Proof of Theorem 3.5

Proof. Recall Eq. (11), we could also obtain

$$|\mathrm{Err}| \leq \frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{E}_{\Delta L_i,U_i}\left[(-1)^{U_i}\Delta L_i\right]\right| = \frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{E}_{\Delta L_i,U_i}\left[(-1)^{U_i}\Delta L_i\right] - \mathbb{E}_{\Delta L_i',U_i}\left[(-1)^{U_i}\Delta L_i'\right]\right|,$$

where $\Delta L_i'$ is an independent copy of ΔL_i (i.e. $\Delta L_i' \sim P_{\Delta L_i}$ and $\Delta L_i' \perp\!\!\!\perp U_i$) and the second equality holds since $\mathbb{E}_{\Delta L_i',U_i}\left[(-1)^{U_i}\Delta L_i'\right] = 0$.

Then, by Jensen's inequality, we move the expectation over U_i and the average outside the absolute function,

$$|\mathrm{Err}| \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{U_i}\left[\left|\mathbb{E}_{\Delta L_i|U_i}\left[(-1)^{U_i}\Delta L_i\right] - \mathbb{E}_{\Delta L_i'}\left[(-1)^{U_i}\Delta L_i'\right]\right|\Big|U_i = u_i\right].$$

Notice that for any fixed u_i, we have

$$\begin{aligned}
\mathbb{E}_{\Delta L_i|U_i=u_i}\left[(-1)^{u_i}\Delta L_i\right] &= \int_{-1}^{1}(-1)^{u_i}\Delta\ell_i dP_{\Delta L_i|U_i=u_i}(\Delta\ell_i),\\
\mathbb{E}_{\Delta L_i'}\left[(-1)^{u_i}\Delta L_i'\right] &= \int_{-1}^{1}(-1)^{u_i}\Delta\ell_i' dP_{\Delta L_i}(\Delta\ell_i').
\end{aligned}$$

Also, noting that $f(x) = x$ is a 1-Lipschitz function, i.e. $|(-1)^{u_i}\Delta L_i - (-1)^{u_i}\Delta L_i'| \leq |\Delta L_i - \Delta L_i'|$ holds trivially.

Recall the KR duality of Wasserstein distance (i.e. Lemma A.3), we have

$$|\mathrm{Err}| \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{U_i}\left[\left|\mathbb{E}_{\Delta L_i|U_i}\left[(-1)^{U_i}\Delta L_i\right] - \mathbb{E}_{\Delta L_i'}\left[(-1)^{U_i}\Delta L_i'\right]\right|\Big|U_i = u_i\right] \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{U_i}\left[\mathbb{W}(P_{\Delta L_i|U_i}, P_{\Delta L_i})\right].$$

This concludes the proof. □

B.4. Proof of Theorem 3.3

Proof. The channel capacity can be obtained from Lemma D.1 by letting $\epsilon_i = 0$ and changing the unit of bit to the unit of nat (i.e. replacing logarithm base of 2 to the base of e by $\ln x = \ln 2 \log_2 x$).

Furthermore, the value of $1 - \alpha_i$ reflects the chance that the interpolating learning algorithm $\mathcal{A}$ will make an error (i.e. $\ell(W, Z_i') = 1$) for a testing instance, or equivalently,

$$\mathbb{E}_{W,Z_i'}\left[\mathbb{1}_{f_W(X_i')\neq Y'}\right] = \mathbb{E}_{U_i,L_i}\left[L_{i,\overline{U}_i}\right] = \frac{\mathbb{E}_{L_i^-|U_i=0}\left[L_i^-\right] + \mathbb{E}_{L_i^+|U_i=1}\left[L_i^+\right]}{2}$$
$$= \frac{P(\Delta L_i = 1|U_i = 0) + P(\Delta L_i = -1|U_i = 1)}{2} = 1 - \alpha_i.$$

Thus, combining the equality above with $C = I(U_i; \Delta L_i) = (1 - \alpha_i) \cdot \ln 2$, we have

$$|\mathrm{Err}| = \mathbb{E}_W\left[L_\mu(W)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{W,Z_i'}\left[\mathbb{1}_{f_W(X_i')\neq Y_i'}\right] = \frac{1}{n}\sum_{i=1}^{n}(1-\alpha_i) = \frac{1}{n\ln 2}\sum_{i=1}^{n} I(U_i; \Delta L_i).$$

This completes the proof. $\square$

B.5. Proof of Theorem 3.4

Proof. Let $E^i_{\Delta L_i} = \mathrm{Err}^i(\Delta L_i) = (-1)^{U_i}\Delta L_i$, then for any integers k_1 and k_0 such that $k_1 > k_0$, we have

$$E^i_{\Delta L_i} = E^i_{\Delta L_{i,k_0}} + \sum_{k=k_0+1}^{k_1}\left(E^i_{\Delta L_{i,k}} - E^i_{\Delta L_{i,k-1}}\right) + E^i_{\Delta L_i} - E^i_{\Delta L_{i,k_1}}.$$

By the definition of the stochastic chain (i.e. Definition A.4), we know that $\mathbb{E}_{E^i_{\Delta L_{i,k_0}}}\left[E^i_{\Delta L_{i,k_0}}\right] = 0$ and $\lim_{k_1\to\infty} E^i_{\Delta L_{i,k_1}} = E^i_{\Delta L_i}$. Therefore, let $k_1 \to \infty$ and taking expectation over $(U_i, \Delta L_i) \sim P_{U_i,\Delta L_i}$ for both sides of the equation above, we have

$$\mathbb{E}_{U_i,\Delta L_i}\left[E^i_{\Delta L_i}\right] = \sum_{k=k_0+1}^{\infty}\mathbb{E}_{U_i,\Delta L_{i,k},\Delta L_{i,k-1}}\left[E^i_{\Delta L_{i,k}} - E^i_{\Delta L_{i,k-1}}\right]. \tag{12}$$

Let U_i' be an independent copy of U_i and recall Lemma A.1, we have

$$\begin{aligned}
&\mathbb{E}_{\Delta L_{i,k},\Delta L_{i,k-1}}\left[\mathrm{D_{KL}}\left(P_{U_i|\Delta L_{i,k},\Delta L_{i,k-1}}||P_{U_i'}\right)\right]\\
\geq&\mathbb{E}_{\Delta L_{i,k},\Delta L_{i,k-1}}\left[\sup_{t>0} t\mathbb{E}_{U_i|\Delta L_{i,k},\Delta L_{i,k-1}}\left[E^i_{\Delta L_{i,k}} - E^i_{\Delta L_{i,k-1}}\right] - \ln\mathbb{E}_{U_i'}\left[e^{t\left(E^i_{\Delta L_{i,k}} - E^i_{\Delta L_{i,k-1}}\right)}\right]\right]\\
\geq&\sup_{t>0} t\mathbb{E}_{U_i,\Delta L_{i,k},\Delta L_{i,k-1}}\left[E^i_{\Delta L_{i,k}} - E^i_{\Delta L_{i,k-1}}\right] - \mathbb{E}_{\Delta L_{i,k},\Delta L_{i,k-1}}\ln\mathbb{E}_{U_i'}\left[e^{t(-1)^{U_i'}(\Delta L_{i,k}-\Delta L_{i,k-1})}\right],
\end{aligned}$$

where the second inequality is by applying Jensen's inequality to the supremum.

