Towards Generalization beyond Pointwise Learning: A Unified Information-theoretic Perspective

Yuxin Dong [1] **Tieliang Gong** [1] **Hong Chen** [2] **Zhongjiang He** [3] **Mengxiang Li** [3] **Shuangyong Song** [3] **Chen Li** [1]

Abstract

The recent surge in contrastive learning has intensified the interest in understanding the generalization of non-pointwise learning paradigms. While information-theoretic analysis achieves remarkable success in characterizing the generalization behavior of learning algorithms, its applicability is largely confined to pointwise learning, with extensions to the simplest pairwise settings remaining unexplored due to the challenges of non-i.i.d losses and dimensionality explosion. In this paper, we develop the first series of information-theoretic bounds extending beyond pointwise scenarios, encompassing pointwise, pairwise, triplet, quadruplet, and higher-order scenarios, all within a unified framework. Specifically, our hypothesis-based bounds elucidate the generalization behavior of iterative and noisy learning algorithms via gradient covariance analysis, and our prediction-based bounds accurately estimate the generalization gap with computationally tractable low-dimensional information metrics. Comprehensive numerical studies then demonstrate the effectiveness of our bounds in capturing the generalization dynamics across diverse learning scenarios.

1. Introduction

It is frequently seen in modern deep-learning scenarios that the loss functions encompass more than one data point. Such paradigms are evident in areas including contrastive representation learning (Chen et al., 2020; Khosla et al., 2020; Radford et al., 2021), deep metric learning (Oh Song et al., 2016; Sohn, 2016; Ge, 2018), AUC maximization (Ying et al., 2016; Liu et al., 2018), and ranking algorithms (Clémençon et al., 2008; Agarwal & Niyogi, 2009). Besides the extensive applications, theoretical foundations of these methodologies have been explored through the lens of uniform convergence (Wang et al., 2012; Kar et al., 2013; Cao et al., 2016) and algorithm stability (Lei et al., 2020; 2021; Yang et al., 2021b). However, existing generalization studies primarily focus on pairwise (Li & Liu, 2023; Wang et al., 2023; Huang et al., 2023) and triplet (Chen et al., 2023) learning scenarios, with limited exploration in quadruplet (Chen et al., 2017) or higher-order (Sohn, 2016; Chen et al., 2020) contexts. Additionally, these analytical methods often hinge on the complexity of hypothesis spaces or stringent assumptions like Lipschitz continuity, smoothness, and convexity, resulting in vacuous or computationally intractable bounds when applied to deep neural networks.

Recent trends in information theory (Xu & Raginsky, 2017) offer a promising alternative for analyzing the generalization properties of noisy and iterative learning algorithms (Negrea et al., 2019; Haghifam et al., 2020; Neu et al., 2021; Tang & Liu, 2023). These bounds are advantageous in being simultaneously contingent on data distributions and learning algorithms, and operate under considerably more relaxed assumptions than algorithm stability approaches. Further advancements in this domain provide refined bounds by utilizing information quantities involving network predictions (Harutyunyan et al., 2021), loss pairs (Hellström & Durisi, 2022b), or loss differences (Wang & Mao, 2023) under the supersample framework (Steinke & Zakynthinou, 2020). These bounds not only exhibit computational tractability due to the lower dimensionality of the associated random variables but also provide quantitatively tighter estimates, particularly for large neural networks. However, to our best knowledge, these analyses are primarily confined to pointwise learning scenarios, with extensions to even the simplest pairwise settings remaining unexplored.

The significant challenge in extending these information-theoretic generalization bounds is that the empirical risk no longer represents a sum of i.i.d random variables, due to the existence of overlapping training samples in individual losses. This property is crucial for deriving bounds through input-output mutual information (Xu & Raginsky, 2017; Bu et al., 2020), a cornerstone in the analysis of noisy and

[1]School of Computer Science and Technology, Xi'an Jiaotong University [2]College of Science, Huazhong Agriculture University [3]China Telecom Corporation Limited. Correspondence to: Tieliang Gong <adidasgtl@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

iterative learning algorithms (Negrea et al., 2019; Wang & Mao, 2021). The generalization of prediction-based bounds within the supersample framework is also highly non-trivial, as straightforward adaptations involve information quantities whose dimensionality grows exponentially with the number of samples m taken by the loss function, resulting in the loss of computational feasibility. In this paper, we overcome these obstacles and provide information-theoretic generalization analysis for a variety of learning scenarios. The main contributions are summarized as follows:

- We introduce the first information-theoretic generalization bound that extends beyond pointwise learning scenarios, accommodating a variety of bounded multivariate loss functions taking arbitrary numbers of data points. Our analysis encompasses prevailing learning paradigms including pointwise, pairwise, triplet, quadruplet, and higher-order cases, all within a unified framework.
- We generalize current information-theoretic generalization bounds built upon input-output mutual information and conditional mutual information measures under the supersample setting, by employing a bottom-to-top reduction of these hypothesis-based information metrics to overcome the non-i.i.d challenge. These bounds shed light on the understanding of iterative and noisy learning algorithms within non-pointwise learning contexts.
- We advance prediction-based bounds while circumventing the issue of dimensionality explosion. By exploring a novel decomposition of the supersample variables into independent lower-dimensional ones, we establish enhanced bounds using only 2-dimensional information quantities, regardless of the value m. These bounds are not only strictly tighter than the straightforward adaptations but also ensure direct computational tractability.
- Extensive experimental results demonstrate the effectiveness of our bounds in capturing the generalization dynamics across various deep-learning configurations, utilizing both synthetic and large-scale real-world datasets.

2. Preliminaries

We denote random variables by capitalized letters (X), their specific realizations by lowercase letters (x), and the corresponding spaces by calligraphic letters ($\mathcal{X}$). Let P_X denote the distribution of variable X, $P_{X|Y}$ be the conditional distribution of X given Y, and $P_{X|Y=y}$ (or $P_{X|y}$) be the one conditioning on a specific realization. Similarly, denote $\mathbb{E}_X$, Var_X, and Cov_X as the expectation, variance, and covariance matrix taken over $X \sim P_X$. Let $H(X)$ be Shannon's (differential) entropy and $D(P \,\|\, Q)$ be the KL divergence of P w.r.t Q. We further refer to $d(p \,\|\, q) = p \log(\frac{p}{q}) + (1-p)\log(\frac{1-p}{1-q})$ as the binary KL divergence. Let $I(X;Y)$ be the mutual information (MI) between variables X and Y, and $I(X;Y|Z)$ be their conditional mutual information (CMI) given Z. We further denote $I^z(X;Y) = D(P_{X,Y|z} \,\|\, P_{X|z}P_{Y|z})$ as the disintegrated MI. Let $\mathbb{W}(\cdot,\cdot)$ be the Wasserstein distance, and $\log$ be the logarithmic function with base e.

2.1. Generalization Error

Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ denote the instance space of interest, where $\mathcal{X}$ and $\mathcal{Y}$ represent the input and label spaces, respectively. The training dataset $Z = \{Z_i\}_{i=1}^n \in \mathcal{Z}^n$ is constructed by i.i.d sampling from the unknown data-generating distribution μ. The learning algorithm $\mathcal{A}$ then takes Z as the input and produces a hypothesis $W \in \mathcal{W}$, characterized by the conditional distribution $P_{W|Z}$. Our analysis adopts a unified framework for generalizing across diverse learning paradigms, which are categorized by the number of samples m taken by the respective loss functions:

- Pointwise Learning ($m = 1$): cross-entropy, MSE;
- Pairwise Learning ($m = 2$): contrastive loss;
- Triplet Learning ($m = 3$): triplet loss;
- Quadruplet Learning ($m = 4$): quadruplet loss;
- Higher-order Cases ($m \geq 5$): N-pair loss, NT-Xent loss.

Let $\ell : \mathcal{W} \times \mathcal{Z}^m \mapsto \mathbb{R}^+$ be the loss function used to assess the quality of the output hypothesis W. Then for a given $w \in \mathcal{W}$, the population risk can be defined as $L(w) \triangleq \mathbb{E}_{Z_{1:m}}[\ell(w, Z_{1:m})]$, where $Z_{1:m} \sim \mu^m$ is a set of test instances. The expected population risk is denoted as $L = \mathbb{E}_W[L(W)]$. Let P_n^m be the set of m-permutations of n, and let C_n^m be the set of m-combinations. Given a sequence of indices $u = \{u_i\}_{i=1}^m \in [1,n]^m$, let $Z_u = \{Z_{u_i}\}_{i=1}^m$ be the sequence of training samples indexed by u. The empirical risk can then be defined as $L_Z(w) \triangleq \frac{1}{|\mathsf{P}_n^m|}\sum_{u \in \mathsf{P}_n^m} \ell(w, Z_u)$. Similarly, let $L_n = \mathbb{E}_{W,Z}[L_Z(W)]$ be the expected empirical risk. The generalization error $\overline{\mathrm{gen}} \triangleq L - L_n$ quantifies the discrepancy between empirical and population risks. Some prevalent settings of non-pointwise learning are discussed in Appendix D.1.

2.2. Supersample Setting

The CMI framework is initially explored in (Steinke & Zakynthinou, 2020) for generalization analysis. Let $\widetilde{Z} = \{\widetilde{Z}_i\}_{i=1}^n \in \mathcal{Z}^{n\times 2}$ be the supersample dataset i.i.d sampled from μ, where each element $\widetilde{Z}_i = (\widetilde{Z}_i^0, \widetilde{Z}_i^1)$ comprises a pair of samples. A set of binary random variables $S = \{S_i\}_{i=1}^n \sim \mathrm{Unif}(\{0,1\}^n)$ is used to segregate training and test samples, such that $\widetilde{Z}_S = \{\widetilde{Z}_i^{S_i}\}_{i=1}^n$ and $\widetilde{Z}_{\overline{S}} = \{\widetilde{Z}_i^{\overline{S}_i}\}_{i=1}^n$ form the training and test datasets, respectively. The empirical and population risks are then formulated as $L_n = \mathbb{E}_{W,\widetilde{Z},S}[L_{\widetilde{Z}_S}(W)]$ and $L = \mathbb{E}_{W,\widetilde{Z},S}[L_{\widetilde{Z}_{\overline{S}}}(W)]$.

Let $\mathsf{B}_m = \{0,1\}^m$ be the set of binary sequences of length m. Given $u \in \mathsf{P}_n^m$ and $b \in \mathsf{B}_m$, we denote $\widetilde{Z}_u^b$ as the sub-

set of samples $\{\widetilde{Z}_{u_i}^{b_i}\}_{i=1}^m$, and L_u^b be the loss evaluated by $\ell(W, \widetilde{Z}_u^b)$. The set of losses computed across all combinations of $b \in \mathsf{B}_m$ is represented as $L_u = \{L_u^b\}_{b\in\mathsf{B}_m}$. Let $S_u = \{S_{u_i}\}_{i=1}^m \in \mathsf{B}_m$ be the sequence of supersample variables indexed by u, and let $\Phi_u = \{S_{u_1} \oplus S_{u_i}\}_{i=2}^m \in \mathsf{B}_{m-1}$, where $\oplus$ is the XOR operation. Given binary value $b \in \{0, 1\}$, define $b \otimes \Phi_u = (b, \{\Phi_{u_i} \oplus b\}_{i=1}^{m-1}) \in \mathsf{B}_m$. To simplify the notations, we denote $0 \otimes \Phi_u$ and $1 \otimes \Phi_u$ as Φ_u^- and Φ_u^+, respectively. $L_u^{\Phi_u} = (L_u^{\Phi_u^-}, L_u^{\Phi_u^+})$ then represents a pair of losses, and $\Delta_u^{\Phi_u} = L_u^{\Phi_u^+} - L_u^{\Phi_u^-}$ is their difference.

3. Hypothesis-based Generalization Bounds

3.1. Generalization Bounds with Input-output MI

The foundational work of (Xu & Raginsky, 2017) introduced a methodology for bounding the expected generalization error through the MI between the input dataset Z and the output hypothesis W, termed the input-output MI. This concept was further refined and expanded upon in subsequent studies (Bu et al., 2020; Harutyunyan et al., 2021) through the incorporation of random subsets:

Theorem 3.1. *(Theorem 2.2 in (Harutyunyan et al., 2021)) Assume that $m = 1$ and the loss function $\ell(\cdot, \cdot)$ is bounded within $[0, 1]$, it holds that for any $k \in [1, n]$,*

$$|\overline{\text{gen}}| \leq \frac{1}{|\mathsf{C}_n^k|} \sum_{u\in\mathsf{C}_n^k} \sqrt{\frac{1}{2k} I(W; Z_u)}.$$

Selecting $k = n$ simplifies the bound by noticing that $|\mathsf{C}_n^k| = 1$, leading to a convergence rate of $O(\sqrt{1/n})$. This result is contingent on the assumption of i.i.d loss terms for a fixed hypothesis $w \in \mathcal{W}$. Under this condition, the empirical risk $L_Z(w)$, representing the mean of n i.i.d $\frac{1}{2}$-subgaussian variables, can be proved to be $\frac{1}{2\sqrt{n}}$-subgaussian. However, this independence assumption is only valid in pointwise learning scenarios ($m = 1$). For multivariate loss functions ($m > 1$), the individual loss terms are no longer independent, owing to the overlapping subsets of the training samples. Inspired by (Bu et al., 2020) who introduced the concept of point-wise stabilities $I(W; Z_i)$ to measure generalization, we extend this notion to group-wise stability in multivariate learning contexts:

$$|\mathbb{E}_{W,Z_u}[\ell(W, Z_u)] - L| \leq \sqrt{\frac{1}{2} I(W; Z_u)},$$

where $u \in \mathsf{P}_n^m$ denotes a subset of the training dataset Z comprising m samples. Leveraging the superadditivity of MI for independent random variables, the group-wise stabilities serve as a surrogate for the on-average stability measured by the input-output MI $I(W; Z)$. We then propose a generalized theorem applicable for arbitrary $m \geq 1$:

Theorem 3.2. *Assume that the loss function $\ell(\cdot, \cdot)$ is bounded within $[0, 1]$, then for any $k \in \left[1, \frac{n}{m}\right]$,*

$$|\overline{\text{gen}}| \leq \frac{1}{|\mathsf{C}_n^{km}|} \sum_{u\in\mathsf{C}_n^{km}} \sqrt{\frac{1}{2k} I(W; Z_u)}.$$

This result coincides with the premise established by (Xu & Raginsky, 2017): the less dependent the output hypothesis W is on the input samples Z, the more effectively the learning algorithm generalizes. In Theorem 3.2, this principle is extended to multivariate loss functions, encompassing diverse learning paradigms with arbitrary $m \geq 1$ including the prevalent pairwise and triplet learning settings. By assuming $n \bmod m = 0$ and taking $k = \frac{n}{m}$, we achieve a convergence rate of $O(\sqrt{m/n})$ in Theorem 3.2, aligning well with the original bound for pointwise learning in Theorem 3.1 when setting $m = 1$. Such a convergence rate also corroborates previous works investigating the uniform stability of pairwise (Lei et al., 2020; Yang et al., 2021b) and triplet (Chen et al., 2023) learning. Importantly, it reveals that the convergence rate is adversely affected by a factor of $\sqrt{m}$, attributable to the correlations among individual loss terms. We highlight that this theorem provides the first explicit linkage between the generalization error and the number of instances m involved in the loss function.

Building on the advancements made by (Hellström & Durisi, 2022b), we further develop our approach to establishing generalization bounds on the binary KL divergence between the expected empirical and population risks:

Theorem 3.3. *Assume that the loss function $\ell(\cdot, \cdot)$ is bounded within $[0, 1]$, then for any $k \in \left[1, \frac{n}{m}\right]$,*

$$d(L_n \,\|\, L) \leq \frac{1}{k|\mathsf{C}_n^{km}|} \sum_{u\in\mathsf{C}_n^{km}} I(W; Z_u).$$

In the interpolating setting, i.e. $L_n = 0$, we further have

$$L \leq \frac{1}{k|\mathsf{C}_n^{km}|} \sum_{u\in\mathsf{C}_n^{km}} I(W; Z_u).$$

Theorem 3.3 implies that a fast convergence rate $O(m/n)$ can be achieved as training risk approaches or equals zero, significantly improving the previous $O(\sqrt{m/n})$ rate. The removal of the square root enables attaining tighter bounds when the input-output MI terms are relatively small. Specifically, for any $k \in [1, \frac{n}{m}]$, the bound under the interpolating setting stipulated above becomes tighter than the prior square-root bound in Theorem 3.2 when $I(W; Z_u) \leq \frac{k}{2}$ for all $u \in \mathsf{P}_n^{km}$. Such a criterion is typically met in conventional settings, as we anticipate the upper bound to diminish to zero as $n \to \infty$, implying that $I(W; Z) = o(n)$.

The selection of an optimal k value in these bounds is not immediately apparent, as choosing a smaller k reduces both

the MI terms and the denominator. Notably, the research of (Harutyunyan et al., 2021) indicates that the upper bound in Theorem 3.1 is non-decreasing w.r.t k, suggesting that the smallest k, namely $k = 1$, provides the tightest bound. We extend this conclusion to the settings beyond pointwise learning with the following proposition:

Proposition 3.4. *Let $\phi : \mathbb{R} \mapsto \mathbb{R}$ be any non-decreasing concave function, then for any $k \in [1, \frac{n}{m} - 1]$, we have*

$$\frac{1}{|\mathsf{C}_n^{km}|} \sum_{u \in \mathsf{C}_n^{km}} \phi\left(\frac{1}{2k} I(W; Z_u)\right) \leq \frac{1}{|\mathsf{C}_n^{km+m}|} \sum_{u \in \mathsf{C}_n^{km+m}} \phi\left(\frac{1}{2(k+1)} I(W; Z_u)\right).$$

Applying $\phi(x) = \sqrt{x}$ to the proposition above confirms that $k = 1$ is indeed the optimal choice for minimizing the upper bound in Theorem 3.2. Such a conclusion also applies to Theorem 3.3 by selecting $\phi(x) = x$. While the choice of $k = \frac{n}{m}$ results in less favorable upper bounds, these findings remain instrumental in analyzing the generalization properties for iterative and noisy learning algorithms, as will be discussed in Section 3.3.

3.2. Generalization Bounds with CMI

The pioneering work of (Steinke & Zakynthinou, 2020) introduced the supersample setting as a novel approach to bound the expected generalization gap. This method involves measuring the CMI between the hypothesis W and supersample variables S, given the supersample dataset $\widetilde{Z}$. Subsequent studies by (Haghifam et al., 2020; Harutyunyan et al., 2021) further tightened and generalized this bound:

Theorem 3.5. *(Theorem 2.6 in (Harutyunyan et al., 2021)) Assume that the loss function $\ell(\cdot, \cdot)$ is bounded within $[0, 1]$, then for any $k \in [1, n]$,*

$$|\overline{\text{gen}}| \leq \frac{1}{|\mathsf{C}_n^k|} \sum_{u \in \mathsf{C}_n^k} \mathbb{E}_{\widetilde{Z}} \sqrt{\frac{2}{k} I^{\widetilde{Z}}(W; S_u)}.$$

Setting $k = n$ in Theorem 3.5 yields a convergence rate of $O(\sqrt{1/n})$. Extending this result to multivariate loss functions meets the same non-i.i.d challenge due to the dependencies among individual loss terms. Following the reduction techniques explored in the previous section, we are ready to present the following theorem which implies a convergence rate of $O(\sqrt{m/n})$ for such learning paradigms:

Theorem 3.6. *Assume that the loss function $\ell(\cdot, \cdot)$ is bounded within $[0, 1]$, then for any $k \in \left[1, \frac{n}{m}\right]$,*

$$|\overline{\text{gen}}| \leq \frac{1}{|\mathsf{C}_n^{km}|} \sum_{u \in \mathsf{C}_n^{km}} \mathbb{E}_{\widetilde{Z}} \sqrt{\frac{2}{k} I^{\widetilde{Z}}(W; S_u)}.$$

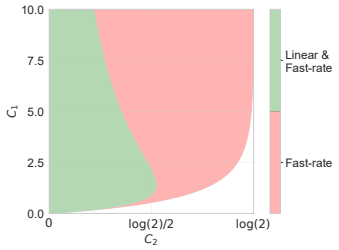


Figure 1: Comparison of the available ranges of C_1 and C_2 in Theorem 3.9 (Linear) and Theorem 3.10 (Fast-rate).

