A New Family of Generalization Bounds Using Samplewise Evaluated CMI

Fredrik Hellström
Chalmers University of Technology
Gothenburg, Sweden
frehells@chalmers.se

Giuseppe Durisi
Chalmers University of Technology
Gothenburg, Sweden
durisi@chalmers.se

Abstract

We present a new family of information-theoretic generalization bounds, in which the training loss and the population loss are compared through a jointly convex function. This function is upper-bounded in terms of the disintegrated, samplewise, evaluated conditional mutual information (CMI), an information measure that depends on the losses incurred by the selected hypothesis, rather than on the hypothesis itself, as is common in probably approximately correct (PAC)-Bayesian results. We demonstrate the generality of this framework by recovering and extending previously known information-theoretic bounds. Furthermore, using the evaluated CMI, we derive a samplewise, average version of Seeger's PAC-Bayesian bound, where the convex function is the binary KL divergence. In some scenarios, this novel bound results in a tighter characterization of the population loss of deep neural networks than previous bounds. Finally, we derive high-probability versions of some of these average bounds. We demonstrate the unifying nature of the evaluated CMI bounds by using them to recover average and high-probability generalization bounds for multiclass classification with finite Natarajan dimension.

1 Introduction

Information-theoretic generalization bounds, i.e., generalization bounds that are expressed in terms of information-theoretic metrics such as the Kullback-Leibler (KL) divergence, mutual information, and conditional mutual information (CMI), have emerged as useful tools to obtain an accurate characterization of the performance of deep neural networks. To obtain such bounds, one compares a *posterior*, that is, the distribution over the hypotheses induced by the learning algorithm, to a reference distribution called the *prior*, an approach first introduced to provide probably approximately correct (PAC) bounds for Bayesian classifiers [1, 2, 3, 4]. The connection between these results and classic information-theoretic metrics was clarified in [5, 6], where the generalization gap, averaged with respect to the joint posterior and data distribution, is bounded in terms of the mutual information between the training data and the hypothesis. These bounds have since been extended in several ways [7, 8, 9, 10, 11, 12, 13, 14, 15].

A major step was taken in [16], in which a setting where the training set is randomly selected from a larger supersample is considered. We refer to this setup as the CMI setting, as it leads to generalization bounds in terms of the CMI between the hypothesis and training set selection, given the supersample. It turns out that these bounds can be further tightened by observing that the selected hypothesis enters the derivation only through the loss that it induces on the supersample. Using this observation, [16] also derives bounds in terms of the information encoded in these losses rather than in the hypothesis itself, a quantity referred to as the *evaluated CMI* (e-CMI). Due to the data-processing inequality, these bounds are always tighter than the regular CMI bounds. This observation was recently further

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

exploited by [17], which derived samplewise versions of these bounds.[1] Intriguingly, these bounds are both easier to evaluate than their hypothesis-based counterparts, due to the lower dimensionality of the random variables involved, and quantitatively tighter for deep neural networks. In particular, while bounds involving information measures based on the hypothesis space tend to increase as training progresses [11, 18], the bound of [17] remains stable.

The results that have so far been derived for the CMI setting pertain only to the (weighted) difference between population loss and training loss, or to its squared or absolute value [8, 9, 14, 16, 17, 18]. In the PAC-Bayesian literature, other types of discrepancy measures have been considered, and shown to result in tighter bounds on the population loss. For example, [19, 20, 21, 22] consider the binary KL divergence (i.e., the KL divergence between two Bernoulli distributions) with parameters given by the training and population loss, respectively, while [23, 24] consider arbitrary jointly convex functions. Finally, [25] allows for arbitrary functions. It should be noted that in all of these results, a moment-generating function that depends on the selected function has to be controlled for the bound to be computable.

Recently, the e-CMI framework was proven to be expressive enough to allow one to rederive known results in learning theory, e.g., generalization bounds expressed in terms of algorithmic stability, VC dimension, and related complexity measures [17, 26]. Tightening and extending e-CMI bounds, which is the main objective of this paper, has the potential to further increase the unifying nature of the e-CMI framework.