Notice that the LHS above is equivalent to $I(\Delta L_{i,k}, \Delta L_{i,k-1}; U_i)$. Since $(-1)^{U_i'}$ is bounded between $[-1, 1]$, by Lemma A.2, we have

$$I(\Delta L_{i,k}, \Delta L_{i,k-1}; U_i) \geq \sup_{t>0} t\mathbb{E}_{U_i,\Delta L_{i,k},\Delta L_{i,k-1}}\left[E^i_{\Delta L_{i,k}} - E^i_{\Delta L_{i,k-1}}\right] - \mathbb{E}_{\Delta L_{i,k},\Delta L_{i,k-1}}\ln e^{\frac{t^2(\Delta L_{i,k}-\Delta L_{i,k-1})^2}{2}}.$$

Thus, let $d(\Delta L_{i,k}, \Delta L_{i,k-1}) = |\Delta L_{i,k} - \Delta L_{i,k-1}|$, we have

$$\mathbb{E}_{U_i,\Delta L_{i,k},\Delta L_{i,k-1}}\left[E^i_{\Delta L_{i,k}} - E^i_{\Delta L_{i,k-1}}\right] \leq \sqrt{2\mathbb{E}_{\Delta L_{i,k},\Delta L_{i,k-1}}\left[d^2(\Delta L_{i,k}, \Delta L_{i,k-1})\right] I(\Delta L_{i,k}, \Delta L_{i,k-1}; U_i)}.$$

Plugging the inequality above into Eq. (12) and taking average over i, we have,

$$\begin{aligned}\mathrm{Err} =&\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{U_i,\Delta L_i}\left[E^i_{\Delta L_i}\right]\\ \leq&\frac{1}{n}\sum_{i=1}^n\sum_{k=k_0+1}^{\infty}\sqrt{2\mathbb{E}_{\Delta L_{i,k},\Delta L_{i,k-1}}\left[d^2(\Delta L_{i,k},\Delta L_{i,k-1})\right]I(\Delta L_{i,k},\Delta L_{i,k-1};U_i)}.\end{aligned}$$

From the third point of Definition A.4, we know that $U_i - \Delta L_i - \Delta L_{i,k} - \Delta L_{i,k-1}$ is a Markov chain, so $I(\Delta L_{i,k},\Delta L_{i,k-1};U_i) = I(\Delta L_{i,k};U_i) + I(\Delta L_{i,k-1};U_i|\Delta L_{i,k}) = I(\Delta L_{i,k};U_i)$. This gives us the final form of the bound,

$$\mathrm{Err} \leq \frac{1}{n}\sum_{i=1}^n\sum_{k=k_0+1}^{\infty}\sqrt{2\mathbb{E}_{\Delta L_{i,k},\Delta L_{i,k-1}}\left[d^2(\Delta L_{i,k},\Delta L_{i,k-1})\right]I(\Delta L_{i,k};U_i)}.$$

This concludes the proof. □

B.6. Additional Result: Chained MI Bound for Bounded Loss

Corollary B.1. *Let $2^{-k_0} \geq \mathrm{diam}(\Gamma)$ and let $\{\mathcal{P}_k\}_{k=k_0}^{\infty}$ be an increasing sequence of partitions of Γ, where for each $k \geq k_0$, $\mathcal{P}_k$ is a 2^{-k}-partition of (Γ, d). Let $\Delta L_{i,k}$ be the center of the covering ball of the partition cell that ΔL_i belongs to the partition $\mathcal{P}_k$, then*

$$\mathrm{Err} \leq \frac{3}{n}\sum_{i=1}^n\sum_{k=k_0}^{\infty}2^{-k}\sqrt{2I(\Delta L_{i,k};U_i)}.$$

Proof. By the triangle inequality, we have $d(\Delta L_{i,k},\Delta L_{i,k-1}) \leq d(\Delta L_{i,k},\Delta L_i) + d(\Delta L_i,\Delta L_{i,k-1})$.

Since $\mathcal{P}_k$ is a 2^{-k} partition, $d(\Delta L_{i,k},\Delta L_i) \leq 2^{-k}$, then $d(\Delta L_{i,k},\Delta L_i) + d(\Delta L_i,\Delta L_{i,k-1}) \leq 2^{-k} + 2^{-(k-1)} = 3\times 2^{-k}$. Plugging this into Theorem 3.4, we have

$$\mathrm{Err} \leq \frac{1}{n}\sum_{i=1}^n\sum_{k=k_0+1}^{\infty}\sqrt{2\mathbb{E}_{\Delta L_{i,k},\Delta L_{i,k-1}}\left[d^2(\Delta L_{i,k},\Delta L_{i,k-1})\right]I(\Delta L_{i,k};U_i)} \leq \frac{1}{n}\sum_{i=1}^n\sum_{k=k_0+1}^{\infty}3\times 2^{-k}\sqrt{2I(\Delta L_{i,k};U_i)}.$$

This completes the proof. □

C. Omitted Proofs and Additional Results in Section 4

C.1. Proof of Lemma 4.1

Proof. By the definition of Err, we can decompose it into two terms,

$$\begin{aligned}\mathrm{Err} = \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{L_i^+,L_i^-,U_i}\left[L_{i,\overline{U}_i} - L_{i,U_i}\right] =&\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{L_i^+,L_i^-,U_i}\left[(-1)^{U_i}(L_i^- - L_i^+)\right]\\ =&\frac{1}{n}\sum_{i=1}^n\left[\mathbb{E}_{L_i^-,U_i}\left[(-1)^{U_i}L_i^-\right] + \mathbb{E}_{L_i^+,U_i}\left[-(-1)^{U_i}L_i^+\right]\right]\\ =&\frac{1}{n}\sum_{i=1}^n\left[\mathbb{E}_{L_i^-,U_i}\left[(-1)^{U_i}L_i^-\right] + \mathbb{E}_{L_i^+,\overline{U}_i}\left[(-1)^{\overline{U}_i}L_i^+\right]\right],\end{aligned}$$

where the last equality is by $-(-1)^{U_i}L_i^+ = (-1)^{\overline{U}_i}L_i^+$. We now show that the following holds

$$\mathbb{E}_{L_i^-,U_i}\left[(-1)^{U_i}L_i^-\right] = \mathbb{E}_{L_i^+,\overline{U}_i}\left[(-1)^{\overline{U}_i}L_i^+\right].$$

Recall that $\widetilde{Z}$ and U are i.i.d drawn from μ^{2n} and the Bernoulli distribution, respectively, and $\widetilde{Z} \perp\!\!\!\perp U$. Usually, a learning algorithm may depend on the order of training instances (i.e. for $i \neq j$, even if two training instances satisfy

$z_i = z_j$, $P_{W|z_i}$ may not be the same with $P_{W|z_j}$), but it should be invariant to the *row index* of the supersample $\widetilde{Z}$, then the distribution $P_{L_i^-,U_i}$ and the distribution $P_{L_i^+,U_i}$ have some symmetric property, namely, $P_{L_i^-|U_i=1} = P_{L_i^+|U_i=0}$ and $P_{L_i^-|U_i=0} = P_{L_i^+|U_i=1}$. Or equivalently, the distribution of the training loss (or testing loss) of the ith training instance (or testing instance) is invariant to U_i. Hence, we have $P_{L_i^-} = P_{L_i^+}$, we say L_i^- and L_i^+ are identically but not independently distributed. Then,

$$\mathbb{E}_{L_i^-|U_i=0}\left[L_i^-\right] = \int_0^1 \ell_i^- dP_{L_i^-|U_i=0}(\ell_i^-) = \int_0^1 \ell_i^+ dP_{L_i^+|U_i=1}(\ell_i^+) = \mathbb{E}_{L_i^+|U_i=1}\left[L_i^+\right],$$
$$\mathbb{E}_{L_i^-|U_i=1}\left[-L_i^-\right] = \int_0^1 -\ell_i^- dP_{L_i^-|U_i=1}(\ell_i^-) = \int_0^1 -\ell_i^+ dP_{L_i^+|U_i=0}(\ell_i^+) = \mathbb{E}_{L_i^+|U_i=0}\left[-L_i^+\right].$$

These give us

$$\mathbb{E}_{L_i^-,U_i}\left[(-1)^{U_i}L_i^-\right] = \frac{\mathbb{E}_{L_i^-|U_i=0}\left[L_i^-\right] + \mathbb{E}_{L_i^-|U_i=1}\left[-L_i^-\right]}{2} = \frac{\mathbb{E}_{L_i^+|U_i=1}\left[L_i^+\right] + \mathbb{E}_{L_i^+|U_i=0}\left[-L_i^+\right]}{2}$$
$$= \mathbb{E}_{L_i^+,\overline{U}_i}\left[(-1)^{\overline{U}_i}L_i^+\right].$$