Our Theorem 3.6 generalizes Theorem 3.5 by incorporating scenarios where $m > 1$. Note that these bounds are tighter compared to the original ones in (Steinke & Zakynthinou, 2020), as we move the expectation over $\widetilde{Z}$ outside the square root. This adjustment allows for a straightforward relaxation to the CMI $I(W; S_u|\widetilde{Z})$ using Jensen's inequality. In parallel with the development in Theorem 3.3, we further upper bound the binary KL divergence between the empirical risk L_n and the mean of the empirical and population risks $(L_n + L)/2$ as follows:

Theorem 3.7. *Assume that the loss function $\ell(\cdot, \cdot)$ is bounded within $[0, 1]$, then for any $k \in \left[1, \frac{n}{m}\right]$,*

$$d\left(L_n \,\middle\|\, \frac{L_n + L}{2}\right) \leq \frac{1}{k|\mathsf{C}_n^{km}|} \sum_{u \in \mathsf{C}_n^{km}} I(W; S_u|\widetilde{Z}).$$

The inherent properties of CMI ensure that $I(W; S_u|\widetilde{Z}) \leq H(S_u) = km \log 2$, thus guaranteeing the finiteness of these CMI-based generalization bounds. Moreover, it can be shown that $I(W; S|\widetilde{Z})$ is consistently tighter than the input-output MI $I(W; \widetilde{Z}_S)$: Notice the Markov chain $(\widetilde{Z}, S) - \widetilde{Z}_S - W$, we find $I(W; \widetilde{Z}, S|\widetilde{Z}_S) = 0$, which leads to $I(W; \widetilde{Z}_S) = I(W; \widetilde{Z}, S) = I(W; S|\widetilde{Z}) + I(W; \widetilde{Z})$.

Building upon the previous methodologies outlined in Proposition 3.4, we extend the analysis to the disintegrated MI $I^{\widetilde{z}}(W; S_u)$. By employing the functions $\phi(x) = \sqrt{x}$ or $\phi(x) = x$, and subsequently taking an expectation over $\widetilde{Z}$, we establish that Theorems 3.6 and 3.7 are both non-decreasing w.r.t the parameter k. Consequently, selecting $k = 1$ emerges as the optimal choice for attaining the tightest estimates of the generalization error.

Proposition 3.8. *Let $\phi : \mathbb{R} \mapsto \mathbb{R}$ be any non-decreasing concave function, then for any $k \in [1, \frac{n}{m} - 1]$ and $\widetilde{z} \in \mathcal{Z}^{2n}$,*

$$\frac{1}{|\mathsf{C}_n^{km}|} \sum_{u \in \mathsf{C}_n^{km}} \phi\left(\frac{2}{k} I^{\widetilde{z}}(W; S_u)\right) \leq \frac{1}{|\mathsf{C}_n^{km+m}|} \sum_{u \in \mathsf{C}_n^{km+m}} \phi\left(\frac{2}{k+1} I^{\widetilde{z}}(W; S_u)\right).$$

In a parallel development, we examine the fast-rate bounds developed in (Hellström & Durisi, 2021), which utilize the weighted generalization error, $\overline{\text{gen}}_{C_1} \triangleq L - (1 + C_1)L_n$, where C_1 is a predefined constant. This framework facilitates the attainment of fast-rate bounds for the expected generalization error, which exhibits a faster scaling rate of $1/n$ as opposed to the conventional $\sqrt{1/n}$.

Theorem 3.9. *(Corollary 3 in (Hellström & Durisi, 2021)) Assume $m = 1$ and $\ell(\cdot,\cdot) \in [0,1]$, then for any $C_1, C_2 > 0$ satisfying $(C_1^2 + 2C_1 + 2)(e^{C_2} - 1 - C_2) - C_1C_2 \leq 0$,*

$$\overline{\text{gen}} \leq C_1 L_n + \frac{1}{n}\sum_{i=1}^{n} \frac{I(W; S_i|\widetilde{Z})}{C_2}.$$

We then extend this fast-rate bound to encompass multivariate loss functions and permit measuring information within subsets rather than each individual supersample variable:

Theorem 3.10. *Assume that $\ell(\cdot,\cdot) \in [0,1]$, then for any $C_2 \in (0, \log 2)$, $C_1 \geq -\frac{\log(2-e^{C_2})}{C_2} - 1$ and $k \in \left[1, \frac{n}{m}\right]$,*

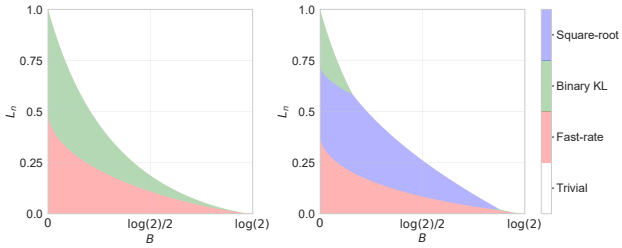
$$\overline{\text{gen}} \leq C_1 L_n + \frac{1}{|\mathsf{C}_n^{km}|}\sum_{u \in \mathsf{C}_n^{km}} \frac{I(W; S_u|\widetilde{Z})}{kC_2}.$$

In the interpolating setting, i.e. $L_n = 0$, we further have

$$L \leq \frac{1}{|\mathsf{C}_n^{km}|}\sum_{u \in \mathsf{C}_n^{km}} \frac{I(W; S_u|\widetilde{Z})}{k \log 2}.$$

Significantly, Theorem 3.10 enhances the bounds established in Theorem 3.9 by broadening the admissible intervals for C_1 and C_2. An intuitive comparison between these bounds is depicted in Figure 1. It is evident that Theorem 3.10 not only permits the selection of C_2 values greater than $\log 2/2$ but also allows for smaller values of C_1 for a given C_2, compared to Theorem 3.9. Additionally, Theorem 3.10 facilitates a seamless transition to the interpolating regime by selecting $C_2 \to \log 2$.

In this section, we introduced multiple generalization bounds without a definitive hierarchy in terms of their tightness. To facilitate a meaningful comparison, we consider each bound in its tightest form by setting $k = 1$. Additionally, we assume that the learning algorithm exhibits average indifference to permutations of the training samples, such that the values of $I^{\widetilde{z}}(W; S_u)$ are independent of the index u. This assumption simplifies our analysis and is often reasonable in traditional learning settings. In Figure 2, we conduct a comparative analysis of the square-root bound (Theorem 3.6), the binary KL bound (Theorem 3.7), and the fast-rate bound (Theorem 3.10) across wide ranges of the training risk L_n and the value of MI quantities B. Our analysis reveals that the fast-rate bound exhibits superior tightness



(a) Constant $I^{\widetilde{z}}(W; S_u)$ (b) Stochastic $I^{\widetilde{Z}}(W; S_u)$

Figure 2: Comparison between the square-root bound (Theorem 3.6), the binary KL bound (Theorem 3.7) and the fast-rate bound (Theorem 3.10) with $k = 1$. The color represents the tightest bound to characterize L. For the "Trivial" region, no bound outperforms the trivial bound: $L \leq 1$. (a) Assume that $I^{\widetilde{z}}(W; S_u) = B$ for any $\widetilde{z} \in \mathcal{Z}^{2n}$. (b) Assume that $I^{\widetilde{Z}}(W; S_u)$ forms an exponential distribution over $\widetilde{Z} \sim \mu^{2n}$, such that $\mathbb{E}_{\widetilde{Z}}[I^{\widetilde{Z}}(W; S_u)] = B$.

for smaller training risks, a situation frequently encountered in practical applications. Conversely, as L_n increases, the binary KL bound becomes more advantageous. The efficacy of the square-root bound, however, is dependent on the distribution of the disintegrated MI $I^{\widetilde{Z}}(W; S_u)$. In scenarios where $I^{\widetilde{Z}}(W; S_u)$ values are diverse, the square-root bound demonstrates predominance, particularly in intermediate regions of the L_n and B spectrum. This comparative exploration underscores the importance of context-specific selection between these bounds.

3.3. Algorithm-based Generalization Bounds

We now associate with the mini-batched iterative and noisy learning algorithms, focusing particularly on stochastic gradient Langevin dynamics (SGLD). We denote the training trajectory of SGLD as $\{W_t\}_{t=0}^{T}$, where $W_0 \in \mathbb{R}^d$ represents the randomly initialized model parameters. In the t-th update, a batch of indices $B_t \in [n]^{b \times m}$ is independently selected given the batch size b. The average gradient for this batch, G_t, is computed as follows:

$$G_t = -\frac{1}{b}\sum_{u \in B_t} \nabla_w \ell(W_{t-1}, Z_u).$$

The updating rule of SGLD can then be formalized by

$$W_t = W_{t-1} + \eta_t G_t + N_t, \quad N_t \sim N(0, \sigma_t^2 I_d),$$

where η_t denotes the learning rate, and N_t is the isotropic Gaussian noise injected in each step.

Given the complexity of multivariate loss functions, it is no longer practical to segment the training dataset into disjoint mini-batches as in (Wang et al., 2021). Consequently, techniques based on bounding point-wise stabilities $I(W; Z_i)$

become inadequate for non-pointwise learning contexts. Instead, we adopt the on-average stability approach through the input-output MI $I(W;Z)$, following the techniques outlined in (Dong et al., 2023). It is demonstrated that the input-output MI for iterative and noisy learning algorithm could be upper bounded by leveraging the determinant trajectory of the gradient covariance matrices, a more precise metric compared to the gradient variance terms explored in (Negrea et al., 2019; Wang et al., 2021):

Theorem 3.11. *(Theorem 2 in (Dong et al., 2023)) For the SGLD algorithm output W after T iterations with $m=1$, the following bound holds:*

$$I(W;Z) \leq \sum_{t=1}^{T} \frac{1}{2} \log \left| \frac{\eta_t^2}{\sigma_t^2} \mathrm{Cov}_{W_{t-1},B_t}[G_t] + I_d \right|.$$

We enhance this analysis by incorporating conditional gradient covariance, as well as extending our conclusions to non-pointwise learning scenarios:

Theorem 3.12. *For the SGLD algorithm output W after T iterations, the following bound holds:*

$$I(W;Z) \leq \sum_{t=1}^{T} \frac{1}{2} \log \left| \frac{\eta_t^2}{\sigma_t^2} \mathbb{E}_{W_{t-1}}[\Sigma_t] + I_d \right|,$$

where $\Sigma_t = \mathrm{Cov}_{B_t}[G_t]$.

According to the law of total variance, the conditional covariance measure Σ_t is strictly tighter than the unconditional one in (Dong et al., 2023). Theorem 3.12 can then be combined with generalization bounds discussed in Section 3.1 to obtain upper bounds for the SGLD algorithm:

Corollary 3.13. *Assume that $\ell(\cdot,\cdot) \in [0,1]$ and $n \bmod m = 0$, then the population risk of SGLD satisfies*

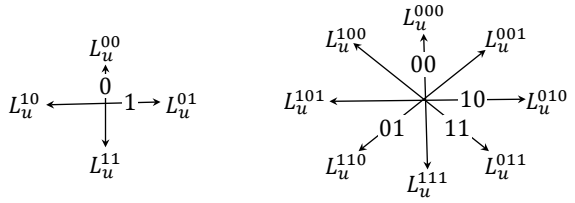
$$d(L_n \,\|\, L) \leq \frac{m}{2n} \sum_{t=1}^{T} \log \left| \frac{\eta_t^2}{\sigma_t^2} \mathbb{E}_{W_{t-1}}[\Sigma_t] + I_d \right|.$$

While our focus here is on SGLD, we note that analogous generalization bounds for the standard stochastic gradient descent (SGD) algorithm and adaptive optimization methods (e.g., AdaGrad) can also be acquired. Such extensions involve integrating the auxiliary weight process as explored in (Neu et al., 2021; Wang & Mao, 2021), following the same analytical methods developed in this paper.

4. Prediction-based Generalization Bounds

4.1. Loss-difference Generalization Bounds

The seminal work of (Harutyunyan et al., 2021) provides generalization bounds using the CMI between model outputs and supersample variables given the supersample dataset.



(a) $m=2$ (b) $m=3$

Figure 3: Demonstration of the selected pair of losses $L_u^{\Phi_u}$, according to the value of $\Phi_u \in \mathsf{B}_{m-1}$.

This methodology was further refined in (Hellström & Durisi, 2022b) by focusing on the information contained within losses, termed evaluated CMI (e-CMI). A straightforward extension of the e-CMI bound to encompass multivariate loss functions is provided as follows:

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

$$|\overline{\mathrm{gen}}| \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \sqrt{2I(L_u;S_u)}.$$

However, a computational issue arises as the dimensionality of the MI terms above scales exponentially with m. Specifically, considering that $|L_u| = |\mathsf{B}_m| = 2^m$ encapsulates losses evaluated across all combinations of training and test samples, the dimensionality of $I(L_u;S_u)$ becomes $2^m + m$. An alternative approach involves employing the loss-difference technique (Wang & Mao, 2023) to replace the losses L_u with their differences, potentially halving the dimensionality. Despite this, the computational feasibility of the bound remains daunting with the increase of m.

Such an issue is completely solved in the following theorem which achieves generalization bounds involving only two 1-dimensional variables irrespective of the value of m, by capitalizing on the independence between S_{u_1} and Φ_u:

Theorem 4.2. *Assume that $\ell(\cdot,\cdot) \in [0,1]$, then*

$$|\overline{\mathrm{gen}}| \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \sqrt{2I(\Delta_u^{\Phi_u};S_{u_1})}.$$

As illustrated in Figure 3, the variable Φ_u selects a pair of losses from L_u to calculate $\Delta_u^{\Phi_u}$, forming the Markov chain $S_{u_1} - (L_u, \Phi_u) - \Delta_u^{\Phi_u}$ and maintaining the lowest possible dimensionality for the MI terms. Utilizing the data-processing inequality and the independence between S_{u_1} and Φ_u, we then establish that:

$$\begin{aligned} I(\Delta_u^{\Phi_u};S_{u_1}) &\leq I(L_u,\Phi_u;S_{u_1}) = I(L_u;S_{u_1}|\Phi_u) \\ &= I(L_u,S_u) - I(L_u;\Phi_u). \end{aligned}$$

Thus, the bound in Theorem 4.2 is strictly tighter than the e-CMI bound in Theorem 4.1. Moreover, the MI term

$I(\Delta_u^{\Phi_u}; S_{u_1})$ can be interpreted as the rate of reliable communication over a memoryless channel with input S_{u_1} and output $\Delta_u^{\Phi_u}$, as discussed in (Wang & Mao, 2023). This conceptualization leads to a precise generalization bound for interpolating scenarios with binary loss functions:

Theorem 4.3. *Assume that $\ell(\cdot,\cdot) \in \{0,1\}$. In the interpolating setting when $L_n = 0$, we have*

$$L = \sum_{u \in \mathsf{P}_n^m} \frac{I(\Delta_u^{\Phi_u}; S_{u_1})}{|\mathsf{P}_n^m| \log 2} = \sum_{u \in \mathsf{P}_n^m} \frac{I(L_u^{\Phi_u}; S_{u_1})}{|\mathsf{P}_n^m| \log 2}.$$

Therefore, in the case of binary loss with interpolating learning algorithms, the expected population risk can be precisely characterized by the summation of sample-wise MI between S_{u_1} and either the pair of selected losses $L_u^{\Phi_u}$ or their difference $\Delta_u^{\Phi_u}$. Further refinement of the square-root bound in Theorem 4.2 is also achievable with the same development in Theorem 3.6, by additionally conditioning on $\widetilde{Z}$ and moving the expectation outside the square root:

Theorem 4.4. *Assume that $\ell(\cdot,\cdot) \in [0,1]$, then*

$$|\overline{\text{gen}}| \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{\widetilde{Z}} \sqrt{2 I^{\widetilde{Z}}(\Delta_u^{\Phi_u}; S_{u_1})}.$$

By $I(\Delta_u^{\Phi_u}; S_{u_1}) \leq I(\widetilde{Z}, \Delta_u^{\Phi_u}; S_{u_1}) = I(\Delta_u^{\Phi_u}; S_{u_1}|\widetilde{Z}) + I(\widetilde{Z}; S_{u_1})$ and considering the independence between $\widetilde{Z}$ and S_{u_1}, the MI term $I(\Delta_u^{\Phi_u}; S_{u_1})$ in Theorem 4.2 is tighter than its conditional counterpart $I(\Delta_u^{\Phi_u}; S_{u_1}|\widetilde{Z})$. Nonetheless, Theorem 4.4 may still offer tighter upper bounds when the values of $I^{\widetilde{z}}(\Delta_u^{\Phi_u}; S_{u_1})$ are scattered w.r.t $\widetilde{z} \sim \mu^{2n}$, as previously evidenced in Figure 2.

4.2. Fast-rate Generalization Bounds

We further enhance the analysis of prediction-based generalization bounds by incorporating the weighted generalization error, thereby improving the convergence rate of these bounds. It is demonstrated by (Wang & Mao, 2023) that

Theorem 4.5. *(Theorem 4.3 in (Wang & Mao, 2023)) Assume that $m = 1$ and $\ell(\cdot,\cdot) \in [0,1]$, then there exist $C_1, C_2 > 0$ such that*

$$\overline{\text{gen}} \leq C_1 L_n + \frac{1}{n} \sum_{i=1}^{n} \frac{I(L_i^0; S_i)}{C_2}.$$

Building on this premise, we extend the fast-rate bound to multivariate loss functions, tightening the result by simultaneously considering the minimum between single-loss MI $2I(L_u^{\Phi_u^+}; S_{u_1})$ and paired-loss MI $I(L_u^{\Phi_u}; S_{u_1})$. The discrepancy between these two quantities is characterized by the interaction information $I(L_u^{\Phi_u^+}; L_u^{\Phi_u^-}; S_{u_1})$, which can be either positive or negative. Consequently, there is no definitive ordering between them, and a more stringent bound can be derived by evaluating both concurrently:

Theorem 4.6. *Assume that $\ell(\cdot,\cdot) \in [0,1]$, then for any $C_2 \in (0, \log 2)$ and $C_1 \geq -\frac{\log(2-e^{C_2})}{C_2} - 1$,*

$$\overline{\text{gen}} \leq C_1 L_n + \sum_{u \in \mathsf{P}_n^m} \frac{\min\{I(L_u^{\Phi_u}; S_{u_1}), 2I(L_u^{\Phi_u^+}; S_{u_1})\}}{|\mathsf{P}_n^m| C_2}.$$

In the interpolating setting, i.e. $L_n = 0$, we further have

$$L \leq \sum_{u \in \mathsf{P}_n^m} \frac{\min\{I(L_u^{\Phi_u}; S_{u_1}), 2I(L_u^{\Phi_u^+}; S_{u_1})\}}{|\mathsf{P}_n^m| \log 2}.$$

The fast-rate bounds are typically useful when the empirical risk approaches or equals zero. Inspired by the work of (Wang & Mao, 2023) utilizing the empirical loss variance to obtain tighter generalization bounds, we further redefine loss variance for multivariate loss functions as:

$$V(\gamma) \triangleq \mathbb{E}_{W,Z}\left[\sum_{u \in \mathsf{P}_n^m} \frac{(\ell(W, Z_u) - (1+\gamma) L_Z(W))^2}{|\mathsf{P}_n^m|}\right].$$

Theorem 4.7. *Assume that $\ell(\cdot,\cdot) \in \{0,1\}$ and $\gamma \in (0,1)$, then for any $C_2 \in (0, \log 2)$ and $C_1 \geq -\frac{\log(2-e^{C_2})}{C_2 \gamma^2} - \frac{1}{\gamma^2}$,*

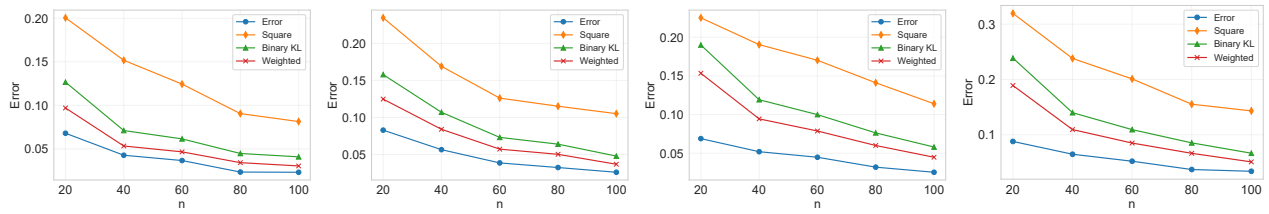
$$\overline{\text{gen}} \leq C_1 V(\gamma) + \sum_{u \in \mathsf{P}_n^m} \frac{\min\{I(L_u^{\Phi_u}; S_{u_1}), 2I(L_u^{\Phi_u^+}; S_{u_1})\}}{|\mathsf{P}_n^m| C_2}.$$

For binary loss functions, it is proven that $V(\gamma) = L_n - (1-\gamma^2)\mathbb{E}_{W,Z}[L_Z^2(W)]$ for any $\gamma \in (0,1)$. By substituting L_n with $V(\gamma)$, the loss variance bound above is tighter than Theorem 4.6 by at least $C_1(1-\gamma^2)\mathbb{E}_{W,Z}[L_Z^2(W)]$ with the same constants C_1 and C_2. Hence, Theorem 4.7 is particularly effective when the training risk is near but not exactly zero, ensuring $L_Z(W) > 0$. Alternatively, when L_n is substantially high, our analysis in Figure 2 demonstrates that the binary KL divergence bound provides the most accurate estimate for the population risk:

Theorem 4.8. *Assume that $\ell(\cdot,\cdot) \in [0,1]$, then*

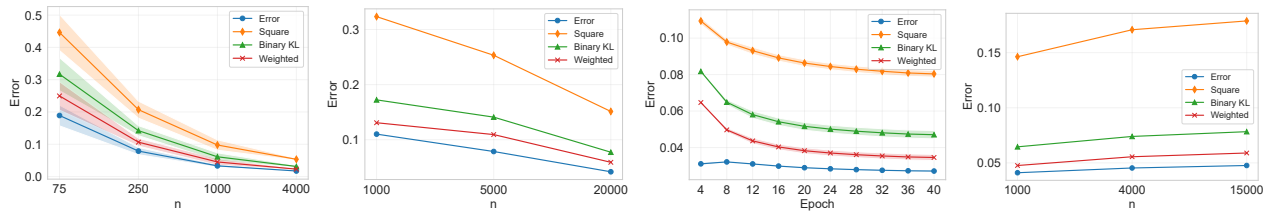
$$d\left(L_n \,\middle\|\, \frac{L_n + L}{2}\right) \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} I(L_u^{\Phi_u}; S_{u_1}).$$

Notably, the unconditional MI in Theorem 4.8 is strictly tighter than its conditional counterpart (the e-CMI bounds in (Hellström & Durisi, 2022b) for $m = 1$): due to the independence between $\widetilde{Z}$ and S_{u_1}, we have $I(L_u^{\Phi_u}; S_{u_1}) \leq I(\widetilde{Z}, L_u^{\Phi_u}; S_{u_1}) = I(L_u^{\Phi_u}; S_{u_1}|\widetilde{Z})$. An intuitive comparison between these generalization bounds is depicted in Figure 2, through examining diverse values of the training risk L_n and MI quantities B.