Contributions Leveraging a basic inequality involving a generic convex function of two random variables (Lemma 1), we establish several novel disintegrated, samplewise, e-CMI bounds on the average generalization error. Specifically, we present i) a square-root bound (Theorem 1) on the generalization error, together with a mean-squared error extension, which tightens the bound recently reported in [17]; ii) a linear bound (Theorem 3) that tightens the bound given in [18]; iii) a binary KL bound (Theorem 4), which is a natural extension to the e-CMI setting of a well-known bound in the PAC-Bayes literature [22]. While the derivation of the first two bounds involves an adaptation of results available in the literature, to obtain the binary KL bound we need a novel concentration inequality involving independent but not identically distributed random variables (Lemma 2). As an additional contribution, we show how to adapt the techniques presented in the paper to obtain high-probability (rather than average) e-CMI bounds (Theorem 7). Furthermore, we illustrate the expressiveness of the e-CMI framework by using our bounds to recover average and high-probability generalization bounds for multiclass classification with finite Natarajan dimension (Theorem 8). Finally, we conduct numerical experiments on MNIST and CIFAR10, which reveal that the binary KL bound results in a tighter characterization of the population loss compared to the square-root bound and linear bound for some deep learning scenarios.

Preliminaries and Notation

We let $D(P \,||\, Q)$ denote the KL divergence between the two probability measures P and Q. This quantity is well-defined if P is absolutely continuous with respect to Q. When P and Q are Bernoulli distributions with parameters p and q respectively, we let $D(P \,||\, Q) = d(p \,||\, q) = p\log(p/q) + (1-p)\log((1-p)/(1-q))$, and we refer to $d(p \,||\, q)$ as the binary KL divergence. For two random variables X and Y with joint distribution P_{XY} and respective marginals P_X and P_Y, we let $I(X;Y) = D(P_{XY} \,||\, P_X P_Y)$ denote their mutual information. Throughout the paper, we use uppercase letters to denote random variables and lowercase letters to denote their realizations. We denote the conditional joint distribution of X and Y given an instance $Z = z$ by $P_{XY|Z=z}$ and the corresponding conditional distribution for the product of marginals by $P_{X|Z=z}P_{Y|Z=z}$. Furthermore, we let $I^z(X;Y) = D(P_{XY|Z=z} \,||\, P_{X|Z=z}P_{Y|Z=z})$, which is referred to as the *disintegrated* mutual information [11]. Its expectation is the conditional mutual information $I(X;Y|Z) = \mathbb{E}_Z\big[I^Z(X;Y)\big]$.

CMI Setting Let $\mathcal{Z}$ be the sample space and let $\mathcal{D}$ denote the data-generating distribution. Consider a supersample $\widetilde{Z} \in \mathcal{Z}^{n\times 2}$, where each entry is generated independently from $\mathcal{D}$. For convenience, we index the columns of $\widetilde{Z}$ starting from 0 and the rows starting from 1. Furthermore, we denote the ith

[1] While [17] considers the information stored in predictions rather than losses, referred to as the f-CMI, the derivations therein can be adapted to obtain bounds that depend on the losses, as we clarify in Section 2.2.