Therefore,

$$\mathrm{Err} = \frac{1}{n}\sum_{i=1}^n \left[\mathbb{E}_{L_i^-,U_i}\left[(-1)^{U_i}L_i^-\right] + \mathbb{E}_{L_i^+,\overline{U}_i}\left[(-1)^{\overline{U}_i}L_i^+\right]\right] = \frac{2}{n}\sum_{i=1}^n \mathbb{E}_{L_i^-,U_i}\left[(-1)^{U_i}L_i^-\right] = \frac{2}{n}\sum_{i=1}^n \mathbb{E}_{L_i^+,\overline{U}_i}\left[(-1)^{\overline{U}_i}L_i^+\right].$$

Notice that both $(-1)^{U_i}$ and $(-1)^{\overline{U}_i}$ are Rademacher variables. This conclude the proof. $\square$

C.2. Proof of Theorem 4.1

Proof. Notice that $2(-1)^{U_i'}L_i^+$ is bounded between $[-2, 2]$, it is a subgaussian random variable with the variance proxy $\sigma = 2$. Let the measurable function $g(L_i^+, U_i)$ in Lemma A.4 be $2(-1)^{U_i}L_i^+$, then $g(L_i^+, U_i)$ is 2-subgaussian under $P_{U_i}P_{L_i^+}$, we have

$$\left|2\mathbb{E}_{L_i^+,U_i}\left[(-1)^{U_i}L_i^+\right] - 2\mathbb{E}_{L_i^+,U_i'}\left[(-1)^{U_i'}L_i^+\right]\right| \leq 2\sqrt{2I(L_i^+;U_i)}.$$

Since $\mathbb{E}_{L_i^+,U_i'}\left[(-1)^{U_i'}L_i^+\right] = \frac{\mathbb{E}_{L_i^+}[L_i^+] - \mathbb{E}_{L_i^+}[L_i^+]}{2} = 0$, then

$$\left|2\mathbb{E}_{L_i^+,U_i}\left[(-1)^{U_i}L_i^+\right]\right| \leq 2\sqrt{2I(L_i^+;U_i)}.$$

Recall Lemma 4.1,

$$\mathrm{Err} = \frac{2}{n}\sum_{i=1}^n \mathbb{E}_{L_i^+,\overline{U}_i}\left[(-1)^{\overline{U}_i}L_i^+\right].$$

Notice that $\overline{U}_i$ and U_i are one-to-one mapping, so using any of them will give the same mutual information. Thus, by applying Jensen's inequality to the absolute function, we have

$$|\mathrm{Err}| \leq \frac{1}{n}\sum_{i=1}^n \left|2\mathbb{E}_{L_i^+,U_i}\left[(-1)^{U_i}L_i^+\right]\right| \leq \frac{2}{n}\sum_{i=1}^n \sqrt{2I(L_i^+;U_i)}.$$

In addition, to obtain the second inequality in the theorem, we first invoke the independence between $\widetilde{Z}$ and U_i, $I(L_i^+;U_i) + I(\widetilde{Z};U_i|L_i^+) = I(L_i^+,\widetilde{Z};U_i) = I(L_i^+;U_i|\widetilde{Z})$ and then use the DPI, $I(L_i^+;U_i|\widetilde{Z}) \leq I(f_W(X);U_i|\widetilde{Z})$, we have

$$|\mathrm{Err}| \leq \frac{2}{n}\sum_{i=1}^n \sqrt{2I(L_i^+;U_i)} \leq \frac{2}{n}\sum_{i=1}^n \sqrt{2I(f_W(X_i^+);U_i|\widetilde{Z})}.$$

This completes the proof. $\square$

C.3. Proof of Theorem 4.2

Proof. Notice that

$$\mathbb{E}_{W,Z'_i}\left[\mathbb{1}_{f_W(X'_i)\neq Y'}\right]=\mathbb{E}_{U_i,L_i}\left[L_{i,\overline{U}_i}\right]=\frac{\mathbb{E}_{L^-_i|U_i=0}\left[L^-_i\right]+\mathbb{E}_{L^+_i|U_i=1}\left[L^+_i\right]}{2}=\mathbb{E}_{L^+_i|U_i=1}\left[L^+_i\right]=P(L^+_i=1|U_i=1)=1-q_i.$$

Hence, $L_\mu=\sum_{i=1}^n\frac{1-q_i}{n}$.

For each i, we have $I(L^+_i;U_i)=H(L^+_i)-H(L^+_i|U_i)=H(\frac{1-q_i}{2})-\frac{1}{2}H(1-q_i)\leq H(\frac{1-q_i}{2})$. Since the entropy function $H(\cdot)$ is a concave function, we have $\frac{1}{n}\sum_{i=1}^n I(L^+_i;U_i)\leq\frac{1}{n}\sum_{i=1}^n H(\frac{1-q_i}{2})\leq H(\frac{L_\mu}{2})$. □

C.4. Additional Result: Chained MI Bound Based on Single-Loss

When the loss is not discrete or even not bounded, let $\xi^i(\ell^+_i)\triangleq\varepsilon_i\ell^+_i$ be a random process and let $\mathcal{L}$ be the domain of L^+_i. Similar to Theorem 3.4, we can also have the corresponding chained bound of Theorem 4.1.

Theorem C.1. *For each $i\in[n]$, we assume $\{L^+_{i,k}\}_{k=k_0}^\infty$ is a stochastic chain of $(\{\xi^i(\ell^+_i)\}_{\ell^+_i\in\mathcal{L}},L^+_i)$, then*

$$\mathrm{Err}\leq\frac{2}{n}\sum_{i=1}^n\sum_{k=k_0}^\infty\sqrt{2\mathbb{E}\left[d^2(L^+_{i,k},L^+_{i,k-1})\right]I(L^+_{i,k};U_i)},$$

where the RHS expectation is taken over $(L^+_{i,k},L^+_{i,k-1})$.

This theorem can be obtained by following the same development with the proof in Section B.5.

C.5. Proof of Lemma 4.2

Proof. A key step is the second equality of the following

$$\begin{aligned}\mathrm{Err}_{C_1}=&\frac{1}{n}\sum_{i=1}^n\mathbb{E}_{L^-_i,L^+_i,U_i}\left[L_{i,\overline{U}_i}-(1+C_1)L_{i,U_i}\right]\\=&\frac{1}{n}\sum_{i=1}^n\mathbb{E}_{L^-_i,L^+_i,U_i}\left[\left(1+\frac{C_1}{2}\right)(L_{i,\overline{U}_i}-L_{i,U_i})-\frac{C_1}{2}L_{i,\overline{U}_i}-\frac{C_1}{2}L_{i,U_i}\right]\\=&\frac{2+C_1}{2n}\sum_{i=1}^n\left[\mathbb{E}_{L^-_i,U_i}\left[(-1)^{U_i}L^-_i-\frac{C_1}{C_1+2}L^-_i\right]+\mathbb{E}_{L^+_i,U_i}\left[-(-1)^{U_i}L^+_i-\frac{C_1}{C_1+2}L^-_i\right]\right].\end{aligned}$$

Recall that $P_{L^-_i}=P_{L^+_i}$, we have $\mathbb{E}_{L^-_i}\left[\frac{C_1}{C_1+2}L^-_i\right]=\mathbb{E}_{L^+_i}\left[\frac{C_1}{C_1+2}L^+_i\right]$. Also, noting that $\mathbb{E}_{L^-_i|U_i=0}\left[L^-_i\right]=\mathbb{E}_{L^+_i|U_i=1}\left[L^+_i\right]$ and $\mathbb{E}_{L^-_i|U_i=1}\left[-L^-_i\right]=\mathbb{E}_{L^+_i|U_i=0}\left[-L^+_i\right]$, we have

$$\mathrm{Err}_{C_1}=\frac{2+C_1}{n}\sum_{i=1}^n\mathbb{E}_{L^+_i,\overline{U}_i}\left[(-1)^{\overline{U}_i}L^+_i-\frac{C_1}{C_1+2}L^+_i\right]=\frac{2+C_1}{n}\sum_{i=1}^n\mathbb{E}_{L^+_i,\tilde{\varepsilon}_i}\left[\tilde{\varepsilon}_iL^+_i\right],\tag{13}$$

where $\tilde{\varepsilon}_i=\varepsilon_i-\frac{C_1}{C_1+2}$ and $\varepsilon_i\sim\mathrm{Unif}\{-1,+1\}$ is the Rademacher variable. In this case, $\tilde{\varepsilon}_i$ is called the *shifted* Rademacher variable and its mean is $-\frac{C_1}{C_1+2}$. □