(a) Cross-entropy ($m = 1$) (b) Contrastive Loss ($m = 2$) (c) Triplet Loss ($m = 3$) (d) N-pair Loss ($m = 4$)

Figure 4: Comparison of the generalization bounds on synthetic Gaussian datasets, where a simple MLP network is trained with various loss functions. (a) Cross-entropy loss for pointwise learning, (b) Contrastive loss for pairwise learning, (c) Triplet loss for triplet learning, (d) N-pair loss for contrastive learning with multiple negative samples.



(a) CNN (MNIST), Adam (b) ResNet (CIFAR10), SGD (c) CNN (MNIST), SGLD (d) CLIP (Flickr30k), Adam

Figure 5: Comparison of the generalization bounds in multiple real-world learning scenarios. (a) CNN model trained on binary MNIST (4 vs 9) using Adam, (b) pretrained ResNet-50 fine-tuned on CIFAR-10 using SGD, (c) CNN model trained on binary MNIST (4 vs 9) using SGLD, (d) pretrained CLIP (ViT-B/32) fine-tuned on Flickr30k using Adam.

5. Numerical Results

In this section, we evaluate different generalization bounds developed in Section 4, utilizing a variety of synthetic and practical deep-learning settings[1]. We focus on comparing the square-root bound (Theorem 4.2), the binary KL bound (Theorem 4.8), and the fast-rate bound (Theorem 4.6). The variance-based bound (Theorem 4.7) is excluded from this comparative analysis due to its negligible difference from Theorem 4.6 within the current graphical resolution. These experiments employ a binary loss function to quantify empirical and population risks, with detailed settings and additional results delineated in Appendix E.

5.1. Synthetic Datasets

Our initial experiment encompasses a 5-class classification task, employing a simple MLP network trained on synthetic Gaussian datasets within a supervised contrastive learning framework (Khosla et al., 2020). The class centers are randomly chosen from vertices of a 5-dimensional unit hypercube. The evaluation of the generalization gap and the derived bounds are illustrated in Figure 4. The visualization results indicate that our generalization bounds adeptly adapt to diverse values of m and align well with the trend of the generalization gap: the bounds decrease as the increase of sample size n. Notably, the fast-rate bound (Theorem 4.6) emerges as the most stringent among these comparisons, corroborating our analysis of its effectiveness when the model achieves low training risks.

[1]The source code is available at `https://github.com/Yuxin-Dong/Pairwise`.

5.2. Real-world Learning Tasks

The subsequent experiment extends our analysis to multiple typical deep-learning scenarios encountered in practice. Following experiment settings in (Hellström & Durisi, 2022b), we first train a 4-layer CNN on a binarized version of the MNIST dataset, specifically focusing on the digits 4 and 9. Subsequently, we fine-tune a pretrained ResNet-50 network on the CIFAR-10 dataset. These setups follow the fully supervised learning settings. To assess the scalability of the presented generalization bounds, we additionally examine fine-tuning a CLIP (ViT-B/32) model (Radford et al., 2021) on the Flickr30k dataset in a self-supervised setting.

The comparison of the generalization gap and the upper bounds is presented in Figure 5. Notably, the empirical and population risks in self-supervised contrastive learning scenarios (e.g. CLIP and SimCLR (Chen et al., 2020)) represent a mixture of pointwise and pairwise risks, which can be effectively upper bounded by amalgamating bounds for both $m \in \{1, 2\}$. Across these experiments, the networks fit the training datasets well and the fast-rate bound (Theorem 4.6) consistently yields the tightest estimates. Interest-

ingly, in the case of CLIP, an increase in the generalization gap is observed for larger n, attributable to the presence of false negatives in ground truth labels under self-supervised contexts. Despite this, our bounds effectively capture the generalization dynamics under both fully-supervised and self-supervised learning contexts.

6. Conclusion

In this work, we develop the first series of information-theoretic generalization bounds for non-pointwise learning paradigms. These bounds augment the analysis of conventional learning algorithms and also ensure direct computational tractability. Our analysis sheds new light on understanding the generalization behavior across a spectrum of pairwise, triplet, quadruplet learning settings, and beyond.

Acknowledgements

This work was supported by National Key Research and Development Program of China(No. 2021ZD0110700), National Natural Science Foundation of China (62106191, 12071166, 62192781, 61721002).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Agarwal, S. and Niyogi, P. Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research*, 10(2), 2009.

Aminian, G., Abroshan, M., Khalili, M. M., Toni, L., and Rodrigues, M. An information-theoretical approach to semi-supervised learning under covariate-shift. In *International Conference on Artificial Intelligence and Statistics*, pp. 7433–7449. PMLR, 2022.

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

Bartlett, P., Bousquet, O., and Mendelson, S. Local rademacher complexities. *Annals of Statistics*, 33(4): 1497–1537, 2005.

Bousquet, O., Klochkov, Y., and Zhivotovskiy, N. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pp. 610–626. PMLR, 2020.

Bu, Y., Zou, S., and Veeravalli, V. V. Tightening mutual information-based bounds on generalization error. *IEEE Journal on Selected Areas in Information Theory*, 1(1): 121–130, 2020.

Bu, Y., Aminian, G., Toni, L., Wornell, G. W., and Rodrigues, M. Characterizing and understanding the generalization error of transfer learning with gibbs algorithm. In *International Conference on Artificial Intelligence and Statistics*, pp. 8673–8699. PMLR, 2022.

Cao, Q., Guo, Z.-C., and Ying, Y. Generalization bounds for metric and similarity learning. *Machine Learning*, 102 (1):115–132, 2016.

Chen, J., Chen, H., Jiang, X., Gu, B., Li, W., Gong, T., and Zheng, F. On the stability and generalization of triplet learning. *arXiv preprint arXiv:2302.09815*, 2023.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. Simclr: A simple framework for contrastive learning of visual representations. In *International Conference on Learning Representations*, volume 2, pp. 4, 2020.

Chen, W., Chen, X., Zhang, J., and Huang, K. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 403–412, 2017.

Clémençon, S., Lugosi, G., and Vayatis, N. Ranking and empirical minimization of u-statistics. *The Annals of Statistics*, pp. 844–874, 2008.

Clerico, E., Shidani, A., Deligiannidis, G., and Doucet, A. Chained generalisation bounds. In *Conference on Learning Theory*, pp. 4212–4257. PMLR, 2022.

Dong, Y., Gong, T., Chen, H., and Li, C. Understanding the generalization ability of deep learning algorithms: A kernelized rényi's entropy perspective. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 3642–3650, 2023.

Foster, D. J., Sekhari, A., and Sridharan, K. Uniform convergence of gradients for non-convex learning and optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

Ge, W. Deep metric learning with hierarchical triplet loss. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 269–285, 2018.

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*, 33:9925–9935, 2020.

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.

Harutyunyan, H., Raginsky, M., Ver Steeg, G., and Galstyan, A. Information-theoretic generalization bounds for black-box learning algorithms. *Advances in Neural Information Processing Systems*, 34:24670–24682, 2021.

He, H., Yan, H., and Tan, V. Y. Information-theoretic characterization of the generalization error for iterative semi-supervised learning. *The Journal of Machine Learning Research*, 23(1):13041–13092, 2022.

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. Evaluated cmi bounds for meta learning: Tightness and expressiveness. *Advances in Neural Information Processing Systems*, 35:20648–20660, 2022a.

Hellström, F. and Durisi, G. A new family of generalization bounds using samplewise evaluated cmi. *Advances in Neural Information Processing Systems*, 35:10108–10121, 2022b.

Huang, S., Zhou, J., Feng, H., and Zhou, D.-X. Generalization analysis of pairwise learning for ranking with deep neural networks. *Neural Computation*, pp. 1–24, 2023.

Jose, S. T., Simeone, O., and Durisi, G. Transfer meta-learning: Information-theoretic bounds and information meta-risk minimization. *IEEE Transactions on Information Theory*, 68(1):474–501, 2021.

Kar, P., Sriperumbudur, B., Jain, P., and Karnick, H. On the generalization ability of online learning algorithms for pairwise loss functions. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 441–449. PMLR, 2013.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Klochkov, Y. and Zhivotovskiy, N. Stability and deviation optimal risk bounds with convergence rate $o(1/n)$. *Advances in Neural Information Processing Systems*, 34: 5065–5076, 2021.

Lei, Y. and Ying, Y. Stochastic proximal auc maximization. *The Journal of Machine Learning Research*, 22(1):2832–2876, 2021.

Lei, Y., Lin, S.-B., and Tang, K. Generalization bounds for regularized pairwise learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 2376–2382, 2018.

Lei, Y., Ledent, A., and Kloft, M. Sharper generalization bounds for pairwise learning. *Advances in Neural Information Processing Systems*, 33:21236–21246, 2020.

Lei, Y., Liu, M., and Ying, Y. Generalization guarantee of sgd for pairwise learning. *Advances in Neural Information Processing Systems*, 34:21216–21228, 2021.

Li, S. and Liu, Y. Learning rates for nonconvex pairwise learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Liu, M., Zhang, X., Chen, Z., Wang, X., and Yang, T. Fast stochastic auc maximization with $o(1/n)$-convergence rate. In *International Conference on Machine Learning*, pp. 3189–3197. PMLR, 2018.

Masiha, M. S., Gohari, A., Yassaee, M. H., and Aref, M. R. Learning under distribution mismatch and model misspecification. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pp. 2912–2917. IEEE, 2021.

Mei, S., Bai, Y., and Montanari, A. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

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*, 32, 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*, pp. 3526–3545. PMLR, 2021.

Oh Song, H., Xiang, Y., Jegelka, S., and Savarese, S. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4004–4012, 2016.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Rammal, M. R., Achille, A., Golatkar, A., Diggavi, S., and Soatto, S. On leave-one-out conditional mutual information for generalization. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=vfCd1Vt8BGq.

Rezazadeh, A., Jose, S. T., Durisi, G., and Simeone, O. Conditional mutual information-based generalization bound for meta learning. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pp. 1176–1181. IEEE, 2021.

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*, 34:19109–19121, 2021.

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.

Schroff, F., Kalenichenko, D., and Philbin, J. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

Steinke, T. and Zakynthinou, L. Reasoning about generalization via conditional mutual information. In *Conference on Learning Theory*, pp. 3437–3452. PMLR, 2020.

Tang, H. and Liu, Y. Information-theoretic generalization bounds for transductive learning and its applications. *arXiv preprint arXiv:2311.04561*, 2023.

Wang, B., Zhang, H., Liu, P., Shen, Z., and Pineau, J. Multi-task metric learning: Theory and algorithm. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3362–3371. PMLR, 2019.

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, J., Chen, J., Chen, H., Gu, B., Li, W., and Tang, X. Stability-based generalization analysis for mixtures of pointwise and pairwise learning. *arXiv preprint arXiv:2302.09967*, 2023.

Wang, Y., Khardon, R., Pechyony, D., and Jones, R. Generalization bounds for online learning algorithms with pairwise loss functions. In *Conference on Learning Theory*, pp. 13–1. JMLR Workshop and Conference Proceedings, 2012.

Wang, Z. and Mao, Y. On the generalization of models trained with sgd: Information-theoretic bounds and implications. In *International Conference on Learning Representations*, 2021.

Wang, Z. and Mao, Y. Information-theoretic analysis of unsupervised domain adaptation. In *The Eleventh International Conference on Learning Representations*, 2022.

Wang, Z. and Mao, Y. Tighter information-theoretic generalization bounds from supersamples. *arXiv preprint arXiv:2302.02432*, 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*, 30, 2017.

Yang, Z., Lei, Y., Lyu, S., and Ying, Y. Stability and differential privacy of stochastic gradient descent for pairwise learning with non-smooth loss. In *International Conference on Artificial Intelligence and Statistics*, pp. 2026–2034. PMLR, 2021a.

Yang, Z., Lei, Y., Wang, P., Yang, T., and Ying, Y. Simple stochastic and online gradient descent algorithms for pairwise learning. *Advances in Neural Information Processing Systems*, 34:20160–20171, 2021b.

Ying, Y., Wen, L., and Lyu, S. Stochastic online auc maximization. *Advances in neural information processing systems*, 29, 2016.

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.

Zhou, R., Tian, C., and Liu, T. Stochastic chaining and strengthened information-theoretic generalization bounds. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pp. 690–695. IEEE, 2022.

A. Prerequisite Definitions and Lemmas

Definition A.1. (Subgaussian) A random variable X is σ-subgaussian if for any $\rho \in \mathbb{R}$, $\mathbb{E}[\exp(\rho(X - \mathbb{E}[X]))] \leq \exp(\rho^2\sigma^2/2)$.

Definition A.2. (Kullback-Leibler Divergence) Let P and Q be probability measures on the same space $\mathcal{X}$, the KL divergence from P to Q is defined as $D(P \,\|\, Q) \triangleq \int_{\mathcal{X}} p(x) \log(p(x)/q(x)) \,\mathrm{d}x$.

Definition A.3. (Mutual Information) Let (X, Y) be a pair of random variables with values over the space $\mathcal{X} \times \mathcal{Y}$. Let their joint distribution be $P_{X,Y}$ and the marginal distributions be P_X and P_Y respectively, the mutual information between X and Y is defined as $I(X;Y) = D(P_{X,Y} \,\|\, P_X P_Y)$.

Definition A.4. (Wasserstein Distance) Let $c(\cdot,\cdot)$ 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 p between P and Q is defined as $\mathbb{W}_p(P,Q) \triangleq \left(\inf_{\gamma \in \Gamma(P,Q)} \int_{\mathcal{X}\times\mathcal{X}} c(x,x')^p \,\mathrm{d}\gamma(x,x')\right)^{1/p}$.

Unless otherwise noted, we use $\mathbb{W}(\cdot,\cdot)$ to denote the Wasserstein distance of order 1.

Definition A.5. (Binary Relative Entropy) Let $p, q \in [0,1]$, then $d(q \,\|\, p)$ denotes the relative entropy between two Bernoulli random variables with parameters q and p respectively: $d(q \,\|\, p) = q\log(\frac{q}{p}) + (1-q)\log(\frac{1-q}{1-p})$. Given $\gamma \in \mathbb{R}$, the relaxed version of binary relative entropy is defined as $d_\gamma(q \,\|\, p) = \gamma q - \log(1 - p + pe^\gamma)$. One can verify that $\sup_\gamma d_\gamma(q \,\|\, p) = d(q \,\|\, p)$.

Lemma A.6. *(Lemma 1 in (Harutyunyan et al., 2021)) Let (X,Y) be a pair of random variables with joint distribution $P_{X,Y}$ and let $\bar{Y}$ be an independent copy of Y. If $f(x,y)$ is a measurable function such that $E_{X,Y}[f(X,Y)]$ exists and $f(X,\bar{Y})$ is σ-subgaussian, then*

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

Lemma A.7. *(Donsker-Varadhan formula) Let P and Q be probability measures defined on the same measurable space $\mathcal{X}$, where P is absolutely continuous with respect to Q. Then for any bounded measurable function $f : \mathcal{X} \mapsto \mathbb{R}$,*

$$D(P \,\|\, Q) = \sup_f \left\{\mathbb{E}_{x\sim P}[f(x)] - \log \mathbb{E}_{x\sim Q}[e^{f(x)}]\right\},$$

where X is any random variable such that e^X is Q-integrable and $\mathbb{E}_P[X]$ exists.

Lemma A.8. *(Kantorovich-Rubinstein Duality) Let P and Q be probability measures defined on the same measurable space $\mathcal{X}$, then*

$$\mathbb{W}(P,Q) = \sup_{f \in \mathrm{Lip}_1} \left\{\int_{\mathcal{X}} f \,\mathrm{d}P - \int_{\mathcal{X}} f \,\mathrm{d}Q\right\},$$

where Lip_1 denotes the set of 1-Lipschitz functions in the metric c, i.e. $|f(x) - f(x')| \leq c(x,x')$ for any $f \in \mathrm{Lip}_1$ and $x, x' \in \mathcal{X}$.