row of $\widetilde{Z}$ as $\widetilde{Z}_i$. Let $S = (S_1, \ldots, S_n)$ denote a membership vector, with entries generated independently according to a $\mathrm{Bern}(1/2)$ distribution, independently of $\widetilde{Z}$. Let $\bar{S} = (1 - S_1, \ldots, 1 - S_n)$ denote the modulo-2 complement of S. We refer to S as a membership vector because it is used to divide the supersample into an n-dimensional training vector $\widetilde{Z}_S$ with entries $[\widetilde{Z}_S]_i = \widetilde{Z}_{i,S_i}$ and an n-dimensional test vector $\widetilde{Z}_{\bar{S}}$ with entries $[\widetilde{Z}_{\bar{S}}]_i = \widetilde{Z}_{i,\bar{S}_i}$. To be able to handle arbitrary learning settings, we consider learning algorithms as maps $\mathcal{A} : \mathcal{Z}^n \times \mathcal{R} \to \mathcal{F}$, where $R \in \mathcal{R}$ is a random variable (independent of $\widetilde{Z}$ and S) that captures the stochasticity of the algorithm and $\mathcal{F}$ is a space, for instance, a parameter space or function space. For a fixed $R = r$ and $\widetilde{Z}_S = \tilde{z}_s$, $\mathcal{A}(\tilde{z}_s, r)$ is a deterministic function of $\tilde{z}_s$. The quality of the learning algorithm's output is evaluated using a bounded loss function $\ell : \mathcal{F} \times \mathcal{Z} \to [0, 1]$. We let U be a vector of size m, the elements of which are sampled without replacement uniformly at random from $(1, \ldots, n)$. For a given realization $U = u = (u_1, \ldots, u_m)$, we let $\tilde{z}_u$ denote the $m \times 2$ matrix obtained by stacking the vectors $\tilde{z}_{u_i}$ for $i = 1, \ldots, m$. Similarly, $\ell(\mathcal{A}(\tilde{z}_s, R), \tilde{z}_u)$ denotes the $m \times 2$ matrix of losses obtained by applying $\ell(\mathcal{A}(\tilde{z}_s, R), \cdot)$ elementwise to $\tilde{z}_u$. We denote the population loss as $L_{\mathcal{D}}(\mathcal{A}, \tilde{z}_s, r) = \mathbb{E}_{Z'}[\ell(\mathcal{A}(\tilde{z}_s, r), Z')]$, where $Z' \sim \mathcal{D}$. The training loss is $L_{\tilde{z}_s}(\mathcal{A}, \tilde{z}_s, r) = \frac{1}{n}\sum_{i=1}^{n} \ell(\mathcal{A}(\tilde{z}_s, r), [\tilde{z}_s]_i)$. In general, for any p-dimensional vector of data points $\hat{z}$, we let $L_{\hat{z}}(\mathcal{A}, \tilde{z}_s, r) = \frac{1}{p}\sum_{i=1}^{p} \ell(\mathcal{A}(\tilde{z}_s, r), \hat{z}_i)$. Since $\mathcal{A}(\tilde{z}_s, r)$ does not depend on $\tilde{z}_{\bar{s}}$, $L_{\tilde{z}_{\bar{s}}}(\mathcal{A}, \tilde{z}_s, r)$ is a test loss. Finally, we let $L_{\mathcal{D}} = \mathbb{E}_{\widetilde{Z},S,R}[L_{\mathcal{D}}(\mathcal{A}, \widetilde{Z}_S, R)]$ denote the average population loss and $\hat{L} = \mathbb{E}_{\widetilde{Z},S,R}[L_{\widetilde{Z}_S}(\mathcal{A}, \widetilde{Z}_S, R)]$ denote the average training loss.

2 Average Generalization Bounds

In this section, we present the main results of this paper: a new family of disintegrated samplewise e-CMI bounds for the average population loss. High-probability versions of these bounds are given in Section 3.

2.1 Main Lemma

We now present the generic inequality upon which the bounds in this section are based. This inequality, which is similar in nature to the ones provided in [23] and [25], gives us a generic framework to derive generalization bounds, as it allows for a wide choice of functions for comparing training loss and population loss. A crucial difference compared to [23] and [25] is that our focus in this section is on average rather than PAC-Bayesian generalization bounds, and on bounds based on the disintegrated, samplewise e-CMI, rather than traditional KL-based bounds.

Lemma 1. *Let $f_\gamma : [0,1]^2 \to \mathbb{R}$ be a function that is jointly convex in its arguments and is parameterized by γ. Let X and Y be two random variables, and let Y' be a random variable with the same marginal distribution as Y such that Y' and X are independent. Assume that the joint distribution of X, Y is absolutely continuous with respect to the joint distribution of X, Y'. Let $g_1(X, Y)$ and $g_2(X, Y)$ be measurable functions with range $[0, 1]$ and finite first moments, such that, for all γ, $\mathbb{E}_{X,Y}[f_\gamma(g_1(X, Y), g_2(X, Y))]$ is finite. Let*

$$\xi_\gamma = \log \mathbb{E}_{X,Y'}\left[e^{f_\gamma\left(g_1(X,Y'), g_2(X,Y')\right)}\right] \tag{1}$$

and assume that ξ_γ is finite. Then,

$$\begin{aligned}\sup_\gamma f_\gamma(\mathbb{E}_{X,Y}[g_1(X, Y)], \mathbb{E}_{X,Y}[g_2(X, Y)]) - \xi_\gamma &\leq \sup_\gamma \mathbb{E}_{X,Y}[f_\gamma(g_1(X, Y), g_2(X, Y))] - \xi_\gamma \\ &\leq I(X; Y). \end{aligned} \tag{2}$$

The proof, which is an application of Jensen's inequality and Donsker-Varadhan's variational representation of the KL divergence, is given in Appendix A, along with the proofs of all other results provided in this paper.