C.6. Proof of Theorem 4.3

Proof. Since $\tilde{\varepsilon}_i$ is obtained by a bijection function of U_i, they two can be replaced by each other in the mutual information. Recall Lemma A.1 and let the measurable function g be $t(C_1+2)\tilde{\varepsilon}_iL^+_i$, we have

$$I(L^+_i;U_i)=I(L^+_i;\tilde{\varepsilon}_i)=\mathrm{D_{KL}}\left(P_{L^+_i,\tilde{\varepsilon}_i}||P_{L^+_i}P_{\tilde{\varepsilon}'_i}\right)\tag{14}$$

$$\geq\sup_{t>0}\mathbb{E}_{L^+_i,\tilde{\varepsilon}_i}\left[t(C_1+2)\tilde{\varepsilon}_iL^+_i\right]-\ln\mathbb{E}_{L^+_i,\tilde{\varepsilon}'_i}\left[e^{t(C_1+2)\tilde{\varepsilon}'_iL^+_i}\right].\tag{15}$$

We hope to have

$$\mathbb{E}_{L_i^+,\tilde{\varepsilon}_i'}\left[e^{t(C_1+2)\tilde{\varepsilon}_i' L_i^+}\right] \leq 1. \tag{16}$$

Since $\tilde{\varepsilon}_i'$ is independent of L_i^+, and $P(\tilde{\varepsilon}_i = \frac{2}{C_1+2}) = P(\tilde{\varepsilon}_i = \frac{-2(C_1+1)}{C_1+2}) = \frac{1}{2}$, then

$$\mathbb{E}_{L_i^+,\tilde{\varepsilon}_i'}\left[e^{t(C_1+2)\tilde{\varepsilon}_i' L_i^+}\right] = \frac{\mathbb{E}_{L_i^+}\left[e^{-2t(C_1+1)L_i^+} + e^{2tL_i^+}\right]}{2}.$$

Notice that $e^{-2t(C_1+1)L_i^+} + e^{2tL_i^+}$ is the summation of two convex function, which is still a convex function, so the maximum value of this function is achieved at the endpoints of the bounded domain. Recall that $L_i^+ \in [0,1]$, we now consider two cases, i) when $L_i^+ = 0$, we have $e^{-2t(C_1+1)L_i^+} + e^{2tL_i^+} = 2$; ii)when $L_i^+ = 1$, we need to require $e^{-2t(C_1+1)} + e^{2t} \leq 2$ s.t. Eq (16) can hold. Note that this inequality implies that $t \leq \frac{\ln 2}{2}$.

Replacing t by C_2, let the values of C_1, C_2 be taken from the domain of $\{C_1, C_2 | C_1, C_2 > 0, e^{-2C_2(C_1+1)} + e^{2C_2} \leq 2\}$, so Eq. (16) will hold. Under this condition, by re-arranging the terms in Eq. (15), we have

$$(C_1+2)\mathbb{E}_{L_i^+,\tilde{\varepsilon}_i}\left[\tilde{\varepsilon}_i L_i^+\right] \leq \frac{I(L_i^+;U_i)}{C_2}.$$

Plugging the inequality above into Eq. (13), we have

$$\mathrm{Err}_{C_1} = L_\mu - (1+C_1)L_n = \frac{2+C_1}{n}\sum_{i=1}^{n}\left[\mathbb{E}_{L_i^+,\tilde{\varepsilon}_i}\left[\tilde{\varepsilon}_i L_i^+\right]\right] \leq \sum_{i=1}^{n}\frac{I(L_i^+;U_i)}{C_2 n}.$$

Thus, the following inequality can be obtained,

$$L_\mu \leq \min_{C_1,C_2>0, e^{2C_2}+e^{-2C_2(C_1+1)}\leq 2}(1+C_1)L_n + \sum_{i=1}^{n}\frac{I(L_i^+;U_i)}{C_2 n}. \tag{17}$$

We can also optimize the parameters C_1, C_2 by relaxing the condition of $e^{-2C_2(C_1+1)} + e^{2C_2} \leq 2$. By invoking $e^x \geq x+1$ and $e^{-x} \leq \frac{1}{1+x}$ for $x > -1$, and $e^x \leq \frac{1}{1-x}$ for $x < 1$, it's sufficient to have

$$\frac{1}{1+2C_2(C_1+1)} + \frac{1}{1-2C_2} \leq 2, \qquad \text{and } 0 < C_2 < \frac{1}{2}. \tag{18}$$

Solving Eq (18) gives us $0 < C_2 \leq \frac{C_1}{4(C_1+1)}$. Since $\frac{C_1}{4(C_1+1)} \leq \frac{1}{4} < \frac{1}{2}$, then Eq (18) holds when $C_2 \in (0, \frac{C_1}{4(C_1+1)}]$. Notice that $\frac{C_1}{4(C_1+1)}$ is also smaller than $\frac{\ln 2}{2}$.

Therefore, we obtain

$$\begin{aligned}
L_\mu \leq & \min_{C_1,C_2>0, e^{2C_2}+e^{-2C_2(C_1+1)}\leq 2}(1+C_1)L_n + \sum_{i=1}^{n}\frac{I(L_i^+;U_i)}{C_2 n} \\
\leq & \min_{C_1>0, 0<C_2\leq\frac{C_1}{4(C_1+1)}}(1+C_1)L_n + \sum_{i=1}^{n}\frac{I(L_i^+;U_i)}{C_2 n} \\
= & L_n + \sum_{i=1}^{n}\frac{4I(L_i^+;U_i)}{n} + 4\sqrt{\sum_{i=1}^{n}\frac{L_n I(L_i^+;U_i)}{n}},
\end{aligned}$$

where the last equality is achieved when $C_1 = 2\sqrt{\sum_{i=1}^{n}\frac{I(L_i^+;U_i)}{nL_n}}$ and $C_2 = \frac{C_1}{4(C_1+1)}$.

For the second part of the theorem, if $\mathcal{A}$ is an interpolating algorithm, then $L_n = 0$, in which case we can let C_1 be arbitrarily large.

Recall that we hope

$$e^{-2C_2(C_1+1)} + e^{2C_2} \leq 2.$$

This can be satisfied by letting $C_2 = \frac{\ln 2}{2}$ and $C_1 \to \infty$.

Thus, the interpolating single-loss MI bound is

$$L_\mu \leq \sum_{i=1}^{n} \frac{2I(L_i^+; U_i)}{n \ln 2}.$$

This completes the proof. □

C.7. Proof of Lemma 4.3

Proof. By the definition of γ-variance, and notice that $L_S(W) = \frac{1}{n}\sum_{i=1}^{n} \ell(W, Z_i)$, we have

$$\begin{aligned}
V(\gamma) =& \mathbb{E}_{W,S}\left[\frac{1}{n}\sum_{i=1}^{n} \left(\ell(W, Z_i) - (1+\gamma)L_S(W)\right)^2\right] \\
=& \mathbb{E}_{W,S}\left[\frac{1}{n}\sum_{i=1}^{n} \left(\ell^2(W, Z_i) - 2(1+\gamma)\ell(W, Z_i)L_S(W) + (1+\gamma)^2 L_S^2(W)\right)\right] \\
=& \mathbb{E}_{W,S}\left[\frac{1}{n}\sum_{i=1}^{n} \ell^2(W, Z_i)\right] - \mathbb{E}_{W,S}\left[(1-\gamma^2)L_S^2(W)\right] \\
=& L_n - (1-\gamma^2)\mathbb{E}_{W,S}\left[L_S^2(W)\right]
\end{aligned}$$

where the last equality is due to the fact that the loss is the zero-one loss (i.e. $\ell^2(\cdot,\cdot) = \ell(\cdot,\cdot)$). □

C.8. Proof of Lemma 4.4

Proof. By Lemma 4.3 and $\gamma \in (0, 1)$, we notice that,

$$\begin{aligned}
\mathrm{Err} - C_1 V(\gamma) =& L_\mu - L_n - C_1 L_n + C_1(1-\gamma^2)\mathbb{E}_{W,S}\left[L_S^2(W)\right] \\
\leq& L_\mu - (1+C_1)L_n + C_1(1-\gamma^2)\mathbb{E}_{W,S}\left[L_S(W)\right] && (19) \\
=& L_\mu - (1+C_1\gamma^2)L_n, && (20)
\end{aligned}$$

where the inequality is because that $L_S(W) \in [0, 1]$ (i.e. $L_S^2(W) \leq L_S(W)$).