Lemma A.9. *(Lemma 2 in (Hellström & Durisi, 2022b)) Let X be a random variable that $X \in [0,1]$ almost surely and $\mathbb{E}[X] = \mu$. Then for any $\gamma \in \mathbb{R}$,*

$$\mathbb{E}\left[e^{d_\gamma(X \,\|\, \mu)}\right] \leq 1.$$

Lemma A.10. *Let $X_1, \cdots, X_n$ be independent random variables, then for any random variable Y,*

$$I(X_1;Y) + \cdots + I(X_n;Y) \leq I(X_1, \cdots, X_n; Y).$$

Proof. Since $X_1, \cdots, X_n$ are independent, we have $I(X_2, \cdots, X_n; X_1) = 0$ and

$$\begin{aligned}
I(X_1, \cdots, X_n; Y) &= I(X_1;Y) + I(X_2, \cdots, X_n; Y|X_1) \\
&= I(X_1;Y) + I(X_2, \cdots, X_n; Y) + I(X_2, \cdots, X_n; X_1|Y) - I(X_2, \cdots, X_n; X_1) \\
&= I(X_1;Y) + I(X_2, \cdots, X_n; Y) + I(X_2, \cdots, X_n; X_1|Y) \\
&\geq I(X_1;Y) + I(X_2, \cdots, X_n; Y).
\end{aligned}$$

The proof is complete by repeating the reduction steps above. Similar results also hold for disintegrated mutual information and conditional mutual information metrics. □

Lemma A.11. *Let $X \sim N(0, \Sigma)$ and Y be any zero-mean random vector satisfying $\mathrm{Cov}_Y[Y] = \Sigma$, then $H(Y) \leq H(X)$.*

Proof. From the condition $\mathrm{Cov}_Y[Y] = \Sigma$, we know that when X and Y are d-dimensional variables,

$$\int p_Y(x) x^\top \Sigma^{-1} x \,\mathrm{d}x = \int p_Y(x) \mathrm{tr}(x x^\top \Sigma^{-1}) \,\mathrm{d}x = \mathrm{tr}(\Sigma \Sigma^{-1}) = d = \int p_X(x) x^\top \Sigma^{-1} x \,\mathrm{d}x.$$

Therefore,

$$\begin{aligned}
0 \leq D(P_Y \,\|\, P_X) &= \int p_Y(x) \log \frac{p_Y(x)}{p_X(x)} \,\mathrm{d}x \\
&= -H(Y) - \int p_Y(x) \log p_X(x) \,\mathrm{d}x \\
&= -H(Y) + \int p_Y(x) \left(\frac{d}{2} \log(2\pi) + \frac{1}{2} \log|\Sigma| + \frac{1}{2} x^\top \Sigma^{-1} x \right) \mathrm{d}x \\
&= -H(Y) + \int p_X(x) \left(\frac{d}{2} \log(2\pi) + \frac{1}{2} \log|\Sigma| + \frac{1}{2} x^\top \Sigma^{-1} x \right) \mathrm{d}x \\
&= -H(Y) - H(X).
\end{aligned}$$

The proof is complete. □

Lemma A.12. *(Lemma 9 in (Dong et al., 2023)) For any symmetric positive-definite matrix A, let $A = \begin{bmatrix} B & D^\top \\ D & C \end{bmatrix}$ be a partition of A, where B and C are square matrices, then $|A| \leq |B||C|$.*

B. Omitted Proofs in Section 3

B.1. Proof of Theorem 3.2

Theorem 3.2 (Restate)**.** *Assume that $\ell(\cdot, \cdot) \in [0, 1]$, then for any $k \in \left[1, \frac{n}{m}\right]$,*

$$|\overline{\mathrm{gen}}| \leq \frac{1}{|\mathsf{C}_n^{km}|} \sum_{u \in \mathsf{C}_n^{km}} \sqrt{\frac{1}{2k} I(W; Z_u)}.$$

Proof. By the definition of the expected generalization error, we have that given i.i.d samples $Z'_{1:m} \sim \mu^m$,

$$\begin{aligned}
|\overline{\mathrm{gen}}| &= |\mathbb{E}_{W,Z}[L(W) - L_Z(W)]| \\
&= \left| \mathbb{E}_{W, Z'_{1:m}}[\ell(W, Z'_{1:m})] - \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{W, Z_u}[\ell(W, Z_u)] \right| \\
&\leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \left| \mathbb{E}_{W, Z'_{1:m}}[\ell(W, Z'_{1:m})] - \mathbb{E}_{W, Z_u}[\ell(W, Z_u)] \right|. \qquad (1)
\end{aligned}$$

Recall that $\ell(\cdot, \cdot)$ is bounded between $[0, 1]$, we know that $\ell(W, Z'_{1:m})$ is $\frac{1}{2}$-subgaussian. For any $u \in \mathsf{P}_n^m$, since Z_u consists of i.i.d samples, $Z'_{1:m}$ is an independent copy of Z_u. Then by applying Lemma A.6 with $f(W, Z_u) = \ell(W, Z_u)$, we have

$$\left| \mathbb{E}_{W, Z'_{1:m}}[\ell(W, Z'_{1:m})] - \mathbb{E}_{W, Z_u}[\ell(W, Z_u)] \right| \leq \sqrt{\frac{1}{2} I(W; Z_u)}.$$

Plugging this inequality into (1), we then have

$$|\overline{\mathrm{gen}}| \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \left| \mathbb{E}_{W, Z'_{1:m}}[\ell(W, Z'_{1:m})] - \mathbb{E}_{W, Z_u}[\ell(W, Z_u)] \right| \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \sqrt{\frac{1}{2} I(W; Z_u)}.$$

Since $I(W;Z_u)$ is invariant against permutations of samples in Z_u, we have

$$|\overline{\text{gen}}| \leq \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\sqrt{\frac{1}{2}I(W;Z_u)} = \frac{1}{|\mathsf{C}_n^m|}\sum_{u\in\mathsf{C}_n^m}\sqrt{\frac{1}{2}I(W;Z_u)}.$$

Then for any $k \in \left[1, \frac{n}{m}\right]$,

$$\begin{aligned}
|\overline{\text{gen}}| &\leq \frac{1}{|\mathsf{C}_n^m|}\sum_{u\in\mathsf{C}_n^m}\sqrt{\frac{1}{2}I(W;Z_u)} = \frac{1}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\frac{1}{|\mathsf{P}_{km}^m|}\sum_{v\in\mathsf{P}_{km}^m}\sqrt{\frac{1}{2}I(W;(Z_u)_v)} \\
&= \frac{1}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\frac{1}{k}\underbrace{\left(\frac{1}{|\mathsf{P}_{km}^m|}\sum_{v\in\mathsf{P}_{km}^m}\sqrt{\frac{1}{2}I(W;(Z_u)_v)} + \cdots + \frac{1}{|\mathsf{P}_{km}^m|}\sum_{v\in\mathsf{P}_{km}^m}\sqrt{\frac{1}{2}I(W;(Z_u)_v)}\right)}_{\times k} \\
&= \frac{1}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\frac{1}{|\mathsf{P}_{km}^{km}|}\sum_{v\in\mathsf{P}_{km}^{km}}\frac{1}{k}\left(\sqrt{\frac{1}{2}I(W;((Z_u)_v)_{1:m})} + \cdots + \sqrt{\frac{1}{2}I(W;((Z_u)_v)_{(k-1)m+1:km})}\right) \\
&\leq \frac{1}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\frac{1}{|\mathsf{P}_{km}^{km}|}\sum_{v\in\mathsf{P}_{km}^{km}}\sqrt{\frac{1}{2k}\left(I(W;((Z_u)_v)_{1:m}) + \cdots + I(W;((Z_u)_v)_{(k-1)m+1:km})\right)} \qquad (2) \\
&\leq \frac{1}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\frac{1}{|\mathsf{P}_{km}^{km}|}\sum_{v\in\mathsf{P}_{km}^{km}}\sqrt{\frac{1}{2k}I(W;(Z_u)_v)} \qquad (3) \\
&= \frac{1}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\sqrt{\frac{1}{2k}I(W;Z_u)},
\end{aligned}$$

where (2) follows by applying Jensen's inequality on the concave square-root function, and (3) follows by applying Lemma A.10. The proof is complete. □

B.2. Proof of Theorem 3.3

Theorem 3.3 (Restate)**.** *Assume that $\ell(\cdot,\cdot) \in [0,1]$, then for any $k \in \left[1, \frac{n}{m}\right]$,*

$$d(L_n \,\|\, L) \leq \frac{1}{k|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}} I(W;Z_u).$$

Furthermore, in the interpolating setting that $L_n = 0$, we have

$$L \leq \frac{1}{k|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}} I(W;Z_u).$$

Proof. By applying Jensen's inequality on the joint convexity of $d_\gamma(\cdot \,\|\, \cdot)$, we have

$$\begin{aligned}
d(L_n \,\|\, L) = \sup_\gamma d_\gamma(L_n \,\|\, L) &= \sup_\gamma d_\gamma\left(\frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\mathbb{E}_{W,Z_u}[\ell(W,Z_u)] \,\middle\|\, \mathbb{E}_W[L(W)]\right) \\
&\leq \sup_\gamma \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\mathbb{E}_{W,Z_u}[d_\gamma(\ell(W,Z_u) \,\|\, L(W))]. \qquad (4)
\end{aligned}$$

Given i.i.d samples $Z'_{1:m} \sim \mu^m$. For any $u \in \mathsf{P}_n^m$, by applying Lemma A.7 with $P = P_{W,Z_u}$, $Q = P_W P_{Z'_{1:m}}$ and $f = d_\gamma$, we know that

$$I(W;Z_u) \geq \mathbb{E}_{W,Z_u}[d_\gamma(\ell(W,Z_u) \,\|\, L(W))] - \log \mathbb{E}_{W,Z'_{1:m}}\left[e^{d_\gamma\left(\ell(W,Z'_{1:m}) \,\|\, L(W)\right)}\right]. \qquad (5)$$

For any $w \in \mathcal{W}$, we have $\mathbb{E}_{Z'_{1:m}}[\ell(w, Z'_{1:m})] = L(w)$. Recall that $\ell(\cdot,\cdot) \in [0,1]$, then by applying Lemma A.9, we have that for any $\gamma \in \mathbb{R}$:

$$\mathbb{E}_{Z'_{1:m}}\left[e^{d_\gamma\left(\ell(w,Z'_{1:m}) \,\|\, L(w)\right)}\right] \leq 1.$$

Since W is independent of $Z'_{1:m}$, this further implies that

$$\mathbb{E}_{W,Z'_{1:m}}\left[e^{d_\gamma\left(\ell(W,Z'_{1:m}) \,\|\, L(W)\right)}\right] = \mathbb{E}_W\left[\mathbb{E}_{Z'_{1:m}}\left[e^{d_\gamma\left(\ell(W,Z'_{1:m}) \,\|\, L(W)\right)}\right]\right] \leq 1.$$

Plugging the inequality above into (5), we then get

$$\mathbb{E}_{W,Z_u}[d_\gamma(\ell(W,Z_u) \,\|\, L(W))] \leq I(W;Z_u).$$

Plugging this into (4), we obtain

$$d(L_n \,\|\, L) \leq \sup_\gamma \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\mathbb{E}_{W,Z_u}[d_\gamma(\ell(W,Z_u) \,\|\, L(W))] \leq \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m} I(W;Z_u).$$

Similarly, utilizing the permutation-invariant property of $I(W;Z_u)$ against Z_u, we have that for any $k \in \left[1, \frac{n}{m}\right]$,

$$\begin{aligned}
d(L_n \,\|\, L) &\leq \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m} I(W;Z_u) = \frac{1}{|\mathsf{C}_n^m|}\sum_{u\in\mathsf{C}_n^m} I(W;Z_u) \qquad (6)\\
&= \frac{1}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\frac{1}{|\mathsf{P}_{km}^m|}\sum_{v\in\mathsf{P}_{km}^m} I(W;(Z_u)_v)\\
&= \frac{1}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\frac{1}{k}\underbrace{\left(\frac{1}{|\mathsf{P}_{km}^m|}\sum_{v\in\mathsf{P}_{km}^m} I(W;(Z_u)_v) + \cdots + \frac{1}{|\mathsf{P}_{km}^m|}\sum_{v\in\mathsf{P}_{km}^m} I(W;(Z_u)_v)\right)}_{\times k}\\
&= \frac{1}{k|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\frac{1}{|\mathsf{P}_{km}^{km}|}\sum_{v\in\mathsf{P}_{km}^{km}}\left(I(W;((Z_u)_v)_{1:m}) + \cdots + I(W;((Z_u)_v)_{(k-1)m+1:km})\right)\\
&\leq \frac{1}{k|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\frac{1}{|\mathsf{P}_{km}^{km}|}\sum_{v\in\mathsf{P}_{km}^{km}} I(W;(Z_u)_v) \qquad (7)\\
&= \frac{1}{k|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}} I(W;Z_u),
\end{aligned}$$

where (7) is by applying Lemma A.10. Consider the case that $L_n = 0$, we have

$$d(L_n \,\|\, L) = d(0 \,\|\, L) = \log\left(\frac{1}{1-L}\right) \geq L.$$

The proof is complete. □

B.3. Proof of Proposition 3.4

Proposition 3.4 (Restate)**.** *Let $\phi : \mathbb{R} \mapsto \mathbb{R}$ be any non-decreasing concave function, then for any $k \in [1, \frac{n}{m} - 1]$,*

$$\frac{1}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\phi\left(\frac{1}{2k}I(W;Z_u)\right) \leq \frac{1}{|\mathsf{C}_n^{km+m}|}\sum_{u\in\mathsf{C}_n^{km+m}}\phi\left(\frac{1}{2(k+1)}I(W;Z_u)\right).$$

Proof. For any $u \in \mathsf{P}_n^{km+m}$, by applying the chain rule of mutual information, we have

$$I(W;Z_u) = \sum_{i=1}^{k+1} I(W;(Z_u)_{(i-1)m+1:im}|(Z_u)_{1:(i-1)m})$$

$$
\begin{aligned}
&= \sum_{i=1}^{k+1} I(W, (Z_u)_{im+1:(k+1)m}; (Z_u)_{(i-1)m+1:im} | (Z_u)_{1:(i-1)m}) \\
&\quad - I((Z_u)_{(i-1)m+1:im}; (Z_u)_{im+1:(k+1)m} | (Z_u)_{1:(i-1)m}, W) \\
&\leq \sum_{i=1}^{k+1} I(W, (Z_u)_{im+1:(k+1)m}; (Z_u)_{(i-1)m+1:im} | (Z_u)_{1:(i-1)m}) \\
&= \sum_{i=1}^{k+1} I(W; (Z_u)_{(i-1)m+1:im} | (Z_u)_{1:(i-1)m}, (Z_u)_{im+1:(k+1)m}) \\
&\quad + I((Z_u)_{(i-1)m+1:im}; (Z_u)_{im+1:(k+1)m} | (Z_u)_{1:(i-1)m}) \\
&= \sum_{i=1}^{k+1} I(W; (Z_u)_{(i-1)m+1:im} | (Z_u)_{1:(i-1)m}, (Z_u)_{im+1:(k+1)m}).
\end{aligned}
$$

Similarly, for any $u \in \mathsf{C}_n^{km+m}$, by applying the inequality above with $Z_u = (Z_u)_v$, we have

$$
\begin{aligned}
I(W; Z_u) &= \frac{1}{\left|\mathsf{P}_{km+m}^{m}\right|} \sum_{v \in \mathsf{P}_{km+m}^{m}} I(W; Z_u \setminus (Z_u)_v) + I(W; (Z_u)_v | Z_u \setminus (Z_u)_v) \\
&= \frac{1}{\left|\mathsf{P}_{km+m}^{km}\right|} \sum_{v \in \mathsf{P}_{km+m}^{km}} I(W; (Z_u)_v) + \frac{1}{k+1} \sum_{i=1}^{k+1} I(W; (Z_u)_v | Z_u \setminus (Z_u)_v) \\
&= \frac{1}{\left|\mathsf{P}_{km+m}^{km}\right|} \sum_{v \in \mathsf{P}_{km+m}^{km}} I(W; (Z_u)_v) \\
&\quad + \frac{1}{\left|\mathsf{P}_{km+m}^{km+m}\right|} \sum_{v \in \mathsf{P}_{km+m}^{km+m}} \frac{1}{k+1} \sum_{i=1}^{k+1} I(W; ((Z_u)_v)_{(i-1)m+1:im} | Z_u \setminus ((Z_u)_v)_{(i-1)m+1:im}) \\
&\geq \frac{1}{\left|\mathsf{P}_{km+m}^{km}\right|} \sum_{v \in \mathsf{P}_{km+m}^{km}} I(W; (Z_u)_v) + \frac{1}{\left|\mathsf{P}_{km+m}^{km+m}\right|} \sum_{v \in \mathsf{P}_{km+m}^{km+m}} \frac{1}{k+1} I(W; (Z_u)_v).
\end{aligned}
$$

This implies that

$$
\frac{1}{k+1} I(W; Z_u) \geq \frac{1}{k\left|\mathsf{C}_{km+m}^{km}\right|} \sum_{v \in \mathsf{C}_{km+m}^{km}} I(W; (Z_u)_v).
$$

Therefore, by applying Jensen's inequality on the concave function ϕ, we have

$$
\begin{aligned}
&\frac{1}{\left|\mathsf{C}_n^{km+m}\right|} \sum_{u \in \mathsf{C}_n^{km+m}} \phi\left(\frac{1}{2(k+1)} I(W; Z_u)\right) \\
&\geq \frac{1}{\left|\mathsf{C}_n^{km+m}\right|} \sum_{u \in \mathsf{C}_n^{km+m}} \phi\left(\frac{1}{2k\left|\mathsf{C}_{km+m}^{km}\right|} \sum_{v \in \mathsf{C}_{km+m}^{km}} I(W; (Z_u)_v)\right) \\
&\geq \frac{1}{\left|\mathsf{C}_n^{km+m}\right|} \sum_{u \in \mathsf{C}_n^{km+m}} \frac{1}{\left|\mathsf{C}_{km+m}^{km}\right|} \sum_{v \in \mathsf{C}_{km+m}^{km}} \phi\left(\frac{1}{2k} I(W; (Z_u)_v)\right) \\
&= \frac{1}{\left|\mathsf{C}_n^{km}\right|} \sum_{u \in \mathsf{C}_n^{km}} \phi\left(\frac{1}{2k} I(W; Z_u)\right).
\end{aligned}
$$

The proof is complete. □

B.4. Proof of Theorem 3.6

Theorem 3.6 (Restate)**.** *Assume that* $\ell(\cdot,\cdot) \in [0,1]$*, then for any* $k \in \left[1, \frac{n}{m}\right]$,

$$|\overline{\text{gen}}| \leq \frac{1}{|\mathsf{C}_n^{km}|} \sum_{u \in \mathsf{C}_n^{km}} \mathbb{E}_{\widetilde{Z}} \sqrt{\frac{2}{k} I^{\widetilde{Z}}(W; S_u)} \leq \frac{1}{|\mathsf{C}_n^{km}|} \sum_{u \in \mathsf{C}_n^{km}} \sqrt{\frac{2}{k} I(W; S_u | \widetilde{Z})}.$$

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

$$\begin{aligned} |\overline{\text{gen}}| &= \left| \mathbb{E}_{W,\widetilde{Z},S} \left[L_{\widetilde{Z}_{\overline{S}}}(W) - L_{\widetilde{Z}_S}(W) \right] \right| \leq \mathbb{E}_{\widetilde{Z}} \left| \mathbb{E}_{W,S|\widetilde{Z}} \left[L_{\widetilde{Z}_{\overline{S}}}(W) - L_{\widetilde{Z}_S}(W) \right] \right| \\ &\leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{\widetilde{Z}} \left| \mathbb{E}_{W,S_u|\widetilde{Z}} \left[L_u^{\overline{S}_u} - L_u^{S_u} \right] \right|. \end{aligned} \tag{8}$$

Let S' be an independent copy of S such that $S' \perp\!\!\!\perp W | \widetilde{Z} = \widetilde{z}$. For any $u \in \mathsf{P}_n^m$, by applying Lemma A.7 with $P = P_{W,S_u|\widetilde{z}}$, $Q = P_{W|\widetilde{z}} P_{S_u}$ and $f(W, S_u) = L_u^{\overline{S}_u} - L_u^{S_u}$, we have

$$\begin{aligned} I^{\widetilde{z}}(W; S_u) &= D\big(P_{W,S_u|\widetilde{z}} \,\big\|\, P_{W|\widetilde{z}} P_{S_u}\big) \\ &\geq \sup_{t \in \mathbb{R}} \left\{ \mathbb{E}_{W,S_u|\widetilde{z}} \left[t \left(L_u^{\overline{S}_u} - L_u^{S_u} \right) \right] - \log \mathbb{E}_{W,S'_u|\widetilde{z}} \left[e^{t \left(L_u^{\overline{S}'_u} - L_u^{S'_u} \right)} \right] \right\}. \end{aligned} \tag{9}$$

Notice that $f(W, S'_u) \in [-1, 1]$ and $\mathbb{E}_{W,S'_u|\widetilde{z}}[f(W, S'_u)] = 0$, then by the definition of subgaussianity, we have

$$\mathbb{E}_{W,S'_u|\widetilde{z}} \left[e^{t \left(L_u^{\overline{S}'_u} - L_u^{S'_u} \right)} \right] \leq e^{\frac{t^2}{2}}.$$

Plugging this into (9), we can get

$$I^{\widetilde{z}}(W; S_u) \geq \sup_{t \in \mathbb{R}} \left\{ \mathbb{E}_{W,S_u|\widetilde{z}} \left[t \left(L_u^{\overline{S}_u} - L_u^{S_u} \right) \right] - \frac{t^2}{2} \right\},$$

which further implies that

$$\left| \mathbb{E}_{W,S_u|\widetilde{z}} \left[L_u^{\overline{S}_u} - L_u^{S_u} \right] \right| \leq \sqrt{2 I^{\widetilde{z}}(W; S_u)}.$$

Plugging the inequality above into (8), we then have

$$|\overline{\text{gen}}| \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{\widetilde{Z}} \left| \mathbb{E}_{W,S_u|\widetilde{Z}} \left[L_u^{\overline{S}_u} - L_u^{S_u} \right] \right| \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{\widetilde{Z}} \sqrt{2 I^{\widetilde{Z}}(W; S_u)}.$$

Following the same reduction steps in the proof of Theorem 3.2, we can prove that for any $k \in \left[1, \frac{n}{m}\right]$,

$$|\overline{\text{gen}}| \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{\widetilde{Z}} \sqrt{2 I^{\widetilde{Z}}(W; S_u)} \leq \frac{1}{|\mathsf{C}_n^{km}|} \sum_{u \in \mathsf{C}_n^{km}} \mathbb{E}_{\widetilde{Z}} \sqrt{\frac{2}{k} I^{\widetilde{Z}}(W; S_u)}.$$

Finally, by applying Jensen's inequality on the square root function, we have

$$|\overline{\text{gen}}| \leq \frac{1}{|\mathsf{C}_n^{km}|} \sum_{u \in \mathsf{C}_n^{km}} \mathbb{E}_{\widetilde{Z}} \sqrt{\frac{2}{k} I^{\widetilde{Z}}(W; S_u)} \leq \frac{1}{|\mathsf{C}_n^{km}|} \sum_{u \in \mathsf{C}_n^{km}} \sqrt{\frac{2}{k} \mathbb{E}_{\widetilde{Z}} \left[I^{\widetilde{Z}}(W; S_u) \right]} = \frac{1}{|\mathsf{C}_n^{km}|} \sum_{u \in \mathsf{C}_n^{km}} \sqrt{\frac{2}{k} I(W; S_u | \widetilde{Z})}.$$

This completes the proof. □

B.5. Proof of Theorem 3.7

Theorem 3.7 (Restate)**.** *Assume that $\ell(\cdot,\cdot) \in [0,1]$, then for any $k \in \left[1, \frac{n}{m}\right]$,*

$$d\left(L_n \,\middle\|\, \frac{L_n + L}{2}\right) \leq \frac{1}{k|\mathsf{C}_n^{km}|} \sum_{u \in \mathsf{C}_n^{km}} I(W; S_u|\widetilde{Z}).$$

Furthermore, in the interpolating setting that $L_n = 0$, we have

$$L \leq \frac{2}{k|\mathsf{C}_n^{km}|} \sum_{u \in \mathsf{C}_n^{km}} I(W; S_u|\widetilde{Z}).$$

Proof. By Jensen's inequality and the joint convexity of $d_\gamma(\cdot \,\|\, \cdot)$, we have

$$\begin{aligned}
d\left(L_n \,\middle\|\, \frac{L_n + L}{2}\right) &= \sup_\gamma d_\gamma\left(L_n \,\middle\|\, \frac{L_n + L}{2}\right) \\
&\leq \sup_\gamma \mathbb{E}_{\widetilde{Z}}\left[d_\gamma\left(\mathbb{E}_{W,S|\widetilde{Z}}\left[L_{\widetilde{Z}_S}(W)\right] \,\middle\|\, \mathbb{E}_{W,S|\widetilde{Z}}\left[\frac{L_{\widetilde{Z}_S}(W) + L_{\widetilde{Z}_{\overline{S}}}(W)}{2}\right]\right)\right] \\
&= \sup_\gamma \mathbb{E}_{\widetilde{Z},\Phi}\left[d_\gamma\left(\frac{1}{|\mathsf{P}_n^m|}\sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{W,S_u|\widetilde{Z},\Phi_u}\left[L_u^{S_u}\right] \,\middle\|\, \frac{1}{|\mathsf{P}_n^m|}\sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{W|\widetilde{Z},\Phi_u}\left[\frac{L_u^{\Phi_u^+} + L_u^{\Phi_u^-}}{2}\right]\right)\right] \\
&\leq \sup_\gamma \frac{1}{|\mathsf{P}_n^m|}\sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{\widetilde{Z},\Phi_u}\left[d_\gamma\left(\mathbb{E}_{W,S_u|\widetilde{Z},\Phi_u}\left[L_u^{S_u}\right] \,\middle\|\, \mathbb{E}_{W|\widetilde{Z},\Phi_u}\left[\frac{L_u^{\Phi_u^+} + L_u^{\Phi_u^-}}{2}\right]\right)\right] \\
&\leq \sup_\gamma \frac{1}{|\mathsf{P}_n^m|}\sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{\widetilde{Z},\Phi_u}\mathbb{E}_{W,S_u|\widetilde{Z},\Phi_u}\left[d_\gamma\left(L_u^{S_u} \,\middle\|\, \frac{L_u^{\Phi_u^+} + L_u^{\Phi_u^-}}{2}\right)\right].
\end{aligned} \tag{10}$$

Let S' be an independent copy of S such that $S' \perp\!\!\!\perp W|\widetilde{Z} = \widetilde{z}$. For any $u \in \mathsf{P}_n^m$, by applying Lemma A.7 with $P = P_{W,S_u|\widetilde{z},\phi_u}$, $Q = P_{W|\widetilde{z},\phi_u}P_{S_u|\phi_u}$ and $f(W, S_u) = d_\gamma\left(L_u^{S_u} \,\middle\|\, \frac{L_u^{\phi_u^+} + L_u^{\phi_u^-}}{2}\right)$, we have

$$\begin{aligned}
I^{\widetilde{z},\phi_u}(W; S_u) &= D\big(P_{W,S_u|\widetilde{z},\phi_u} \,\big\|\, P_{W|\widetilde{z},\phi_u}P_{S_u|\phi_u}\big) \\
&\geq \mathbb{E}_{W,S_u|\widetilde{z},\phi_u}\left[d_\gamma\left(L_u^{S_u} \,\middle\|\, \frac{L_u^{\phi_u^+} + L_u^{\phi_u^-}}{2}\right)\right] - \log \mathbb{E}_{W,S'_u|\widetilde{z},\phi_u}\left[e^{d_\gamma\left(L_u^{S'_u} \,\middle\|\, \frac{L_u^{\phi_u^+} + L_u^{\phi_u^-}}{2}\right)}\right].
\end{aligned} \tag{11}$$

Notice that $\mathbb{E}_{S'_u|\phi_u}\left[L_u^{S'_u}\right] = \frac{L_u^{\phi_u^+} + L_u^{\phi_u^-}}{2}$. Then by Lemma A.9, we know that for any $\gamma \in \mathbb{R}$:

$$\mathbb{E}_{W,S'_u|\widetilde{z},\phi_u}\left[e^{d_\gamma\left(L_u^{S'_u} \,\middle\|\, \frac{L_u^{\phi_u^+} + L_u^{\phi_u^-}}{2}\right)}\right] \leq 1.$$

Plugging this into (11), we then have

$$\mathbb{E}_{W,S_u|\widetilde{z},\phi_u}\left[d_\gamma\left(L_u^{S_u} \,\middle\|\, \frac{L_u^{\phi_u^+} + L_u^{\phi_u^-}}{2}\right)\right] \leq I^{\widetilde{z},\phi_u}(W; S_u).$$

Plugging this inequality back into (10), we obtain

$$d\left(L_n \,\middle\|\, \frac{L_n + L}{2}\right) \leq \sup_\gamma \frac{1}{|\mathsf{P}_n^m|}\sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{\widetilde{Z},\Phi_u}\mathbb{E}_{W,S_u|\widetilde{Z},\Phi_u}\left[d_\gamma\left(L_u^{S_u} \,\middle\|\, \frac{L_u^{\Phi_u^+} + L_u^{\Phi_u^-}}{2}\right)\right]$$

$$
\begin{aligned}
&\leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{\widetilde{Z},\Phi_u}\Big[I^{\widetilde{Z},\Phi_u}(W;S_u)\Big] \\
&\leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \Big(I(W;S_u|\widetilde{Z},\Phi_u) + I(W;\Phi_u|\widetilde{Z})\Big) \\
&= \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} I(W;S_u,\Phi_u|\widetilde{Z}) \\
&= \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} I(W;S_u|\widetilde{Z}).
\end{aligned}
$$

Following similar reduction steps in the proof of Theorem 3.3, for any $k \in \big[1, \frac{n}{m}\big]$:

$$
d\bigg(L_n \bigg\| \frac{L_n + L}{2}\bigg) \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} I(W;S_u|\widetilde{Z}) \leq \frac{1}{|\mathsf{C}_n^{km}|} \sum_{u \in \mathsf{C}_n^{km}} I(W;S_u|\widetilde{Z}).
$$

When $L_n = 0$, we have

$$
d\bigg(L_n \bigg\| \frac{L_n + L}{2}\bigg) = d\bigg(0 \bigg\| \frac{L}{2}\bigg) \geq \frac{L}{2}.
$$

The proof is complete. □

B.6. Proof of Proposition 3.8

Proposition 3.8 (Restate)**.** *Let $\phi : \mathbb{R} \mapsto \mathbb{R}$ be any non-decreasing concave function, then for any $k \in [1, \frac{n}{m} - 1]$ and $\widetilde{z} \in \mathcal{Z}^{2n}$,*

$$
\frac{1}{|\mathsf{C}_n^{km}|} \sum_{u \in \mathsf{C}_n^{km}} \phi\bigg(\frac{2}{k} I^{\widetilde{z}}(W;S_u)\bigg) \leq \frac{1}{|\mathsf{C}_n^{km+m}|} \sum_{u \in \mathsf{C}_n^{km+m}} \phi\bigg(\frac{2}{k+1} I^{\widetilde{z}}(W;S_u)\bigg).
$$

Proof. The proof follows the same development as the proof of Proposition 3.4, by replacing the mutual information $I(W;Z_u)$ with disintegrated mutual information $I^{\widetilde{z}}(W;S_u)$. □

B.7. Proof of Theorem 3.10

Theorem 3.10 (Restate)**.** *Assume that $\ell(\cdot,\cdot) \in [0,1]$, then for any $C_2 \in (0, \log 2)$, $C_1 \geq -\frac{\log(2-e^{C_2})}{C_2} - 1$ and $k \in \big[1, \frac{n}{m}\big]$,*

$$
\overline{\text{gen}} \leq C_1 L_n + \frac{1}{|\mathsf{C}_n^{km}|} \sum_{u \in \mathsf{C}_n^{km}} \frac{I(W;S_u|\widetilde{Z})}{kC_2}.
$$

Furthermore, in the interpolating regime that $L_n = 0$, we have

$$
L \leq \frac{1}{|\mathsf{C}_n^{km}|} \sum_{u \in \mathsf{C}_n^{km}} \frac{I(W;S_u|\widetilde{Z})}{k \log 2}.
$$

Proof. Notice that

$$
\begin{aligned}
L - (1+C_1)L_n &= \mathbb{E}_{W,\widetilde{Z},S}\Big[L_{\widetilde{Z}_{\overline{S}}}(W) - (1+C_1)L_{\widetilde{Z}_S}(W)\Big] \\
&= \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{W,S_u,\widetilde{Z},\Phi_u}\Big[L_u^{\overline{S}_u} - (1+C_1)L_u^{S_u}\Big] \qquad (12) \\
&= \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{\widetilde{Z},\Phi_u}\mathbb{E}_{W,S_u|\widetilde{Z},\Phi_u}\Big[L_u^{\overline{S}_{u_1}\otimes\Phi_u} - (1+C_1)L_u^{S_{u_1}\otimes\Phi_u}\Big]
\end{aligned}
$$

Let S' be an independent copy of S. For any $u \in \mathsf{P}_n^m$, by applying Lemma A.7 with $P = P_{W,S_{u_1}|\widetilde{z},\phi_u}$, $Q = P_{W|\widetilde{z},\phi_u} P_{S_{u_1}}$ and $f(W, S_{u_1}) = L_u^{\overline{S}_{u_1}\otimes\phi_u} - C_1 L_u^{S_{u_1}\otimes\phi_u}$, we have

$$\begin{aligned}
I^{\widetilde{z},\phi_u}(W; S_{u_1}) &= D\Big(P_{W,S_{u_1}|\widetilde{z},\phi_u} \,\Big\|\, P_{W|\widetilde{z},\phi_u} P_{S_{u_1}}\Big) \\
&\geq \sup_{C_2>0} \Bigg\{ \mathbb{E}_{W,S_{u_1}|\widetilde{z},\phi_u}\Big[C_2\Big(L_u^{\overline{S}_u} - (1+C_1)L_u^{S_u}\Big)\Big] \\
&\qquad - \log \mathbb{E}_{W,S'_{u_1}|\widetilde{z},\phi_u}\Bigg[e^{C_2\left(L_u^{\overline{S}'_{u_1}\otimes\phi_u} - (1+C_1)L_u^{S'_{u_1}\otimes\phi_u}\right)}\Bigg]\Bigg\} \\
&= \sup_{C_2>0} \Bigg\{ \mathbb{E}_{W,S_{u_1}|\widetilde{z},\phi_u}\Big[C_2\Big(L_u^{\overline{S}_u} - (1+C_1)L_u^{S_u}\Big)\Big] \\
&\qquad - \log \frac{\mathbb{E}_{W|\widetilde{z},\phi_u}\Big[e^{C_2 L_u^{\phi_u^+} - C_2(1+C_1)L_u^{\phi_u^-}} + e^{C_2 L_u^{\phi_u^-} - C_2(1+C_1)L_u^{\phi_u^+}}\Big]}{2}\Bigg\}.
\end{aligned} \tag{13}$$

We intend to carefully select the values of C_1 and C_2, such that the second term on the RHS is guaranteed to be less than 0. Notice that $e^{C_2 L_u^{\phi_u^+} - C_2(1+C_1)L_u^{\phi_u^-}}$ is jointly convex w.r.t $L_u^{\phi_u^+}$ and $L_u^{\phi_u^-}$, the maximum value of this term is thus achieved at the endpoints of $L_u^{\phi_u^+}, L_u^{\phi_u^-} \in [0,1]$. When $L_u^{\phi_u^+} = L_u^{\phi_u^-} = 0$, we naturally have $e^{C_2 L_u^{\phi_u^+} - C_2(1+C_1)L_u^{\phi_u^-}} = e^{C_2 L_u^{\phi_u^-} - C_2(1+C_1)L_u^{\phi_u^+}} = 1$. When $L_u^{\phi_u^+} = L_u^{\phi_u^-} = 1$, we also have $e^{C_2 L_u^{\phi_u^+} - C_2(1+C_1)L_u^{\phi_u^-}} = e^{C_2 L_u^{\phi_u^-} - C_2(1+C_1)L_u^{\phi_u^+}} = e^{-C_2 C_1} \leq 1$. Elsewise when $L_u^{\phi_u^+} = 0$ and $L_u^{\phi_u^-} = 1$ (or $L_u^{\phi_u^+} = 1$ and $L_u^{\phi_u^-} = 0$), it suffices to select a large enough C_1 such that

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

Solving the inequality above yields $C_1 \geq -\frac{\log(2-e^{C_2})}{C_2} - 1$ and $C_2 \leq \log 2$. Under these conditions, for any $u \in \mathsf{P}_n^m$,

$$\mathbb{E}_{W|\widetilde{z},\phi_u}\Big[e^{C_2 L_u^{\phi_u^+} - C_2(1+C_1)L_u^{\phi_u^-}} + e^{C_2 L_u^{\phi_u^-} - C_2(1+C_1)L_u^{\phi_u^+}}\Big] \leq 2.$$

Applying this inequality into (13), we then get

$$\mathbb{E}_{W,S_{u_1}|\widetilde{z},\phi_u}\Big[C_2\Big(L_u^{\overline{S}_u} - (1+C_1)L_u^{S_u}\Big)\Big] \leq I^{\widetilde{z},\phi_u}(W; S_{u_1}).$$

Plugging the inequality above into (12), we obtain

$$\begin{aligned}
\overline{\text{gen}} = L - (1+C_1)L_n + C_1 L_n &\leq C_1 L_n + \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m} \mathbb{E}_{\widetilde{Z},\Phi_u}\left[\frac{I^{\widetilde{Z},\Phi_u}(W; S_{u_1})}{C_2}\right] \\
&= C_1 L_n + \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m} \frac{I(W; S_{u_1}|\widetilde{Z}, \Phi_u)}{C_2} \\
&\leq C_1 L_n + \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m} \frac{I(W; S_u|\widetilde{Z})}{C_2}.
\end{aligned}$$

Following a similar development with the proof of Theorem 3.6, we have that for any $k \in [1, \frac{n}{m}]$,

$$\overline{\text{gen}} \leq C_1 L_n + \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m} \frac{I(W; S_u|\widetilde{Z})}{C_2} \leq C_1 L_n + \frac{1}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}} \frac{I(W; S_u|\widetilde{Z})}{kC_2}.$$

In the interpolating regime where $L_n = 0$, by letting $C_2 \to \frac{\log 2}{2}$ and $C_1 \to \infty$, we have

$$L \leq \frac{1}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}} \frac{I(W; S_u|\widetilde{Z})}{k\log 2}.$$

This completes the proof. □

B.8. Proof of Theorem 3.12

Theorem 3.12 (Restate)**.** *Let W be the output of the SGLD algorithm after T updates, then*

$$I(W;Z) \leq \sum_{t=1}^{T} \frac{1}{2} \log\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{E}_{W_{t-1}}[\Sigma_t] + I_d\right|.$$

Proof. For any $t \in [1,T]$, by applying the data-processing inequality on the Markov chain $Z \to (W_{T-1}, \eta_t G_T + N_T) \to W_{T-1} + \eta_t G_T + N_T$, we have

$$\begin{aligned} I(W_T;Z) = I(W_{T-1} + \eta_t G_T + N_T; Z) &\leq I(W_{T-1}, \eta_t G_T + N_T; Z) \\ &= I(W_{T-1};Z) + I(\eta_t G_T + N_T; Z|W_{T-1}) \\ &\quad \cdots \\ &\leq \sum_{t=1}^{T} I(\eta_t G_t + N_t; Z|W_{t-1}). \end{aligned}$$

Since N_t is independent of Z and B_t, we have

$$\mathrm{Cov}_{Z,B_t,N_t}[\eta_t G_t + N_t] = \mathrm{Cov}_{Z,B_t}[\eta_t G_t] + \mathrm{Cov}_{N_t}[N_t] = \eta_t^2 \Sigma_t + \sigma_t^2 I_d.$$

By applying Lemma A.11 with $\Sigma = \eta_t^2 \Sigma_t + \sigma_t^2 I_d$, we obtain

$$\begin{aligned} I^{w_{t-1}}(\eta_t G_t + N_t; Z) &= H(\eta_t G_t + N_t | W_{t-1} = w) - H(\eta_t G_t + N_t | Z, W_{t-1} = w) \\ &\leq H(\eta_t G_t + N_t | W_{t-1} = w) - H(\eta_t G_t + N_t | Z, B_t, W_{t-1} = w) \\ &= H(\eta_t G_t + N_t | W_{t-1} = w) - H(N_t) \\ &\leq \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\left|\eta_t^2 \Sigma_t + \sigma_t^2 I_d\right| - \frac{d}{2}\log(2\pi e \sigma_t^2) \\ &= \frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\Sigma_t + I_d\right|. \end{aligned}$$

Combining the inequalities above yields:

$$\begin{aligned} I(W_T;Z) \leq \sum_{t=1}^{T} I(\eta_t G_t + N_t; Z|W_{t-1}) &= \sum_{t=1}^{T} \mathbb{E}_{W_{t-1}}\left[I^{W_{t-1}}(\eta_t G_t + N_t; Z|W_{t-1})\right] \\ &= \sum_{t=1}^{T} \mathbb{E}_{W_{t-1}}\left[\frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\Sigma_t + I_d\right|\right] \\ &\leq \sum_{t=1}^{T} \frac{1}{2}\log\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{E}_{W_{t-1}}[\Sigma_t] + I_d\right|, \end{aligned}$$

where the last inequality follows by applying Jensen's inequality on the concave log-determinant function. By recursively applying Lemma A.12 to partition the diagonal elements of Σ_t, we then have

$$\begin{aligned} \log\left|\frac{\eta_t^2}{\sigma_t^2}\mathbb{E}_{W_{t-1}}[\Sigma_t] + I_d\right| &\leq \sum_{i=1}^{d} \log\left(\frac{\eta_t^2}{\sigma_t^2}\mathbb{E}_{W_{t-1}}[(\Sigma_t)_{ii}] + 1\right) \leq d\log\left(\frac{\eta_t^2}{d\sigma_t^2}\mathbb{E}_{W_{t-1}}\left[\sum_{i=1}^{d}(\Sigma_t)_{ii}\right] + 1\right) \\ &= d\log\left(\frac{\eta_t^2}{d\sigma_t^2}\mathbb{E}_{W_{t-1}}[\mathrm{tr}(\Sigma_t)] + 1\right) = d\log\left(\frac{\eta_t^2}{d\sigma_t^2}\mathbb{V}_t + 1\right), \end{aligned}$$

where $\mathbb{V}_t$ is the conditional gradient variance:

$$\mathbb{V}_t = \mathbb{E}_{W_{t-1},B_t}\left[\left\|G_t - \mathbb{E}_{B_t|W_{t-1}}[G_t]\right\|_2^2\right].$$

Note that this metric is strictly tighter than the gradient variance defined in (Wang et al., 2021) according to the law of total variance. □

C. Omitted Proofs in Section 4

C.1. Proof of Theorem 4.1

Theorem 4.1 (Restate)**.** *Assume that* $\ell(\cdot,\cdot) \in [0,1]$*, then*

$$|\overline{\text{gen}}| \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \sqrt{2I(L_u; S_u)}.$$

Proof. Notice that

$$\begin{aligned}|\overline{\text{gen}}| = \left|\mathbb{E}_{W,\widetilde{Z},S}\left[L_{\widetilde{Z}_{\overline{S}}}(W) - L_{\widetilde{Z}_S}(W)\right]\right| &= \left|\frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{W,\widetilde{Z}_u,S_u}\left[L_u^{\overline{S}_u} - L_u^{S_u}\right]\right| \\ &\leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \left|\mathbb{E}_{S_u,L_u}\left[L_u^{\overline{S}_u} - L_u^{S_u}\right]\right|. \end{aligned} \tag{14}$$

For any $u \in \mathsf{P}_n^m$, we have $L_u^{\overline{S}_u} - L_u^{S_u} \in [-1,1]$. Therefore, $L_u^{\overline{S}_u} - L_u^{S_u}$ is 1-subgaussian. Then by applying Lemma A.6 with $f(L_u, S_u) = L_u^{\overline{S}_u} - L_u^{S_u}$, we have

$$\left|\mathbb{E}_{S_u,L_u}\left[L_u^{\overline{S}_u} - L_u^{S_u}\right] - \mathbb{E}_{S'_u,L_u}\left[L_u^{\overline{S}'_u} - L_u^{S'_u}\right]\right| \leq \sqrt{2I(L_u, S_u)}.$$