Noting that in Eq (20), $\mathrm{Err} - C_1 V(\gamma)$ is upper bounded by a weighted generalization error. Thus, we can then directly apply Lemma 4.2 by choosing $C_1\gamma^2$ as the trade-off coefficient (i.e. replacing C_1 in Lemma 4.2 by $C_1\gamma^2$), which gives us

$$\mathrm{Err} - C_1 V(\gamma) \leq \frac{2 + C_1\gamma^2}{n} \sum_{i=1}^{n} \mathbb{E}_{L_i^+, \tilde{\varepsilon}_i}\left[\tilde{\varepsilon}_i L_i^+\right],$$

where $\tilde{\varepsilon}_i = \varepsilon_i - \frac{C_1\gamma^2}{C_1\gamma^2+2}$. □

C.9. Proof of Theorem 4.4

Proof. The RHS of Eq. (20) in the proof of Lemma 4.4 has already been bounded in Theorem 4.3 by regarding $C_1\gamma^2$ as the weighted parameter C_1 in Theorem 4.3. Then, there exist $C_1, C_2 > 0$ s.t.

$$\mathrm{Err} - C_1 V(\gamma) \leq L_\mu - (1+C_1\gamma^2)L_n \leq \sum_{i=1}^{n} \frac{I(L_i^+; U_i)}{nC_2}.$$

Furthermore, from the proof of Theorem 4.3, we note that the following is valid

$$\mathrm{Err} \leq \min_{C_1,C_2>0, e^{2C_2}+e^{-2C_2(C_1\gamma^2+1)}\leq 2} C_1 V(\gamma) + \sum_{i=1}^{n} \frac{I(L_i^+;U_i)}{nC_2}.$$

Notice that the original optimization space of the variance based bound should be larger than $\{C_1, C_2 | C_1, C_2 > 0, e^{2C_2} + e^{-2C_2(C_1\gamma^2+1)} \leq 2\}$ because in Eq. (19), we upper bound the most interested quantity $\mathrm{Err} - C_1 V(\gamma)$ by $L_\mu - (1+C_1\gamma^2)L_n$, which restricts the original optimization space.

This completes the proof. □

C.10. Proof of Lemma 4.5

Proof. By the definition of λ-sharpness, we notice that

$$\begin{aligned}
F_i(\lambda) =& \mathbb{E}_{W,Z_i}\left[\ell(W,Z_i) - (1+\lambda)\mathbb{E}_{W|Z_i}\ell(W,Z_i)\right]^2 \\
=& \mathbb{E}_{Z_i}\left[\mathbb{E}_{W|Z_i}\left[\ell(W,Z_i)^2\right] - 2(1+\lambda)\mathbb{E}^2_{W|Z_i}\ell(W,Z_i) + (1+\lambda)^2\mathbb{E}^2_{W|Z_i}\ell(W,Z_i)\right] \\
=& \mathbb{E}_{W,Z_i}\left[\ell(W,Z_i)\right] - (1-\lambda^2)\mathbb{E}_{Z_i}\left[\mathbb{E}^2_{W|Z_i}\ell(W,Z_i)\right],
\end{aligned}$$

where the last equality is due to the fact that the loss is the zero-one loss. □

C.11. Proof of Lemma 4.6

Proof. By Lemma 4.5, we have

$$\begin{aligned}
\mathrm{Err} - \frac{C_1}{n}\sum_{i=1}^n F_i(\lambda) =& L_\mu - (1+C_1)L_n + \frac{(1-\lambda^2)C_1}{n}\sum_{i=1}^n \mathbb{E}_{Z_i}\left[\mathbb{E}^2_{W|Z_i}\ell(W,Z_i)\right] \\
=& \frac{1}{n}\sum_{i=1}^n\left[\mathbb{E}_{U_i,L_i}\left[L_{i,\overline{U}_i} - (1+C_1)L_{i,U_i}\right] + (1-\lambda^2)C_1\mathbb{E}_{\widetilde{Z}_{i,U_i}}\left[\mathbb{E}^2_{L_{i,U_i}|\widetilde{Z}_{i,U_i}} L_{i,U_i}\right]\right].
\end{aligned}$$

Let $\Lambda(\widetilde{Z}_{i,U_i}) = \mathbb{E}^2_{L_{i,U_i}|\widetilde{Z}_{i,U_i}} L_{i,U_i}$ and $\Lambda(\widetilde{Z}_{i,\overline{U}_i}) = \mathbb{E}^2_{L_{i,\overline{U}_i}|\widetilde{Z}_{i,\overline{U}_i}} L_{i,\overline{U}_i}$. Let $\Lambda(\widetilde{Z}_i^+) = \mathbb{E}^2_{L_i^+|\widetilde{Z}_i^+,U_i} L_i^+$ and $\Lambda(\widetilde{Z}_i^-) = \mathbb{E}^2_{L_i^-|\widetilde{Z}_i^-,U_i} L_i^-$. A key observation is the following:

$$\begin{aligned}
& \mathbb{E}_{U_i,L_i}\left[L_{i,\overline{U}_i} - (1+C_1)L_{i,U_i}\right] + (1-\lambda^2)C_1\mathbb{E}_{\widetilde{Z}_{i,U_i}}\left[\Lambda(\widetilde{Z}_{i,U_i})\right] \\
=& (1+\frac{C_1}{2})\mathbb{E}_{U_i,L_i}\left[L_{i,\overline{U}_i} - L_{i,U_i}\right] - \frac{C_1}{2}(1-\lambda^2)\mathbb{E}_{\widetilde{Z}_i,U_i}\left[\Lambda(\widetilde{Z}_{i,\overline{U}_i}) - \Lambda(\widetilde{Z}_{i,U_i})\right] \\
& - \frac{C_1}{2}\left[\mathbb{E}_{U_i,L_i}\left[L_{i,\overline{U}_i}\right] - (1-\lambda^2)\mathbb{E}_{\widetilde{Z}_i,U_i}\left[\Lambda(\widetilde{Z}_{i,\overline{U}_i})\right]\right] \\
& - \frac{C_1}{2}\left[\mathbb{E}_{U_i,L_i}\left[L_{i,U_i}\right] - (1-\lambda^2)\mathbb{E}_{\widetilde{Z}_i,U_i}\left[\Lambda(\widetilde{Z}_{i,U_i})\right]\right] \\
=& \mathbb{E}_{L_i^-,U_i}\left[(-1)^{U_i}\frac{C_1+2}{2}L_i^- - (-1)^{U_i}\frac{C_1(1-\lambda^2)}{2}\mathbb{E}_{\widetilde{Z}_i^-|U_i}\left[\Lambda(\widetilde{Z}_i^-)\right]\right. \\
& \left. - \frac{C_1}{2}L_i^- + \frac{C_1(1-\lambda^2)}{2}\mathbb{E}_{\widetilde{Z}_i^-|U_i}\left[\Lambda(\widetilde{Z}_i^-)\right]\right] \\
& + \mathbb{E}_{L_i^+,U_i}\left[-(-1)^{U_i}\frac{C_1+2}{2}L_i^+ + (-1)^{U_i}\frac{C_1(1-\lambda^2)}{2}\mathbb{E}_{\widetilde{Z}_i^+|U_i}\left[\Lambda(\widetilde{Z}_i^+)\right]\right. \\
& \left. - \frac{C_1}{2}L_i^+ + \frac{C_1(1-\lambda^2)}{2}\mathbb{E}_{\widetilde{Z}_i^+|U_i}\left[\Lambda(\widetilde{Z}_i^+)\right]\right] \\
=& (C_1+2)\mathbb{E}_{L_i^+,U_i}\left[(\varepsilon_i - \frac{C_1}{C_1+2})L_i^+ - \frac{C_1(1-\lambda^2)}{C_1+2}(\varepsilon_i - 1)\mathbb{E}_{\widetilde{Z}_i^+|U_i}\left[\Lambda(\widetilde{Z}_i^+)\right]\right],
\end{aligned}$$

where ε_i is the Rademacher variable.