It is easy to verify that $\mathbb{E}_{S'_u,L_u}\left[L_u^{\overline{S}'_u} - L_u^{S'_u}\right] = 0$. By plugging the inequality above into (14), we then get

$$|\overline{\text{gen}}| \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \sqrt{2I(L_u, S_u)}.$$

The proof is complete. □

C.2. Proof of Theorem 4.2

Theorem 4.2 (Restate)**.** *Assume that* $\ell(\cdot,\cdot) \in [0,1]$*, then*

$$|\overline{\text{gen}}| \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \sqrt{2I(\Delta_u^{\Phi_u}; S_{u_1})}.$$

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

$$\begin{aligned}|\overline{\text{gen}}| = \left|\mathbb{E}_{W,\widetilde{Z},S}\left[L_{\widetilde{Z}_{\overline{S}}}(W) - L_{\widetilde{Z}_S}(W)\right]\right| &= \left|\frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{W,\widetilde{Z}_u,S_u}\left[L_u^{\overline{S}_u} - L_u^{S_u}\right]\right| \\ &\leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \left|\mathbb{E}_{S_u,L_u}\left[L_u^{\overline{S}_u} - L_u^{S_u}\right]\right| \\ &= \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \left|\mathbb{E}_{S_u,L_u,\Phi_u}\left[(-1)^{S_{u_1}}\left(L_u^{\Phi_u^+} - L_u^{\Phi_u^-}\right)\right]\right| \\ &\leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \left|\mathbb{E}_{S_u,L_u,\Phi_u}\left[(-1)^{S_{u_1}}\Delta_u^{\Phi_u}\right]\right|. \end{aligned} \tag{15}$$

For any $u \in \mathsf{P}_n^m$, we have $\Delta_u^{\Phi_u} \in [-1,1]$. Therefore, $(-1)^{S'_{u_1}}\Delta_u^{\Phi_u}$ is 1-subgaussian. Then by applying Lemma A.6 with $f(S_{u_1}, \Delta_u^{\Phi_u}) = (-1)^{S_{u_1}}\Delta_u^{\Phi_u}$, we have

$$\left|\mathbb{E}_{S_u,L_u,\Phi_u}\left[(-1)^{S_{u_1}}\Delta_u^{\Phi_u}\right] - \mathbb{E}_{S'_u,L_u,\Phi_u}\left[(-1)^{S'_{u_1}}\Delta_u^{\Phi_u}\right]\right| \leq \sqrt{2I(\Delta_u^{\Phi_u}; S_{u_1})}.$$

Notice that $\mathbb{E}_{S'_u, L_u, \Phi_u}\left[(-1)^{S'_{u_1}} \Delta_u^{\Phi_u}\right] = 0$, then by plugging the inequality above into (15), we can get

$$|\overline{\text{gen}}| \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \left|\mathbb{E}_{S_u, L_u, \Phi_u}\left[(-1)^{S_{u_1}} \Delta_u^{\Phi_u}\right]\right| \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \sqrt{2I(\Delta_u^{\Phi_u}; S_{u_1})}.$$

The proof is complete. □

C.3. Proof of Theorem 4.3

Theorem 4.3 (Restate)**.** *Assume that $\ell(\cdot,\cdot) \in \{0,1\}$ and $L_n = 0$, then*

$$L = \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \frac{I(\Delta_u^{\Phi_u}; S_{u_1})}{\log 2} = \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \frac{I(L_u^{\Phi_u}; S_{u_1})}{\log 2} \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \frac{2I(L_u^{\Phi_u^+}; S_{u_1})}{\log 2}.$$

Proof. According to the assumption that $\ell(\cdot,\cdot) \in \{0,1\}$ and $L_n = 0$, we have

$$\begin{aligned} L &= \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{S_u, L_u}\left[L_u^{\overline{S}_u}\right] \\ &= \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \frac{\mathbb{E}_{L_u, \Phi_u | S_{u_1}=0}\left[L_u^{\Phi_u^+}\right] + \mathbb{E}_{L_u, \Phi_u | S_{u_1}=1}\left[L_u^{\Phi_u^+}\right]}{2} \\ &= \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \frac{P\big(\Delta_u^{\Phi_u} = 1 | S_{u_1} = 0\big) + P\big(\Delta_u^{\Phi_u} = -1 | S_{u_1} = 1\big)}{2}. \end{aligned} \tag{16}$$

Notice that the distribution of samplewise training loss $L_u^{S_u}$ (or test loss $L_u^{\overline{S}_u}$) should be identical regardless of the value of S_{u_1}. Therefore, the distributions of $L_u^{\Phi_u^+}$ and $L_u^{\Phi_u^-}$ are symmetric given S_{u_1}, i.e. $P_{L_u^{\Phi_u^+}|S_{u_1}=0} = P_{L_u^{\Phi_u^-}|S_{u_1}=1}$ and $P_{L_u^{\Phi_u^+}|S_{u_1}=1} = P_{L_u^{\Phi_u^-}|S_{u_1}=0}$. We then have that $P(\Delta_u^{\Phi_u} = 1 | S_{u_1} = 0) = P(\Delta_u^{\Phi_u} = -1 | S_{u_1} = 1)$, $P(\Delta_u^{\Phi_u} = 0 | S_{u_1} = 0) = P(\Delta_u^{\Phi_u} = 0 | S_{u_1} = 1)$ and $P(\Delta_u^{\Phi_u} = 1 | S_{u_1} = 1) = P(\Delta_u^{\Phi_u} = -1 | S_{u_1} = 0) = 0$.

Let $\alpha_u = P(\Delta_u^{\Phi_u} = 1 | S_{u_1} = 0)$, then $P(\Delta_u^{\Phi_u} = 0 | S_{u_1} = 0) = 1 - \alpha_u$ and

$$\begin{aligned} I(\Delta_u^{\Phi_u}; S_{u_1}) &= H(\Delta_u^{\Phi_u}) - H(\Delta_u^{\Phi_u} | S_{u_1}) = H\Big(\frac{\alpha_u}{2}, 1 - \alpha_u, \frac{\alpha_u}{2}\Big) - H(\alpha_u, 1 - \alpha_u) \\ &= -\alpha_u \log\Big(\frac{\alpha_u}{2}\Big) + \alpha_u \log(\alpha_u) = \alpha_u \log 2. \end{aligned}$$

Plugging this equality into (16), we then have

$$|\overline{\text{gen}}| = \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \frac{P\big(\Delta_u^{\Phi_u} = 1 | S_{u_1} = 0\big) + P\big(\Delta_u^{\Phi_u} = -1 | S_{u_1} = 1\big)}{2} = \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \alpha_u = \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \frac{I(\Delta_u^{\Phi_u}; S_{u_1})}{\log 2}.$$

By assuming $L_n = 0$, we know that $P(L_u^{\Phi_u^+} = 1, L_u^{\Phi_u^-} = 1) = 0$. Therefore, there exists a bijection between $\Delta_u^{\Phi_u}$ and $L_u^{\Phi_u}$: $\Delta_u^{\Phi_u} = 0 \leftrightarrow L_u^{\Phi_u} = \{0,0\}$, $\Delta_u^{\Phi_u} = 1 \leftrightarrow L_u^{\Phi_u} = \{0,1\}$ and $\Delta_u^{\Phi_u} = -1 \leftrightarrow L_u^{\Phi_u} = \{1,0\}$. Then by the data-processing inequality, we know that $I(\Delta_u^{\Phi_u}; S_{u_1}) = I(L_u^{\Phi_u}; S_{u_1})$, and

$$\begin{aligned} I(L_u^{\Phi_u^+}; S_{u_1}) &= H(L_u^{\Phi_u^+}) - H(L_u^{\Phi_u^+} | S_{u_1}) = H\Big(\frac{\alpha_u}{2}, 1 - \frac{\alpha_u}{2}\Big) - \frac{1}{2} H(\alpha_u, 1 - \alpha_u) \\ &= -\frac{\alpha_u}{2} \log\Big(\frac{\alpha_u}{2}\Big) - \Big(1 - \frac{\alpha_u}{2}\Big) \log\Big(1 - \frac{\alpha_u}{2}\Big) + \frac{\alpha_u}{2} \log(\alpha_u) + \frac{1 - \alpha_u}{2} \log(1 - \alpha_u) \\ &\geq -\frac{\alpha_u}{2} \log\Big(\frac{\alpha_u}{2}\Big) + \frac{\alpha_u}{2} \log(\alpha_u) = \frac{\alpha_u}{2} \log 2, \end{aligned}$$

where the inequality above follows by applying Jensen's inequality on the convex function $f(x) = (1 - x)\log(1 - x)$, such that $\frac{f(0) + f(\alpha_u)}{2} \geq f(\frac{\alpha_u}{2})$. The proof is complete. □

C.4. Proof of Theorem 4.4

Theorem 4.4 (Restate)**.** *Assume that* $\ell(\cdot,\cdot) \in [0,1]$*, then*

$$|\overline{\text{gen}}| \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{\widetilde{Z}} \sqrt{2I^{\widetilde{Z}}(\Delta_u^{\Phi_u}; S_{u_1})} \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \sqrt{2I(\Delta_u^{\Phi_u}; S_{u_1}|\widetilde{Z})}.$$

Proof. From (15), we know that

$$|\overline{\text{gen}}| \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \left|\mathbb{E}_{S_u, L_u, \Phi_u}\left[(-1)^{S_{u_1}} \Delta_u^{\Phi_u}\right]\right| \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{\widetilde{Z}} \left|\mathbb{E}_{S_u, L_u, \Phi_u|\widetilde{Z}}\left[(-1)^{S_{u_1}} \Delta_u^{\Phi_u}\right]\right|. \tag{17}$$

Let S' be an independent copy of S such that $S' \perp\!\!\!\perp W|\widetilde{Z} = \widetilde{z}$. For any $u \in \mathsf{P}_n^m$, by applying Lemma A.7 with $P = P_{S_{u_1}, L_u, \Phi_u|\widetilde{z}}$, $Q = P_{L_u, \Phi_u|\widetilde{z}} P_{S_{u_1}}$ and $f(S_{u_1}, \Delta_u^{\Phi_u}) = (-1)^{S_{u_1}} \Delta_u^{\Phi_u}$, we have

$$\begin{aligned} I^{\widetilde{z}}(\Delta_u^{\Phi_u}; S_{u_1}) &= D\Big(P_{S_{u_1}, L_u, \Phi_u|\widetilde{z}} \,\Big\|\, P_{L_u, \Phi_u|\widetilde{z}} P_{S_{u_1}}\Big) \\ &\geq \sup_{t \in \mathbb{R}} \left\{ \mathbb{E}_{S_u, L_u, \Phi_u|\widetilde{Z}}\left[t(-1)^{S_{u_1}} \Delta_u^{\Phi_u}\right] - \log \mathbb{E}_{S'_u, L_u, \Phi_u|\widetilde{Z}}\left[e^{t(-1)^{S'_{u_1}} \Delta_u^{\Phi_u}}\right] \right\}. \end{aligned} \tag{18}$$

Recall that $(-1)^{S'_{u_1}} \Delta_u^{\Phi_u} \in [-1,1]$, then by subgaussianity, we have

$$\mathbb{E}_{S'_u, L_u, \Phi_u|\widetilde{Z}}\left[e^{t(-1)^{S'_{u_1}} \Delta_u^{\Phi_u}}\right] \leq e^{\frac{t^2}{2}}.$$

Plugging this into (18), we have

$$I^{\widetilde{z}}(\Delta_u^{\Phi_u}; S_{u_1}) \geq \sup_{t \in \mathbb{R}} \left\{ \mathbb{E}_{S_u, L_u, \Phi_u|\widetilde{Z}}\left[t(-1)^{S_{u_1}} \Delta_u^{\Phi_u}\right] - \frac{t^2}{2} \right\}.$$

This further implies that

$$\left|\mathbb{E}_{S_u, L_u, \Phi_u|\widetilde{Z}}\left[(-1)^{S_{u_1}} \Delta_u^{\Phi_u}\right]\right| \leq \sqrt{2I^{\widetilde{z}}(\Delta_u^{\Phi_u}; S_{u_1})}.$$

Plugging this inequality into (17), we obtain

$$|\overline{\text{gen}}| \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{\widetilde{Z}} \left|\mathbb{E}_{S_u, L_u, \Phi_u|\widetilde{Z}}\left[(-1)^{S_{u_1}} \Delta_u^{\Phi_u}\right]\right| \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{\widetilde{Z}} \sqrt{2I^{\widetilde{z}}(\Delta_u^{\Phi_u}; S_{u_1})}.$$

Finally, by applying Jensen's inequality on the concave square root function, we finish the proof by

$$|\overline{\text{gen}}| \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{\widetilde{Z}} \sqrt{2I^{\widetilde{z}}(\Delta_u^{\Phi_u}; S_{u_1})} \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \sqrt{\mathbb{E}_{\widetilde{Z}}\left[2I^{\widetilde{z}}(\Delta_u^{\Phi_u}; S_{u_1})\right]} = \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \sqrt{2I(\Delta_u^{\Phi_u}; S_{u_1}|\widetilde{Z})}.$$

□

C.5. Proof of Theorem 4.6

Theorem 4.6 (Restate)**.** *Assume that* $\ell(\cdot,\cdot) \in [0,1]$*, then for any* $C_2 \in (0, \log 2)$ *and* $C_1 \geq -\frac{\log(2-e^{C_2})}{C_2} - 1$*,*

$$\overline{\text{gen}} \leq C_1 L_n + \sum_{u \in \mathsf{P}_n^m} \frac{I(L_u^{\Phi_u}; S_{u_1})}{|\mathsf{P}_n^m| C_2}.$$

Furthermore, in the interpolating regime that $L_n = 0$*, we have*

$$L \leq \sum_{u \in \mathsf{P}_n^m} \frac{I(L_u^{\Phi_u}; S_{u_1})}{|\mathsf{P}_n^m| \log 2}.$$

Proof. This result can be obtained by following the same development as the proof of Theorem 3.10 by replacing the mutual information in (13) by $I(L_u^{\Phi_u}; S_{u_1})$. □

Theorem 4.6 (Restate)**.** *Assume that $\ell(\cdot,\cdot) \in [0,1]$, then for any $C_2 \in (0, \frac{\log 2}{2})$ and $C_1 \geq -\frac{\log(2-e^{2C_2})}{2C_2} - 1$,*

$$\overline{\text{gen}} \leq C_1 L_n + \sum_{u \in \mathsf{P}_n^m} \frac{I(L_u^{\Phi_u^+}; S_{u_1})}{|\mathsf{P}_n^m| C_2}.$$

Furthermore, in the interpolating regime that $L_n = 0$, we have

$$L \leq \sum_{u \in \mathsf{P}_n^m} \frac{2I(L_u^{\Phi_u^+}; S_{u_1})}{|\mathsf{P}_n^m| \log 2}.$$

Proof. Notice that

$$\begin{aligned}
L - (1+C_1)L_n &= \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{S_u, L_u}\left[L_u^{\overline{S}_u} - (1+C_1) L_u^{S_u}\right] \\
&= \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{S_u, L_u}\left[\left(1 + \frac{C_1}{2}\right)\left(L_u^{\overline{S}_u} - L_u^{S_u}\right) - \frac{C_1}{2} L_u^{\overline{S}_u} - \frac{C_1}{2} L_u^{S_u}\right] \\
&= \frac{1}{2|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \Big(\mathbb{E}_{S_{u_1}, L_u, \Phi_u}\left[(C_1+2)(-1)^{S_{u_1}} L_u^{\Phi_u^+} - C_1 L_u^{\Phi_u^+}\right] \\
&\qquad + \mathbb{E}_{S_{u_1}, L_u, \Phi_u}\left[-(C_1+2)(-1)^{S_{u_1}} L_u^{\Phi_u^-} - C_1 L_u^{\Phi_u^-}\right]\Big).
\end{aligned} \tag{19}$$

Recall that in (28) we proved $\mathbb{E}_{S_{u_1}, L_u, \Phi_u}\left[(-1)^{S_{u_1}} L_u^{\Phi_u^+}\right] = -\mathbb{E}_{S_{u_1}, L_u, \Phi_u}\left[(-1)^{S_{u_1}} L_u^{\Phi_u^-}\right]$. Additionally, notice that $P_{L_u^{\Phi_u^+}} = P_{L_u^{\Phi_u^-}}$, we then have $\mathbb{E}_{L_u^{\Phi_u^+}}\left[L_u^{\Phi_u^+}\right] = \mathbb{E}_{L_u^{\Phi_u^-}}\left[L_u^{\Phi_u^-}\right]$. Plugging these into (19), we then have

$$L - (1+C_1)L_n = \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{S_{u_1}, L_u, \Phi_u}\left[(C_1+2)(-1)^{S_{u_1}} L_u^{\Phi_u^+} - C_1 L_u^{\Phi_u^+}\right]. \tag{20}$$

For any $u \in \mathsf{P}_n^m$, by applying Lemma A.7 with $P = P_{L_u^{\Phi_u^+}, S_{u_1}}$, $Q = P_{L_u^{\Phi_u^+}} P_{S_{u_1}}$ and $f(L_u^{\Phi_u^+}, S_{u_1}) = C_2(C_1 + 2)(-1)^{S_{u_1}} L_u^{\Phi_u^+} - C_2 C_1 L_u^{\Phi_u^+}$, we then have

$$\begin{aligned}
I(L_u^{\Phi_u^+}; S_{u_1}) = D\left(P_{L_u^{\Phi_u^+}, S_{u_1}} \,\middle\|\, P_{L_u^{\Phi_u^+}} P_{S_{u_1}}\right) &\geq \sup_{C_2 > 0} \Big\{\mathbb{E}_{S_{u_1}, L_u, \Phi_u}\left[C_2(C_1+2)(-1)^{S_{u_1}} L_u^{\Phi_u^+} - C_2 C_1 L_u^{\Phi_u^+}\right] \\
&\qquad - \log \mathbb{E}_{S'_{u_1}, L_u, \Phi_u}\left[e^{C_2(C_1+2)(-1)^{S'_{u_1}} L_u^{\Phi_u^+} - C_2 C_1 L_u^{\Phi_u^+}}\right]\Big\} \\
&= \sup_{C_2 > 0} \Bigg\{\mathbb{E}_{S_{u_1}, L_u, \Phi_u}\left[C_2(C_1+2)(-1)^{S_{u_1}} L_u^{\Phi_u^+} - C_2 C_1 L_u^{\Phi_u^+}\right] \\
&\qquad - \log \frac{\mathbb{E}_{L_u, \Phi_u}\left[e^{-2C_2(C_1+1) L_u^{\Phi_u^+}} + e^{2C_2 L_u^{\Phi_u^+}}\right]}{2}\Bigg\}.
\end{aligned} \tag{21}$$

We intend to carefully select the values of C_1 and C_2, such that the second term on the RHS is guaranteed to be less than 0. Notice that $e^{-2C_2(C_1+1)L_u^{\Phi_u^+}}$ and $e^{2C_2 L_u^{\Phi_u^+}}$ are both convex functions of $L_u^{\Phi_u^+}$, the maximum value of this term is achieved at the endpoints of $L_u^{\Phi_u^+} \in [0,1]$. When $L_u^{\Phi_u^+} = 0$, we naturally have $e^{-2C_2(C_1+1)L_u^{\Phi_u^+}} + e^{2C_2 L_u^{\Phi_u^+}} = 2$. Elsewise when $L_u^{\Phi_u^+} = 1$, it suffices to select a large enough C_1 such that

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

Solving the inequality above yields $C_1 \geq -\frac{\log(2-e^{2C_2})}{2C_2} - 1$ and $C_2 \leq \frac{\log 2}{2}$. Under these conditions, for any $u \in \mathsf{P}_n^m$,

$$\mathbb{E}_{L_u,\Phi_u}\left[e^{-2C_2(C_1+1)L_u^{\Phi_u^+}} + e^{2C_2 L_u^{\Phi_u^+}}\right] \leq 2.$$

Applying this inequality into (21), we then get

$$\mathbb{E}_{S_{u_1},L_u,\Phi_u}\left[C_2(C_1+2)(-1)^{S_{u_1}} L_u^{\Phi_u^+} - C_2 C_1 L_u^{\Phi_u^+}\right] \leq I(L_u^{\Phi_u^+}; S_{u_1}).$$

Plugging the inequality above into (20), we obtain

$$\overline{\text{gen}} = L - (1+C_1)L_n + C_1 L_n \leq C_1 L_n + \sum_{u\in\mathsf{P}_n^m} \frac{I(L_u^{\Phi_u^+}; S_{u_1})}{|\mathsf{P}_n^m| C_2}.$$

In the interpolating regime where $L_n = 0$, by letting $C_2 \to \frac{\log 2}{2}$ and $C_1 \to \infty$, we have

$$L \leq \sum_{u\in\mathsf{P}_n^m} \frac{2I(L_u^{\Phi_u^+}; S_{u_1})}{|\mathsf{P}_n^m| \log 2}.$$

This completes the proof. □

C.6. Proof of Theorem 4.7

Theorem 4.7 (Restate)**.** *Assume that $\ell(\cdot,\cdot) \in \{0,1\}$ and $\gamma \in (0,1)$, then for any $C_2 \in (0, \frac{\log 2}{2})$ and $C_1 \geq -\frac{\log(2-e^{2C_2})}{2C_2\gamma^2} - \frac{1}{\gamma^2}$,*

$$\overline{\text{gen}} \leq C_1 V(\gamma) + \sum_{u\in\mathsf{P}_n^m} \frac{I(L_u^{\Phi_u^+}; S_{u_1})}{|\mathsf{P}_n^m| C_2}.$$

Proof. By the definition of γ-variance, we have

$$\begin{aligned}
V(\gamma) &= \mathbb{E}_{W,\widetilde{Z},S}\left[\frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\left(\ell(W,\widetilde{Z}_u^{S_u}) - (1+\gamma)L_{\widetilde{Z}_S}(W)\right)^2\right] \\
&= \mathbb{E}_{W,\widetilde{Z},S}\left[\frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\left(\ell^2(W,\widetilde{Z}_u^{S_u}) - 2(1+\gamma)\ell(W,\widetilde{Z}_u^{S_u})L_{\widetilde{Z}_S}(W) + (1+\gamma)^2 L^2_{\widetilde{Z}_S}(W)\right)\right] \\
&= \mathbb{E}_{W,\widetilde{Z},S}\left[\frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\ell(W,\widetilde{Z}_u^{S_u})\right] - 2(1+\gamma)\mathbb{E}_{W,\widetilde{Z},S}\left[L^2_{\widetilde{Z}_S}(W)\right] + (1+\gamma)^2\mathbb{E}_{W,\widetilde{Z},S}\left[L^2_{\widetilde{Z}_S}(W)\right] \\
&= L_n - (1-\gamma^2)\mathbb{E}_{W,\widetilde{Z},S}\left[L^2_{\widetilde{Z}_S}(W)\right].
\end{aligned}$$

Recall that $\ell(\cdot,\cdot) \in \{0,1\}$, we have $L_{\widetilde{Z}_S}(W) \in [0,1]$, $L^2_{\widetilde{Z}_S}(W) \leq L_{\widetilde{Z}_S}(W)$ and

$$\begin{aligned}
\overline{\text{gen}} - C_1 V(\gamma) &= \overline{\text{gen}} - C_1 L_n + C_1(1-\gamma^2)\mathbb{E}_{W,\widetilde{Z},S}\left[L^2_{\widetilde{Z}_S}(W)\right] \\
&\leq \overline{\text{gen}} - C_1 L_n + C_1(1-\gamma^2)\mathbb{E}_{W,\widetilde{Z},S}\left[L_{\widetilde{Z}_S}(W)\right] \\
&= \overline{\text{gen}} - C_1\gamma^2 L_n. \qquad (22)
\end{aligned}$$

By applying Theorem 4.6 with $C_1 = C_1\gamma^2$ and $C_2 = C_2$, we have

$$\overline{\text{gen}} - C_1\gamma^2 L_n \leq \sum_{u\in\mathsf{P}_n^m} \frac{I(L_u^{\Phi_u^+}; S_{u_1})}{|\mathsf{P}_n^m| C_2}, \qquad (23)$$

under the constraints that $C_2 \in (0, \frac{\log 2}{2})$ and $C_1 \geq -\frac{\log(2-e^{2C_2})}{2C_2\gamma^2} - \frac{1}{\gamma^2}$. The proof is complete by combining inequalities (22) and (23). □

C.7. Proof of Theorem 4.8

Theorem 4.8 (Restate)**.** *Assume that* $\ell(\cdot,\cdot) \in [0,1]$*, then*

$$d\left(L_n \,\middle\|\, \frac{L_n + L}{2}\right) \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} I(L_u^{\Phi_u}; S_{u_1}).$$

Furthermore, in the interpolating setting that $L_n = 0$*, we have*

$$L \leq \frac{2}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} I(L_u^{\Phi_u}; S_{u_1}).$$

Proof. Recall that in (10) we proved that

$$d\left(L_n \,\middle\|\, \frac{L_n + L}{2}\right) \leq \sup_{\gamma} \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{L_u^{\Phi_u}, S_{u_1}} \left[d_\gamma \left(L_u^{S_u} \,\middle\|\, \frac{L_u^{\Phi_u^+} + L_u^{\Phi_u^-}}{2} \right) \right]. \tag{24}$$

For any $u \in \mathsf{P}_n^m$, by applying Lemma A.7 with $P = P_{L_u^{\Phi_u}, S_{u_1}}$, $Q = P_{L_u^{\Phi_u}} P_{S_{u_1}}$ and $f(L_u^{\Phi_u}, S_{u_1}) = d_\gamma\left(L_u^{S_u} \,\middle\|\, \frac{L_u^{\Phi_u^+} + L_u^{\Phi_u^-}}{2}\right)$, we have

$$\begin{aligned} I(L_u^{\Phi_u}, S_{u_1}) &= D\left(P_{L_u^{\Phi_u}, S_{u_1}} \,\middle\|\, P_{L_u^{\Phi_u}} P_{S_{u_1}}\right) \\ &\geq \mathbb{E}_{L_u^{\Phi_u}, S_{u_1}} \left[d_\gamma \left(L_u^{S_u} \,\middle\|\, \frac{L_u^{\Phi_u^+} + L_u^{\Phi_u^-}}{2} \right) \right] - \log \mathbb{E}_{L_u^{\Phi_u}, S'_{u_1}} \left[e^{d_\gamma \left(L_u^{S'_{u_1} \otimes \Phi_u} \,\middle\|\, \frac{L_u^{\Phi_u^+} + L_u^{\Phi_u^-}}{2} \right)} \right]. \end{aligned} \tag{25}$$

Since $\mathbb{E}_{S'_{u_1}} \left[L_u^{S'_{u_1} \otimes \Phi_u} \right] = \frac{L_u^{\Phi_u^+} + L_u^{\Phi_u^-}}{2}$, by applying Lemma A.9, we have that for any $\gamma \in \mathbb{R}$,

$$\mathbb{E}_{L_u^{\Phi_u}, S'_{u_1}} \left[e^{d_\gamma \left(L_u^{S'_{u_1} \otimes \Phi_u} \,\middle\|\, \frac{L_u^{\Phi_u^+} + L_u^{\Phi_u^-}}{2} \right)} \right] \leq 1.$$

Plugging this into (25), we can get

$$\mathbb{E}_{L_u^{\Phi_u}, S_{u_1}} \left[d_\gamma \left(L_u^{S_u} \,\middle\|\, \frac{L_u^{\Phi_u^+} + L_u^{\Phi_u^-}}{2} \right) \right] \leq I(L_u^{\Phi_u}, S_{u_1}).$$

Plugging the inequality above into (24), we finally get

$$d\left(L_n \,\middle\|\, \frac{L_n + L}{2}\right) \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} I(L_u^{\Phi_u}, S_{u_1}).$$

The proof is finished. □

D. Additional Discussions and Theoretical Results

D.1. Examples of Non-pointwise Learning

Contrastive representation learning enhances the performance of machine learning models by leveraging the relational contrast between data points. In this methodology, similar samples are drawn closer together in the embedding space, while dissimilar samples are distanced from each other. This process is facilitated by a similarity metric $d : \mathcal{T} \times \mathcal{T} \mapsto \mathbb{R}^+$, which quantifies the proximity between two embeddings. The core of contrastive learning lies in the evaluation of a contrastive loss, calculated based on similarities between feature representations T_i extracted from an encoder network $f : \mathcal{X} \mapsto \mathcal{T}$ as

$T_i = f(X_i)$. The fundamental principle of contrastive learning is to maximize the distance between similar samples while minimizing the distance between dissimilar ones, often formulated as the max-margin contrastive loss:

$$\ell_{\text{contrast}}(X_i, X_j) = \mathbb{1}_{Y_i = Y_j} \cdot d(T_i, T_j) + \mathbb{1}_{Y_i \neq Y_j} \cdot \max\{\epsilon - d(T_i, T_j), 0\},$$

where ϵ is a margin hyperparameter defining the minimum distance between samples of different classes. Through this approach, the model learns robust representations, which are then transferable to downstream tasks.

The triplet loss, introduced by (Schroff et al., 2015), enhances this objective by incorporating the concept of positive and negative samples, allowing simultaneous optimization over similarities and dissimilarities:

$$\ell_{\text{triplet}}(X_i, X_+, X_-) = \max\{d(T_i, T_+) - d(T_i, T_-) + \epsilon, 0\}.$$

This formulation enforces $Y_i = Y_+$ and $Y_i \neq Y_-$, which may appear incompatible with our original problem settings discussed in Section 2. However, this issue can be addressed through a simple adaptation:

$$\ell_{\text{triplet}}(X_i, X_j, X_k) = \begin{cases} \max\{d(T_i, T_j) - d(T_i, T_k) + \epsilon, 0\}, & \text{if } Y_i = Y_j \text{ and } Y_i \neq Y_k, \\ 0, & \text{otherwise.} \end{cases}$$

Under this modified formulation, the empirical and population risks are expressed as the summation of triplet losses evaluated on every subset $u \in \mathsf{P}_n^3$, thereby facilitating the application of our generalization analysis to these learning settings.

Expanding upon this, the quadruplet loss, proposed by (Chen et al., 2017), incorporates a fourth contrastive sample:

$$\ell_{\text{quadruplet}}(X_i, X_j, X_k, X_l) = \begin{cases} \max\{d(T_i, T_j) - d(T_k, T_l) + \epsilon, 0\}, & \text{if } Y_i = Y_j \text{ and } Y_i \neq Y_k \text{ and } Y_i \neq Y_l \text{ and } Y_k \neq Y_l, \\ 0, & \text{otherwise.} \end{cases}$$

Moreover, the n-pair loss, developed by (Sohn, 2016), caters to an arbitrary number of negative samples and introduces smoothing to the maximum operation, leading to the following formulation:

$$\ell_{\text{n-pair}}(X_{1:m}) = \begin{cases} \log(1 + \sum_{i=3}^{m} \exp(d(T_1, T_i) - d(T_1, T_2))), & \text{if } Y_1 = Y_2 \text{ and } Y_1 \neq Y_i, \forall i \in [3, m], \\ 0, & \text{otherwise.} \end{cases}$$

Other prominent loss functions in contrastive learning include the NT-Xent loss (Chen et al., 2020) and the InfoNCE loss (Oord et al., 2018). These loss functions enable the consideration of an arbitrarily large number of contrastive samples. However, current generalization analyses are predominantly confined to pairwise and triplet settings. Our results, in contrast, naturally accommodate arbitrarily large values of m, thus providing a more comprehensive and flexible framework for contrastive learning analysis.

Deep metric learning focuses on quantifying the similarity between data samples. The primary objective of deep metric learning is to develop an embedding encoder, $f : \mathcal{X} \mapsto \mathcal{T}$, coupled with a distance metric, $d : \mathcal{T} \times \mathcal{T} \mapsto \mathbb{R}^+$. This framework is designed such that for any two data samples, X_i, X_j, along with their corresponding labels, Y_i, Y_j, the computed distance $d(T_i, T_j)$ yields smaller values when the labels are identical and larger values when they differ.

A significant portion of research in deep metric learning is aligned with the principles of contrastive learning. The problem formulation in deep metric learning closely resembles that of contrastive representation learning, while a key distinction lies in the nature of the distance metric d. Unlike in contrastive representation learning, where d is typically a predefined and fixed metric, deep metric learning treats d as a target of the training process, thereby making it trainable and adaptable to the specific nuances of the given data. Given this conceptual overlap, our generalization analysis is equally applicable and relevant to the settings of deep metric learning.

Ranking algorithms are designed to process sets of feature vectors and predict the optimal ordering between them. These algorithms are critical in various applications, ranging from search engines to recommendation systems. Ranking algorithms can be broadly classified into three primary methodologies, each with its unique approach to ranking items:

- Pointwise ranking: This method involves predicting a score for each individual feature vector. These scores are then used as the basis for sorting and determining the relative order of the items. Pointwise ranking treats the ranking problem as a regression or classification task, predicting scores or classes for individual items independently.

- Pairwise ranking: Pairwise ranking models operate by comparing pairs of items at a time. A typical model, denoted as $f : \mathcal{X} \times \mathcal{X} \mapsto [0, 1]$, receives two data points as input and outputs the probability of the first item being ranked higher than the second. This approach inherently focuses on the relative ordering of item pairs, thereby transforming the ranking problem into a binary classification task.
- Listwise Ranking: Diverging from pointwise and pairwise methods, listwise ranking algorithms handle an entire list of items simultaneously. The input to these models is a complete list of items, and the output is the entire ordering among them. Consequently, the number of samples m considered by the loss function in listwise ranking is dependent on the length of these item lists.

While traditional generalization bounds have been effectively applied to the analysis of pointwise and pairwise ranking algorithms, their extension to listwise ranking algorithms, particularly for large m values, has been less explored. In this regard, we first facilitate the analysis of generalization in listwise ranking algorithms with arbitrarily large list sizes.

D.2. Additional Related Works

Uniform convergence has emerged as a prominent methodology for investigating the generalization performance of pairwise learning algorithms (Bartlett et al., 2005). This approach is notable for yielding meaningful learning rates in nonconvex learning scenarios (Foster et al., 2018; Mei et al., 2018). Uniform convergence analysis has been successfully applied to specific pairwise learning contexts, such as metric learning (Cao et al., 2016), ranking (Clémençon et al., 2008) and AUC maximization (Liu et al., 2018; Lei & Ying, 2021). (Lei et al., 2018) further explored the pairwise learning framework using uniform convergence techniques. More recently, (Lei et al., 2021) developed a uniform convergence analysis of gradients for pairwise learning. Despite its advantages, the uniform convergence approach often relies on the complexity of the hypothesis space, characterized by metrics like VC dimension, covering number, and Rademacher complexity. However, these metrics tend to be scale-sensitive (Zhang et al., 2021) and pose challenges when applied to modern deep neural networks.

Algorithmic stability also plays a crucial role in the analysis of pairwise learning algorithms (Bousquet et al., 2020; Klochkov & Zhivotovskiy, 2021). This approach involves quantifying the variations in the output predictions consequent to modifications of the training dataset. (Agarwal & Niyogi, 2009) investigated the relationship between generalization and stability within ranking algorithms. (Wang et al., 2019) analyzes regularized metric learning through the lens of stability. (Yang et al., 2021a) extended the scope of stability analysis to address differential privacy problems in pairwise settings. (Lei et al., 2020) established learning rates for regularized empirical risk minimizers. Subsequently, (Lei et al., 2021) further offered generalization guarantees for pairwise SGD, broadening the applicability of these concepts under less restrictive assumptions. Despite the extensive application and theoretical significance, algorithmic stability often relies on convexity assumptions, which restricts its applicability predominantly to convex learning scenarios. Moreover, a strong convexity condition is often required when establishing faster learning rates. To our best knowledge, existing generalization studies of algorithmic stability primarily focus on pairwise (Li & Liu, 2023; Wang et al., 2023; Huang et al., 2023) and triplet (Chen et al., 2023) learning scenarios, while exploration in quadruplet learning (Chen et al., 2017) or higher-order cases (Sohn, 2016; Chen et al., 2020) is still lacking.

The utilization of **information-theoretic metrics** for analyzing the generalization properties of learning algorithms has gained significant traction, particularly following the foundational contributions of (Xu & Raginsky, 2017; Russo & Zou, 2019). These seminal works established a pivotal connection between the expected generalization error and the mutual information between the hypothesis and the training dataset. This approach has proven to be highly effective in elucidating the dynamics of noisy and iterative learning algorithms, such as SGLD (Negrea et al., 2019; Wang et al., 2021) and SGD (Neu et al., 2021; Wang & Mao, 2021; Dong et al., 2023). Additionally, this framework has been enriched and expanded through various methodologies, including conditioning (Hafez-Kolahi et al., 2020), the chaining strategy (Asadi et al., 2018; Zhou et al., 2022; Clerico et al., 2022), the random subsets or individual techniques (Bu et al., 2020; Rodríguez-Gálvez et al., 2021), and conditional information measures (Steinke & Zakynthinou, 2020; Haghifam et al., 2020). A noteworthy advancement was made by (Harutyunyan et al., 2021), who introduced an innovative method for establishing generalization bounds by leveraging the conditional mutual information between a model's output and supersample variables. This approach, which conceptualizes the neural network as a "black box", leads to a significant reduction in the dimensionality of the random variables involved, thereby enhancing the computational tractability. Building upon this foundation, subsequent studies (Hellström & Durisi, 2022b; Wang & Mao, 2023) further refined this methodology by integrating evaluated losses and loss differences, resulting in even tighter generalization bounds. Another notable development in the conditional mutual information framework is the leave-one-out setting (Haghifam et al., 2022; Rammal et al., 2022). This variant markedly decreases the sample requirement from $n \times 2$ to just $n + 1$. Beyond supervised learning contexts, information-theoretic

bounds have also been applied to analyze generalization in various learning paradigms including meta-learning (Rezazadeh et al., 2021; Jose et al., 2021; Hellström & Durisi, 2022a), semi-supervised learning (Aminian et al., 2022; He et al., 2022), and transfer learning (Wu et al., 2020; Masiha et al., 2021; Wang & Mao, 2022; Bu et al., 2022). To our best knowledge, information-theoretic generalization analysis is currently confined to pointwise learning scenarios, with extensions to even the simplest pairwise settings remaining unexplored.