Thus,

$$\mathrm{Err}-\frac{C_1}{n}\sum_{i=1}^{n}F_i(\lambda)=\frac{C_1+2}{n}\sum_{i=1}^{n}\mathbb{E}_{L_i^+,U_i}\left[(\varepsilon_i-\frac{C_1}{C_1+2})L_{i,1}-\frac{C_1(1-\lambda^2)}{C_1+2}(\varepsilon_i-1)\mathbb{E}_{\widetilde{Z}_i^+|U_i}\left[\mathbb{E}^2_{L_i^+|\widetilde{Z}_i^+,U_i}L_i^+\right]\right]. \tag{21}$$

This completes the proof. □

C.12. Proof of Theorem 4.5

Proof. Recall Eq. (21),

$$\mathrm{Err}-\frac{C_1}{n}\sum_{i=1}^{n}F_i(\lambda)=\frac{C_1+2}{n}\sum_{i=1}^{n}\mathbb{E}_{L_i^+,U_i}\left[(\varepsilon_i-\frac{C_1}{C_1+2})L_i^+-\frac{C_1(1-\lambda^2)}{C_1+2}(\varepsilon_i-1)\mathbb{E}_{\widetilde{Z}_i^+|U_i}\left[\mathbb{E}^2_{L_i^+|\widetilde{Z}_i^+,U_i}L_i^+\right]\right].$$

Notice that we cannot directly apply Lemma A.1 starting from here since there is a quadratic term, namely, $\mathbb{E}_{\widetilde{Z}_i^+|U_i}\left[\mathbb{E}^2_{L_i^+|\widetilde{Z}_i^+,U_i}L_i^+\right]$ in the RHS.

Inspired by Yang et al. (2019), we now assume that there exists a random variable R_i s.t.

$$\begin{aligned}
&(C_1+2)\mathbb{E}_{L_i^+,U_i}\left[(\varepsilon_i-\frac{C_1}{C_1+2})L_i^+-\frac{C_1(1-\lambda^2)}{C_1+2}(\varepsilon_i-1)\mathbb{E}_{\widetilde{Z}_i^+|U_i}\left[\mathbb{E}^2_{L_i^+|\widetilde{Z}_i^+,U_i}L_i^+\right]\right]\\
\leq&(C_1+2)\mathbb{E}_{L_i^+,U_i}\left[(\varepsilon_i-\frac{C_1}{C_1+2})L_i^+-\frac{C_1(1-\lambda^2)}{C_1+2}(\varepsilon_i-1)\mathbb{E}_{\widetilde{Z}_i^+|U_i}\left[R_i\mathbb{E}_{L_i^+|\widetilde{Z}_i^+,U_i}L_i^+\right]\right]\\
=&(C_1+2)\mathbb{E}_{L_i^+,U_i}\left[(\varepsilon_i-\frac{C_1}{C_1+2})L_i^+-\frac{C_1(1-\lambda^2)}{C_1+2}(\varepsilon_i-1)\mathbb{E}_{L_i^+|U_i}\left[R_iL_i^+\right]\right]\\
=&(C_1+2)\mathbb{E}_{L_i^+,U_i}\left[\left((\varepsilon_i-\frac{C_1}{C_1+2})-\frac{C_1(1-\lambda^2)}{C_1+2}(\varepsilon_i-1)R_i\right)L_i^+\right].
\end{aligned}\tag{22}$$

Such R_i could satisfy $R_i\geq\sup_{\tilde{z}_i^+}\mathbb{E}_{L_i^+|\widetilde{Z}_i^+=\tilde{z}_i^+,U_i=u_i}L_i^+$, for any fixed u_i, and the randomness of R_i is controlled by U_i, i.e. R_i is a function of U_i. A simple choice is to let $R_i=1$ (so R_i always exists), and another choice could be letting $R_i=\mathbb{E}_{L_i^{+\prime}|U_i\sim Q_i}\left[L_i^{+\prime}\right]$ that satisfies the condition, where Q_i is some distribution of L_i^+, and $L_i^{+\prime}$ is independent of L_i^+ and $\widetilde{Z}_i^+$ given U_i.

Recall that the shifted Rademacher varaible $\tilde{\varepsilon}_i=\varepsilon_i-\frac{C_1}{C_1+2}$, and let another shifted Rademacher variable $\hat{\varepsilon}_i=\varepsilon_i-1$. Then we are ready to invoke Lemma A.1,

$$I(L_i^+;U_i)\geq\sup_{t>0}t\mathbb{E}_{L_i^+,U_i}\left[\left((C_1+2)\tilde{\varepsilon}_i-C_1(1-\lambda^2)\hat{\varepsilon}_iR_i\right)L_i^+\right]-\ln\mathbb{E}_{L_i^+,U_i'}\left[e^{t\left((C_1+2)\tilde{\varepsilon}_i'-C_1(1-\lambda^2)\hat{\varepsilon}_i'R_i'\right)L_i^+}\right]. \tag{23}$$

Similar to the proof of Theorem 4.3, we hope the following hold

$$\mathbb{E}_{L_i^+}\mathbb{E}_{U_i'}\left[e^{t\left((C_1+2)\tilde{\varepsilon}_i'-C_1(1-\lambda^2)\hat{\varepsilon}_i'R_i'\right)L_i^+}\right]\leq 1. \tag{24}$$

By the independence, we have

$$\mathbb{E}_{L_i^+}\mathbb{E}_{U_i'}\left[e^{t\left((C_1+2)\tilde{\varepsilon}_i'-C_1(1-\lambda^2)\hat{\varepsilon}_i'R_i'\right)L_i^+}\right]=\mathbb{E}_{L_i^+}\left[\frac{e^{2t\left(C_1(1-\lambda^2)\tilde{r}_i'-C_1-1\right)L_i^+}+e^{2tL_i^+}}{2}\right],$$

where $\tilde{r}_i'\in[0,1]$ is the value (or the realization) of R_i' when $U_i'=0$ (or $\varepsilon_i'=-(-1)^{U_i'}=-1$), e.g., $\tilde{r}_i'=\mathbb{E}_{L_i^{+\prime}|U_i'=0}\left[L_i^{+\prime}\right]$.

Then, since L_i^+ could be either 0 or 1. We now consider the two cases.

(i) When $L_i^+ = 0$, then $\frac{e^{2t\left(C_1(1-\lambda^2)\tilde{r}_i'-C_1-1\right)L_i^+}+e^{2tL_i^+}}{2} = 1$. Therefore, when $L_i^+ = 0$, the value of R_i' has no effect on the moment generating function.

(ii) When $L_i^+ = 1$, we have the formula $\frac{e^{2t\left(C_1(1-\lambda^2)\tilde{r}_i'-C_1-1\right)}+e^{2t}}{2}$. Notably, only when $\varepsilon_i' = -1$ (or $U_i' = 0$) and $L_i^+ = 1$, the value of R_i', viz, $\tilde{r}_i'$, has some impact on the moment generating function. Since $\tilde{r}_i' \in [0,1]$, it's sufficient to upper bound R_i' by the random variable $\frac{1-\varepsilon_i'}{2} = \frac{-\hat{\varepsilon}_i'}{2}$. Thus,

$$\mathbb{E}_{L_i^+}\mathbb{E}_{U_i'}\left[e^{t\left((C_1+2)\tilde{\varepsilon}_i'-C_1(1-\lambda^2)\hat{\varepsilon}_i'R_i'\right)L_i^+}\right] \leq \mathbb{E}_{L_i^+}\mathbb{E}_{U_i'}\left[e^{t\left((C_1+2)\tilde{\varepsilon}_i'+\frac{C_1}{2}(1-\lambda^2)\hat{\varepsilon}_i'^2\right)L_i^+}\right]. \tag{25}$$

By the moment generating function of the Bernoulli random variable L_i^+, we have

$$\begin{aligned}\mathbb{E}_{U_i'}\mathbb{E}_{L_i^+}\left[e^{t\left((C_1+2)\tilde{\varepsilon}_i'+\frac{C_1}{2}(1-\lambda^2)\hat{\varepsilon}_i'^2\right)L_i^+}\right] =&\mathbb{E}_{U_i}\left[1-\mathbb{E}_{L_i^+}\left[L_i^+\right]+\mathbb{E}_{L_i^+}\left[L_i^+\right]e^{t\left((C_1+2)\tilde{\varepsilon}_i'+\frac{C_1}{2}(1-\lambda^2)\hat{\varepsilon}_i'^2\right)}\right]\\ =&1-\mathbb{E}_{L_i^+}\left[L_i^+\right]+\mathbb{E}_{L_i^+}\left[L_i^+\right]\frac{e^{-2t\left(C_1\lambda^2+1\right)}+e^{2t}}{2}.\end{aligned}$$

Since $0 \leq \mathbb{E}_{L_i^+}\left[L_i^+\right] \leq 1$, we only need to require that

$$\frac{e^{-2t\left(C_1\lambda^2+1\right)}+e^{2t}}{2} \leq 1.$$

Replacing t by C_2 and putting everything together (Eq. (21-24)), we have

$$\mathrm{Err} \leq \frac{C_1}{n}\sum_{i=1}^n F_i(\lambda) + \sum_{i=1}^n \frac{I(L_i^+;U_i)}{nC_2}.$$

This completes the proof. □

C.13. Proof of Corollary 4.1

Proof. According to the proofs in Section C.9 and Section C.12, we know that the sufficient conditions to let Eq. (5) and Eq. (6) hold are $e^{2C_2}+e^{-2C_2(C_1\gamma^2+1)} \leq 2$ and $e^{2C_2}+e^{-2C_2(C_1\lambda^2+1)} \leq 2$, respectively. Given that both $\gamma, \lambda \in (0,1)$, there must exist C_1, C_2 to let both Eq. (5) and Eq. (6) hold. Then taking minimum of these two inequalities will give us the desired result. □

Additionally, in any of the following case: (i) $L_n \to 0$; (ii)$V(\gamma) \to 0$ for some $\gamma \in (0,1)$; (iii)$F(\lambda) \to 0$ for some $\lambda \in (0,1)$, we can let $C_1 \to \infty$ and let $C_2 = \frac{\ln 2}{2}$, then $\mathrm{Err} \leq \sum_{i=1}^n \frac{2I(L_i^+;U_i)}{n\ln 2}$. This justifies the remark after Corollary 4.1.

D. Some Background on Channel Capacity of Binary Channel

In this section, we follow the custom of the notations in Cover & Thomas (2006), where the logarithm usually has a base of 2 (i.e. $\log_2$). In addition, for a binary random variable, the entropy function $H(\cdot)$ can be a binary entropy function, for example, the random variable X has the value 0 and 1, and $P(X=1)=p$, $P(X=0)=1-p$, then $H(X)=H(p)=-p\log_2 p-(1-p)\log_2(1-p)$. The channel capacity of a channel between input X and output Y is defined as $C \triangleq \max_{P_X} I(X;Y)$.

D.1. Channel Capacity of Binary Symmetric Channel (BSC)

In a general case, the channel capacity of Figure 1(left) can be computed as in the following lemma.

Lemma D.1. *When $X \sim P_{U_i}$, the channel capacity of the channel in Figure 1(left) is achieved and $C = (1-\alpha)\left(1-H\left(\frac{1-\alpha-\epsilon}{1-\alpha}, \frac{\epsilon}{1-\alpha}\right)\right)$.*

We note that this lemma is an exercise problem in Cover & Thomas (2006, Problem 7.13).

Proof. Let $P(X=0)=\pi$ and $P(X=0)=1-\pi$, then $I(X;Y)=H(Y)-H(Y|X)=H(Y)-H(1-\epsilon-\alpha,\alpha,\epsilon)$. It's easy to see that $H(Y)=H(\pi(1-\alpha-\epsilon)+\epsilon(1-\pi),\alpha,\pi\epsilon+(1-\pi)(1-\epsilon-\alpha))$. We let $A=\pi(1-\alpha-\epsilon)+\epsilon(1-\pi)$ and $B=\pi\epsilon+(1-\pi)(1-\epsilon-\alpha))$. Notice that $A+B=1-\alpha$, then

$$\begin{aligned}
H(Y)=&H(\pi(1-\alpha-\epsilon)+\epsilon(1-\pi),\alpha,\pi\epsilon+(1-\pi)(1-\epsilon-\alpha))\\
=&-\left[(A+B)\log_2(1-\alpha)+\alpha\log_2\alpha+A\log_2\frac{A}{1-\alpha}+B\log_2\frac{B}{1-\alpha}\right]\\
=&H(\alpha)-(1-\alpha)\left[\frac{A}{1-\alpha}\log_2\frac{A}{1-\alpha}+\frac{B}{1-\alpha}\log_2\frac{B}{1-\alpha}\right]\\
=&H(\alpha)+(1-\alpha)H(\frac{A}{1-\alpha},\frac{B}{1-\alpha})\leq H(\alpha)+1-\alpha.
\end{aligned}$$

To achieve the channel capacity (or to let the equality above hold), we need to let $A=B$, which indicates that $\pi=\frac{1}{2}$.

Thus,

$$\begin{aligned}
C=H(\alpha)+1-\alpha-H(1-\epsilon-\alpha,\alpha,\epsilon)=&H(\alpha)+1-\alpha-\left(H(\alpha)+(1-\alpha)H\left(\frac{1-\epsilon-\alpha}{1-\alpha},\frac{\epsilon}{1-\alpha}\right)\right)\\
=&(1-\alpha)\left(1-H\left(\frac{1-\alpha-\epsilon}{1-\alpha},\frac{\epsilon}{1-\alpha}\right)\right),
\end{aligned}$$

which completes the proof. ☐

D.2. Channel Capacity of Binary Asymmetric Channel (BAC)

The channel capacity of the BAC channel in Figure 1(right) is given below.

Lemma D.2. *The channel capacity of the BAC in Figure 1(right) is* $C=\log_2(1+\beta)-\frac{1-q}{1-p-q}H(p)+\frac{p}{1-p-q}H(q)$, *where* $\beta=2^{\frac{H(p)-H(q)}{1-p-q}}$ *and the capacity is achived when* $P(U_i=1)=\frac{1-q(1+\beta)}{(1-p-q)(1+\beta)}$. *Further, if* $p=0$ *(i.e. Z-channel),* $C=\log_2\left(1+2^{\frac{-H(q)}{1-q}}\right)$, *and for small* q, *the capacity can be approximated by* $C\approx 1-\frac{1}{2}H(q)$.

Remark D.1. *Notice that in this case, let* $X\sim\mathrm{Bern}(1/2)$ *(i.e.* $X=U_i$*) will not achieve the channel capacity. Thus, in the interpolating setting, except for Theorem 4.2, we have another upper bound for* $I(L_i^+;U_i)$, *namely* $\frac{1}{n}\sum_{i=1}^n I(L_i^+;U_i)\leq\frac{1}{n}\sum_{i=1}\ln\left(1+2^{\frac{-H(1-q_i)}{1-q_i}}\right)$. *If we further let* q_i *be the same for each* i *(which indeed should be true), then* $I(L_i^+;U_i)\leq\ln\left(1+2^{\frac{-H(L_\mu)}{L_\mu}}\right)$.