D.3. Square-root Bounds with Single-loss MI

According to the Markov chain relationship $S_{u_1} - L_u^{\Phi_u} - \Delta_u^{\Phi_u}$ and by applying the data-processing inequality, the loss-difference MI $I(\Delta_u^{\Phi_u}; S_{u_1})$ is proven to be tighter than the evaluated MI $I(L_u^{\Phi_u}; S_{u_1})$. However, there is no definite ordering between the tightness of $I(\Delta_u^{\Phi_u}; S_{u_1})$ and $2I(L_u^{\Phi_u^+}; S_{u_1})$. Therefore, the square-root bound in Theorem 4.2 could be potentially improved by taking the minimum with the following single-loss bound:

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

$$|\overline{\text{gen}}| \leq \frac{2}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \sqrt{2I(L_u^{\Phi_u^+}; S_{u_1})}.$$

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

$$\begin{aligned}
\overline{\text{gen}} &= \mathbb{E}_{W,\widetilde{Z},S}\Big[L_{\widetilde{Z}_{\overline{S}}}(W) - L_{\widetilde{Z}_S}(W)\Big] \qquad (26)\\
&= \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{S_u,L_u}\Big[L_u^{\overline{S}_u} - L_u^{S_u}\Big]\\
&= \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{S_{u_1},L_u,\Phi_u}\Big[(-1)^{S_{u_1}}\Big(L_u^{\Phi_u^+} - L_u^{\Phi_u^-}\Big)\Big]\\
&= \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \Big(\mathbb{E}_{S_{u_1},L_u,\Phi_u}\Big[(-1)^{S_{u_1}} L_u^{\Phi_u^+}\Big] - \mathbb{E}_{S_{u_1},L_u,\Phi_u}\Big[(-1)^{S_{u_1}} L_u^{\Phi_u^-}\Big]\Big). \qquad (27)
\end{aligned}$$

From the analysis in the proof of Theorem 4.3 that $P_{L_u^{\Phi_u^+}|S_{u_1}=0} = P_{L_u^{\Phi_u^-}|S_{u_1}=1}$ and $P_{L_u^{\Phi_u^+}|S_{u_1}=1} = P_{L_u^{\Phi_u^-}|S_{u_1}=0}$, we then know that $\mathbb{E}_{L_u^{\Phi_u^+}|S_{u_1}=0}\Big[L_u^{\Phi_u^+}\Big] = \mathbb{E}_{L_u^{\Phi_u^-}|S_{u_1}=1}\Big[L_u^{\Phi_u^-}\Big]$ and $\mathbb{E}_{L_u^{\Phi_u^+}|S_{u_1}=1}\Big[L_u^{\Phi_u^+}\Big] = \mathbb{E}_{L_u^{\Phi_u^-}|S_{u_1}=0}\Big[L_u^{\Phi_u^-}\Big]$. Therefore,

$$\begin{aligned}
\mathbb{E}_{S_{u_1},L_u,\Phi_u}\Big[(-1)^{S_{u_1}} L_u^{\Phi_u^+}\Big] &= \frac{\mathbb{E}_{L_u^{\Phi_u^+}|S_{u_1}=0}\Big[L_u^{\Phi_u^+}\Big] - \mathbb{E}_{L_u^{\Phi_u^+}|S_{u_1}=1}\Big[L_u^{\Phi_u^+}\Big]}{2}\\
&= \frac{\mathbb{E}_{L_u^{\Phi_u^-}|S_{u_1}=1}\Big[L_u^{\Phi_u^-}\Big] - \mathbb{E}_{L_u^{\Phi_u^-}|S_{u_1}=0}\Big[L_u^{\Phi_u^-}\Big]}{2}\\
&= -\mathbb{E}_{S_{u_1},L_u,\Phi_u}\Big[(-1)^{S_{u_1}} L_u^{\Phi_u^-}\Big]. \qquad (28)
\end{aligned}$$

Plugging the equality above into (27), we then have

$$\begin{aligned}
\overline{\text{gen}} &= \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \Big(\mathbb{E}_{S_{u_1},L_u,\Phi_u}\Big[(-1)^{S_{u_1}} L_u^{\Phi_u^+}\Big] - \mathbb{E}_{S_{u_1},L_u,\Phi_u}\Big[(-1)^{S_{u_1}} L_u^{\Phi_u^-}\Big]\Big)\\
&= \frac{2}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{S_{u_1},L_u,\Phi_u}\Big[(-1)^{S_{u_1}} L_u^{\Phi_u^+}\Big] = -\frac{2}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{S_{u_1},L_u,\Phi_u}\Big[(-1)^{S_{u_1}} L_u^{\Phi_u^-}\Big]. \qquad (29)
\end{aligned}$$

Notice that for any $u \in \mathsf{P}_n^m$, $(-1)^{S_{u_1}} L_u^{\Phi_u^+} \in [-1,1]$, i.e. $(-1)^{S_{u_1}} L_u^{\Phi_u^+}$ is 1-subgaussian. Then by applying Lemma A.6 with $f(S_{u_1}, L_u^{\Phi_u^+}) = (-1)^{S_{u_1}} L_u^{\Phi_u^+}$, we have

$$\Big|\mathbb{E}_{S_{u_1},L_u,\Phi_u}\Big[(-1)^{S_{u_1}} L_u^{\Phi_u^+}\Big] - \mathbb{E}_{S'_{u_1},L_u,\Phi_u}\Big[(-1)^{S'_{u_1}} L_u^{\Phi_u^+}\Big]\Big| \leq \sqrt{2I(L_u^{\Phi_u^+}; S_{u_1})}.$$

Since $\mathbb{E}_{S'_{u_1},L_u,\Phi_u}\left[(-1)^{S'_{u_1}}L_u^{\Phi_u^+}\right]=0$, by plugging the inequality above into (29),

$$|\overline{\text{gen}}| \leq \frac{2}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\left|\mathbb{E}_{S_{u_1},L_u,\Phi_u}\left[(-1)^{S_{u_1}}L_u^{\Phi_u^+}\right]\right| \leq \frac{2}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\sqrt{2I(L_u^{\Phi_u^+};S_{u_1})}.$$

The proof is complete. □

Similarly, the CMI loss-difference bound in Theorem 4.4 could be improved by considering single-loss CMI metrics:

Theorem D.2. *Assume that $\ell(\cdot,\cdot)\in[0,1]$, then*

$$|\overline{\text{gen}}| \leq \frac{2}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\mathbb{E}_{\widetilde{Z}}\sqrt{2I^{\widetilde{Z}}(L_u^{\Phi_u^+};S_{u_1})} \leq \frac{2}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\sqrt{2I(L_u^{\Phi_u^+};S_{u_1}|\widetilde{Z})}.$$

Proof. From (29), we know that

$$|\overline{\text{gen}}| \leq \frac{2}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\left|\mathbb{E}_{S_{u_1},L_u,\Phi_u}\left[(-1)^{S_{u_1}}L_u^{\Phi_u^+}\right]\right| \leq \frac{2}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\mathbb{E}_{\widetilde{Z}}\left|\mathbb{E}_{S_{u_1},L_u,\Phi_u|\widetilde{Z}}\left[(-1)^{S_{u_1}}L_u^{\Phi_u^+}\right]\right|.$$

The result can then be obtained by following the same development with the proof of Theorem 4.4. □

D.4. Generalization Bounds with Wasserstein Distance

Inspired by the work of (Rodríguez Gálvez et al., 2021), we further present generalization bounds based on Wasserstein distances. These metrics are shown to be tighter than their information-theoretic counterparts (Wang & Mao, 2022), when the utilized distance metric c is discrete in Definition A.4.

Theorem D.3. *Assume that $\ell(w,\cdot)$ is β-Lipschitz w.r.t $w\in\mathcal{W}$, then for any $k\in\left[1,\frac{n}{m}\right]$,*

$$|\overline{\text{gen}}| \leq \frac{\beta}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\mathbb{E}_{Z_u}\left[\mathbb{W}(P_{W|Z_u},P_W)\right].$$

Proof. Given i.i.d samples $Z'_{1:m}\sim\mu^m$ and any $k\in\left[1,\frac{n}{m}\right]$, we have the following samplewise decomposition of the expected generalization error:

$$\begin{aligned}
\overline{\text{gen}} &= \frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\left(\mathbb{E}_{W,Z'_{1:m}}[\ell(W,Z'_{1:m})]-\mathbb{E}_{W,Z_u}[\ell(W,Z_u)]\right)\\
&= \frac{1}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\frac{1}{|\mathsf{P}_{km}^m|}\sum_{v\in\mathsf{P}_{km}^m}\left(\mathbb{E}_{W,Z'_{1:m}}[\ell(W,Z'_{1:m})]-\mathbb{E}_{W,(Z_u)_v}[\ell(W,(Z_u)_v)]\right)\\
&= \frac{1}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\Bigg(\mathbb{E}_{W,Z'_{1:m}}[\ell(W,Z'_{1:m})]\\
&\quad\underbrace{-\frac{1}{k|\mathsf{P}_{km}^m|}\sum_{v\in\mathsf{P}_{km}^m}\mathbb{E}_{W,(Z_u)_v}[\ell(W,(Z_u)_v)]-\cdots-\frac{1}{k|\mathsf{P}_{km}^m|}\sum_{v\in\mathsf{P}_{km}^m}\mathbb{E}_{W,(Z_u)_v}[\ell(W,(Z_u)_v)]}_{\times k}\Bigg)\\
&= \frac{1}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\frac{1}{|\mathsf{P}_{km}^{km}|}\sum_{v\in\mathsf{P}_{km}^{km}}\Bigg(\mathbb{E}_{W,Z'_{1:m}}[\ell(W,Z'_{1:m})]\\
&\quad-\frac{1}{k}\mathbb{E}_{W,Z_u}\left[\ell(W,((Z_u)_v)_{1:m})+\cdots+\ell(W,((Z_u)_v)_{(k-1)m+1:km})\right]\Bigg).
\end{aligned}\tag{30}$$

Recall that $\ell(w,\cdot)$ is β-Lipschitz, then $f(w,\cdot)=\frac{1}{k\beta}\underbrace{(\ell(w,\cdot)+\cdots+\ell(w,\cdot))}_{\times k}$ is 1-Lipschitz. For any $u\in\mathsf{C}_n^{km}$ and $v\in\mathsf{P}_{km}^{km}$, by applying Lemma A.8 with $P=P_W$ and $Q=P_{W|Z_u}$, we have

$$\mathbb{E}_{Z_u}[\mathbb{W}(P_{W|Z_u},P_W)]\geq\frac{1}{\beta}\Bigg(\mathbb{E}_{W,Z'_{1:m}}[\ell(W,Z'_{1:m})]$$
$$-\frac{1}{k}\mathbb{E}_{W,Z_u}\big[\ell(W,((Z_u)_v)_{1:m})+\cdots+\ell(W,((Z_u)_v)_{(k-1)m+1:km})\big]\Bigg).$$

Plugging this inequality into (30), we then get

$$\begin{aligned}\overline{\text{gen}}&\leq\frac{\beta}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\frac{1}{|\mathsf{P}_{km}^{km}|}\sum_{v\in\mathsf{P}_{km}^{km}}\mathbb{E}_{Z_u}[\mathbb{W}(P_{W|Z_u},P_W)]\\&=\frac{\beta}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\mathbb{E}_{Z_u}[\mathbb{W}(P_{W|Z_u},P_W)].\end{aligned}\tag{31}$$

Similarly, we can prove that $-\overline{\text{gen}}\leq\frac{\beta}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\mathbb{E}_{Z_u}[\mathbb{W}(P_{W|Z_u},P_W)]$, which together with (31) complete the proof. □

Theorem D.4. *Assume that $\ell(w,\cdot)$ is β-Lipschitz w.r.t $w\in\mathcal{W}$, then for any $k\in\big[1,\frac{n}{m}\big]$,*

$$|\overline{\text{gen}}|\leq\frac{2\beta}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\mathbb{E}_{\widetilde{Z},S_u}\Big[\mathbb{W}\Big(P_{W|\widetilde{Z},S_u},P_{W|\widetilde{Z}}\Big)\Big].$$

Proof. Let S' be an independent copy of S such that $S'\perp\!\!\!\perp W|\widetilde{Z}$, we have that $\mathbb{E}_{S'}[L_u^{\overline{S}'_u}-L_u^{S'_u}]=0$ and

$$\begin{aligned}\overline{\text{gen}}&=\mathbb{E}_{W,\widetilde{Z},S}\Big[L_{\widetilde{Z}_{\overline{S}}}(W)-L_{\widetilde{Z}_S}(W)\Big]=\mathbb{E}_{\widetilde{Z},S}\mathbb{E}_{W|\widetilde{Z},S}\Big[L_{\widetilde{Z}_{\overline{S}}}(W)-L_{\widetilde{Z}_S}(W)\Big]\\&=\frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\mathbb{E}_{\widetilde{Z},S_u,S'_u}\Big[\mathbb{E}_{W|\widetilde{Z},S_u}\Big[L_u^{\overline{S}_u}-L_u^{S_u}\Big]-\mathbb{E}_{W|\widetilde{Z}}\Big[L_u^{\overline{S}'_u}-L_u^{S'_u}\Big]\Big].\end{aligned}\tag{32}$$

Recall that $\ell(w,\cdot)$ is β-Lipschitz, then $f(w)=L_u^{\overline{S}_u}-L_u^{S_u}$ is 2β-Lipschitz. Following similar reduction steps as the proof of Theorem D.3, we can get

$$\overline{\text{gen}}\leq\frac{2\beta}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\mathbb{E}_{\widetilde{Z},S_u}\Big[\mathbb{W}\Big(P_{W|\widetilde{Z},S_u},P_{W|\widetilde{Z}}\Big)\Big].$$

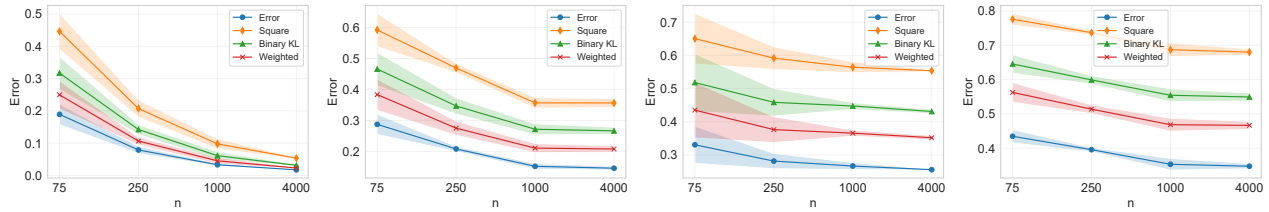
We can similarly prove that $-\overline{\text{gen}}\leq\frac{2\beta}{|\mathsf{C}_n^{km}|}\sum_{u\in\mathsf{C}_n^{km}}\mathbb{E}_{\widetilde{Z},S_u}\Big[\mathbb{W}\Big(P_{W|\widetilde{Z},S_u},P_{W|\widetilde{Z}}\Big)\Big]$, which finishes the proof. □

Theorem D.5. *Assume that $\ell(\cdot,\cdot)\in[0,1]$, then*

$$|\overline{\text{gen}}|\leq\frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\mathbb{E}_{S_{u_1}}\Big[\mathbb{W}\Big(P_{\Delta_u^{\Phi_u}|S_{u_1}},P_{\Delta_u^{\Phi_u}}\Big)\Big].$$

Proof. Let S' be an independent copy of S, then $\mathbb{E}_{S'_{u_1},L_u,\Phi_u}[(-1)^{S'_{u_1}}\Delta_u^{\Phi_u}]=0$. By the definition of the expected generalization error,

$$\begin{aligned}\overline{\text{gen}}&=\mathbb{E}_{W,\widetilde{Z},S}\Big[L_{\widetilde{Z}_{\overline{S}}}(W)-L_{\widetilde{Z}_S}(W)\Big]=\frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\mathbb{E}_{S_u,L_u,\Phi_u}\big[(-1)^{S_{u_1}}\Delta_u^{\Phi_u}\big]\\&=\frac{1}{|\mathsf{P}_n^m|}\sum_{u\in\mathsf{P}_n^m}\mathbb{E}_{S_{u_1},S'_{u_1}}\Big[\mathbb{E}_{L_u,\Phi_u|S_{u_1}}\big[(-1)^{S_{u_1}}\Delta_u^{\Phi_u}\big]-\mathbb{E}_{L_u,\Phi_u}\Big[(-1)^{S'_{u_1}}\Delta_u^{\Phi_u}\Big]\Big].\end{aligned}\tag{33}$$



(a) $\delta = 0$ (b) $\delta = 0.05$ (c) $\delta = 0.1$ (d) $\delta = 0.15$

Figure 6: Comparison of the generalization gap and the upper bounds for the binary MNIST classification task with different levels of label noise, where the labels are randomly flipped with probability δ.

Let $f(\Delta_u^{\Phi_u}) = (-1)^{S_{u_1}} \Delta_u^{\Phi_u}$, then it is 1-Lipschitz since $|f(\Delta_u^{\Phi_u})| = |\Delta_u^{\Phi_u}|$. Then by applying Lemma A.8 with $P = P_{\Delta_u^{\Phi_u}|S_{u_1}}$ and $Q = P_{\Delta_u^{\Phi_u}}$, we have

$$\mathbb{W}\left(P_{\Delta_u^{\Phi_u}|S_{u_1}}, P_{\Delta_u^{\Phi_u}}\right) \geq \mathbb{E}_{L_u,\Phi_u|S_{u_1}}\left[(-1)^{S_{u_1}} \Delta_u^{\Phi_u}\right] - \mathbb{E}_{L_u,\Phi_u}\left[(-1)^{S'_{u_1}} \Delta_u^{\Phi_u}\right].$$

Plugging this inequality into (33), we obtain

$$\overline{\text{gen}} \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{S_{u_1}}\left[\mathbb{W}\left(P_{\Delta_u^{\Phi_u}|S_{u_1}}, P_{\Delta_u^{\Phi_u}}\right)\right].$$

The proof is complete by similarly proving that $-\overline{\text{gen}} \leq \frac{1}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{S_{u_1}}\left[\mathbb{W}\left(P_{\Delta_u^{\Phi_u}|S_{u_1}}, P_{\Delta_u^{\Phi_u}}\right)\right]$. □

Theorem D.6. *Assume that* $\ell(\cdot,\cdot) \in [0,1]$, *then*

$$|\overline{\text{gen}}| \leq \frac{2}{|\mathsf{P}_n^m|} \sum_{u \in \mathsf{P}_n^m} \mathbb{E}_{S_{u_1}}\left[\mathbb{W}\left(P_{L_u^{\Phi_u^+}|S_{u_1}}, P_{L_u^{\Phi_u^+}}\right)\right].$$

Proof. The result can be obtained by following the same development with the proof of Theorem D.5. □

E. Experiment Details and Additional Results

In this section, we present experiment details and additional experimental results that were not included in the main text due to space limitations. The deep learning models are trained with an Intel Xeon CPU (2.10GHz, 48 cores), 256GB memory, and 4 Nvidia Tesla V100 GPUs (32GB).

E.1. Synthetic Experiments

We follow the experimental settings outlined in (Wang & Mao, 2023) to generate synthetic Gaussian datasets using the scikit-learn Python package. Our objective is to train a 5-class classification network, handling 5-dimensional input data points. The class centers for this dataset are randomly allocated from the vertices of a five-dimensional hypercube. Data points are then independently drawn from isotropic Gaussian distributions with a standard deviation of 0.25. For the model, we opt for a simple 4-layer MLP network, employing ReLU as the activation function. The selection of the loss function is contingent on the value of m: for $m = 1$, we utilized the binary 0-1 loss to quantify the generalization gap; for $m > 1$, we implemented a binarized version of the corresponding contrastive losses. Specifically, with a predictive function $f : \mathcal{X}^m \to \mathbb{R}$, the losses are computed based on a given threshold θ, exemplified in the pairwise contrastive loss as follows:

$$L_{ij} = \mathbb{1}_{f(X_i,X_j) \geq \theta} \oplus \mathbb{1}_{Y_i = Y_j}.$$

Here, the threshold θ was adaptively selected to balance precision and recall scores. To ensure statistical robustness, we executed 200 independent trials for each experimental configuration to accurately estimate the mutual information metrics. In our main text, the comparative analysis for the variance-based bound (Theorem 4.7) was omitted due to its minimal

m \ n	20	40	60	80	100
1	0.09701 / 0.09679	0.05337 / 0.05333	0.04662 / 0.04657	0.03422 / 0.03418	0.03046 / 0.03045
2	0.12478 / 0.12452	0.08405 / 0.08380	0.05727 / 0.05712	0.05027 / 0.05010	0.03686 / 0.03681
3	0.15337 / 0.15176	0.09442 / 0.09385	0.07873 / 0.07832	0.05999 / 0.05973	0.04478 / 0.04455
4	0.18916 / 0.18782	0.10948 / 0.10895	0.08510 / 0.08475	0.06653 / 0.06622	0.05102 / 0.05085

Table 1: Comparison between the fast-rate bound (Theorem 4.6, left) and the variance-based bound (Theorem 4.7, right) on synthetic Gaussian datasets using a simple MLP network.

difference at the current graphical resolution in Figure 4. Below, we include a comprehensive comparison between the fast-rate bound (Theorem 4.6) and the variance-based bound (Theorem 4.7) for completeness:

As can be seen in Table 1, the variance-based bound (Theorem 4.7) is consistently tighter than the fast-rate bound (Theorem 4.6) by taking $\gamma = 0.9$. This verifies the advantages of the loss variance when the training risk is close but not equal to zero.

E.2. Real-world Experiments

Following the experiment settings in (Harutyunyan et al., 2021; Hellström & Durisi, 2022b), we conduct 4 distinct real-world learning scenarios to evaluate the generalization bounds presented in this paper: 1) MNIST 4 vs 9 classification using Adam, 2) MNIST 4 vs 9 classification using SGLD, 3) CIFAR-10 classification with fine-tuned ResNet-50. We additionally consider pretraining on Flickr30k with a fine-tuned CLIP (ViT-B/32) model to examine the scalability of our bounds.

For each learning task, we sampled k_1 instances of $\widetilde{Z}$, involving the random selection of $2n$ samples from the respective datasets. Additionally, for each $\widetilde{Z}$, we drew k_2 samples of the supersample variables S, culminating in $k_1 \times k_2$ independent runs in total, with k_1 and k_2 values aligned with those in (Harutyunyan et al., 2021). Notably, in the CLIP model, the empirical and population risks are represented as a combination of pointwise and pairwise risks. Let $\mathcal{I}$ and $\mathcal{T}$ represent the spaces of images and texts respectively, then sample Z_i consists of an image-text pair (I_i, T_i). Let $f : \mathcal{I} \times \mathcal{T} \mapsto \mathbb{R}$ be the predictive function parameterized by the CLIP model, the self-supervised contrastive learning loss is then given as:

$$L_{ij} = \begin{cases} \mathbb{1}_{f(I_i,T_i)\leq\theta}, & \text{if } i = j, \\ \mathbb{1}_{f(I_i,T_j)\geq\theta}, & \text{if } i \neq j. \end{cases}$$

An upper bound for this mixed risk can then be attained by amalgamating bounds for both $m = 1$ and $m = 2$. The threshold θ is dynamically chosen to balance pointwise and pairwise risks. It's important to acknowledge that pretraining generalization performance may not directly correlate with downstream task efficacy, particularly under self-supervised learning paradigms where false negatives exist in ground truth labels. Consequently, an increase in generalization error with larger n is a reasonable observation, as demonstrated in Figure 5.

Furthermore, to examine scenarios with significant overfitting, we introduced random label noise into the binary MNIST dataset. Specifically, the labels were randomly flipped with a specified probability δ. As evidenced in Figure 6, the generalization bounds evaluated in this study consistently provided non-vacuous estimates of the generalization error. Among these, the fast-rate bound (Theorem 4.6) consistently emerged as the most stringent in these comparisons.