Proof. Let $P(X=1)=\pi$, then $I(X;Y)=H(\pi(1-p-q)+q)-\pi\left(H(p)-H(q)\right)-H(q)$. Let $\frac{dI(X;Y)}{d\pi}=(1-p-q)\log_2\left(\frac{1}{\pi(1-p-q)+q}-1\right)-H(p)+H(q)=0$, we have the optimal $\pi^*=\frac{1-q(1+\beta)}{(1-p-q)(1+\beta)}$ where $\beta=2^{\frac{H(p)-H(q)}{1-p-q}}$. Plugging $\pi=\pi^*$ into the formula of $I(X;Y)$, we have $I(X;Y)=\log_2(1+\beta)-\frac{1-q}{1-p-q}H(p)+\frac{p}{1-p-q}H(q)$. ☐

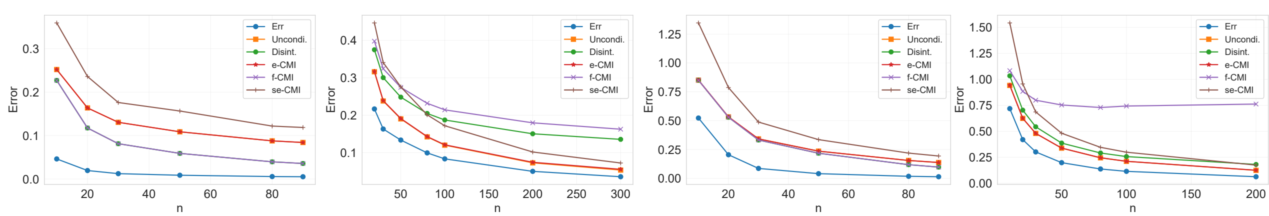
E. Experimental Details and Additional Results

E.1. Experiment Setup

In our linear classifier experiment, we generate synthetic Gaussian data using the widely-used Python package *scikit-learn* (Pedregosa et al., 2011). We draw each dimension (or feature) of X independently from some Gaussian distribution, and let all the features be informative to its class labels Y. Specifically, we choose the dimension of data X to be 5 and we create different class of points normally distributed (with the standard deviation being 1) about vertices of an 5-dimensional hypercube, where its sides of length can be manually controlled. In addition, we utilize full-batch gradient descent with a fixed learning rate of 0.01 to train the linear classifier. We perform training for a total of 500 epochs, and we employ early stopping when the training error reaches a sufficiently low threshold (e.g., $< 0.5\%$). To ensure robustness and statistical significance, we draw 50 different supersamples for each experiment. Within each supersample, we further generate 100 different mask random variables, resulting in a total of $5,000$ runs for each experimental setting. This comprehensive setup enables us to compare both the unconditional MI bounds and the disintegrated MI bounds. Additionally, if the unconditional MI bound is the sole evaluated objective, one has the option to completely restart the training process $5,000$ times.

In the neural networks experiments, we follow the same setup with (Harutyunyan et al., 2021; Hellström & Durisi, 2022a). Specifically, we draw k_1 samples of $\widetilde{Z}$ and k_2 samples of U for each given $\tilde{z}$. For the CNN on the binary MNIST dataset, we set $k_1 = 5$ and $k_2 = 30$. The 4-layer CNN model is trained using the Adam optimizer with a learning rate of 0.001 and a momentum coefficient of $\beta_1 = 0.9$. The training process spans 200 epochs, with a batch size of 128. For ResNet-50 on CIFAR10, we set $k_1 = 2$ and $k_2 = 40$. The ResNet model is trained using stochastic gradient descent (SGD) with a learning rate of 0.01 and a momentum coefficient of 0.9 for a total of 40 epochs. The batch size for this experiment is set to 64. In the SGLD experiment, we once again train a 4-layer CNN on the binary MNIST dataset. The batch size is set to 100, and the training lasts for 40 epochs. The initial learning rate is 0.01 and decays by a factor of 0.9 after every 100 iterations. Let t be the iteration index, the inverse temperature of SGLD is given by $\min\{4000, \max\{100, 10e^{t/100}\}\}$. We set the training sample size to $n = 4000$, and $k_1 = 5$ and $k_2 = 30$. We save checkpoints every 4 epochs. All these experiments are conducted using NVIDIA Tesla V100 GPUs with 32 GB of memory. For more comprehensive details, including model architectures, we recommend referring to (Harutyunyan et al., 2021; Hellström & Durisi, 2022a).

Estimating the γ-variance and λ-sharpness in the CMI setting is a straightforward process. For example, to estimate sharpness, for each fixed $\tilde{z}$, we store the training losses L_i^+ when $U_i = 0$ (corresponding to $\tilde{z}^+$) and the training losses L_i^- when $U_i = 1$ (corresponding to $\tilde{z}^-$) with different weight configurations W. By doing so, we collect the necessary data to compute the second term of the equation in Lemma 4.5.



(a) $|\mathcal{Y}| = 2$ (Realizable) (b) $|\mathcal{Y}| = 2$ (Non-Separable) (c) $|\mathcal{Y}| = 10$ (Realizable) (d) $|\mathcal{Y}| = 10$ (Non-Separable)

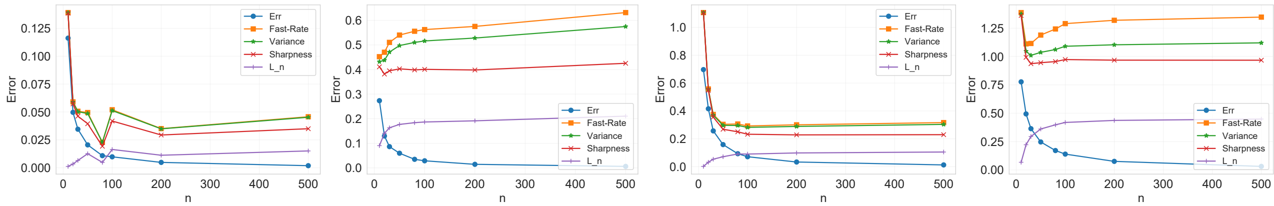
Figure 4. Comparison of the square-root bounds on the synthetic dataset. Here *f-CMI*, *e-CMI* and *se-CMI* refer to the disintegrated f-CMI bound (Harutyunyan et al., 2021), the unconditional e-CMI bound and the single-loss square-root bound in Theorem 4.1, respectively.

E.2. Additional Numerical Results: Comparison of Square-Root Bounds

We conduct a comparison of square-root bounds on the synthetic dataset, where we also include the disintegrated version of the f-CMI bound proposed by Harutyunyan et al. (2021), an improved unconditional e-CMI bound (obtained by replacing $I(L_i; U_i|\widetilde{Z})$ with $I(L_i; U_i)$), and the single-loss square-root bound presented in Theorem 4.1. The results are illustrated in Figure 4. Consistent with the observations in the main text, we find that the disintegrated bounds are tighter than the unconditional MI bounds when the training loss approaches zero, but looser than the unconditional MI bounds when the training loss is large. This suggests that while, according to the DPI, the unconditional e-CMI bound or ld-MI bound should be tighter than the f-CMI bound, in some cases, the disintegrated version of the f-CMI bound may be tighter than the unconditional e-CMI bound or ld-MI bound. For non-separable μ, the f-CMI bound becomes looser as the number of classes increases, which provides justification for the remarks made after Theorem 3.1. In fact, it can be even worse than the single-loss square-root bound in Theorem 4.1, which includes an undesired constant of 2.

E.3. Additional Numerical Results: Comparison of Fast-Rate Bounds

We conduct a comparison of fast-rate bounds, including Eq. (1) in Theorem 4.3, the variance bound in Theorem 4.4, and the sharpness bound in Theorem 4.5, on the synthetic dataset with fixed values of C_1 and C_2. As mentioned in the main text, if $L_n \to 0$, both $V(\gamma)$ and $F(\lambda)$ become zero, resulting in the three bounds being equivalent. However, when $L_n \neq 0$, the variance bound and sharpness bound are always sharper than Eq. (1), as discussed earlier. In Figure 5, we compare these



(a) $|\mathcal{Y}| = 2$ (Small L_n) (b) $|\mathcal{Y}| = 2$ (Large L_n) (c) $|\mathcal{Y}| = 10$ (Small L_n) (d) $|\mathcal{Y}| = 10$ (Large L_n)

Figure 5. Comparison of three fast-rate bounds on the synthetic dataset. Here *Fast-Rate* refers to the fast-rate bound of Eq. (1) in Theorem 4.3.

bounds with fixed values of $C_1 = 3$ and $C_2 = 0.3$. Figures 5a and 5c demonstrate that when L_n is small, the gap between the variance bound and Eq. (1) is small, indicating that the loss variance in this case is also small. However, the sharpness bound clearly outperforms the other two bounds. Furthermore, in Figures 5b and 5d, when L_n is large, both the sharpness bound and the variance bound significantly improve upon Eq. (1). Notably, only the sharpness bound remains non-vacuous in Figure 5d.