Note that we are allowed to optimize γ because, in this section, we only consider average bounds. In contrast, when we derive high-probability bounds in Section 3, we need to fix γ. Optimizing γ over a set of candidate values in the high-probability scenario incurs a union bound cost [27, Sec. 1.2.2].

To apply Lemma 1, we need to identify functions $f_\gamma(\cdot, \cdot)$ for which the moment generating function in (1) can be controlled. A theme throughout this section is that we identify functions $f_\gamma(\cdot, \cdot)$ for

which there exist concentration results that imply that $\xi_\gamma \leq 0$. This allows us to loosen the bound in (2) by discarding ξ_γ.

Throughout this section, we assume that the loss is bounded, so that $\ell(\cdot,\cdot) \in [0,1]$. Note that the reported results can be extended to more general losses by use of scaling. In particular, assume that there is a function $K : \mathcal{F} \to \mathbb{R}_+$ such that, for all $f \in \mathcal{F}$, $\sup_z \ell(f,z) \leq K(f)$ (referred to as the hypothesis-dependent range condition in [28]). Then, all of the results that we present for bounded losses also hold for the scaled loss $\ell(f,z)/K(f)$.

2.2 Extending (f)-CMI Bounds to e-CMI

We now apply Lemma 1 to recover the average bounds of [17, 18]. These works derive bounds using the information contained either in the hypothesis itself or the resulting predictions, i.e., the CMI or the f-CMI. We instead derive bounds in terms of the information captured by the matrix of losses, i.e., the e-CMI. For parametric supervised learning algorithms, we can recover the original bounds via the data-processing inequality. This is formalized in the following remark.

Remark 1. *Consider a parametric supervised learning setting, where $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $\mathcal{A}(\tilde{z}_S, R) = W \in \mathcal{F}$ are the parameters of a function $\phi_W : \mathcal{X} \to \mathcal{Y}$. Let $\tilde{x}_u$ denote the $m \times 2$ matrix obtained by projecting each element of $\tilde{z}_u$ onto $\mathcal{X}$, i.e., the matrix of unlabeled examples. Let $\phi_W(\tilde{x}_u)$ denote the matrix of predictions obtained by elementwise application of ϕ_W to $\tilde{x}_u$. Then, for any fixed $\tilde{z}$ and u,*

$$I^{\tilde{z},u}(\ell(W,\tilde{z}_u); S_u) \leq I^{\tilde{z},u}(\phi_W(\tilde{x}_u); S_u) \leq I^{\tilde{z},u}(W; S_u). \tag{3}$$

We now use Lemma 1 to recover some of the results reported in [17]. Let $\Delta = L_{\tilde{z}_{S_i'}}(\mathcal{A}, \tilde{z}_s, R) - L_{\tilde{z}_{S_i'}}(\mathcal{A}, \tilde{z}_s, R)$, where S_i' is an independent copy of S_i. For a fixed $\tilde{z}$, symmetry implies that $\mathbb{E}_{S_i'}[\Delta] = 0$. Furthermore, since $\ell(\cdot,\cdot) \in [0,1]$, we have that $\Delta \in [-1,1]$. Thus, Δ is 1-sub-Gaussian. By using properties of sub-Gaussian random variables, we can control ξ_γ in Lemma 1 when choosing $f_\gamma(\cdot,\cdot)$ as $\gamma\Delta$ or $\gamma\Delta^2$. The resulting bounds are given in the following theorem.

Theorem 1 (Square-root bound and squared bound)**.** *Consider the CMI setting. Then,*

$$\left|L_{\mathcal{D}} - \hat{L}\right| \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\widetilde{Z}}\left[\sqrt{2I^{\widetilde{Z}}(\ell(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z}_i); S_i)}\right] \tag{4}$$

$$\leq \frac{1}{n}\sum_{i=1}^{n} \sqrt{2I(\ell(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z}_i); S_i|\widetilde{Z})}. \tag{5}$$

Furthermore,

$$\mathbb{E}_{\widetilde{Z},R,S}\left[\left(L_{\mathcal{D}}(\mathcal{A}, \widetilde{Z}_S, R) - L_{\widetilde{Z}_S}(\mathcal{A}, \widetilde{Z}_S, R)\right)^2\right] \leq \frac{8}{m}\Big(I(\ell(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z}_U); S_U|\widetilde{Z}, U) + 2\Big). \tag{6}$$

In (4), the expectation over $\widetilde{Z}$ is taken outside of the square root, and the generalization bound is in terms of the disintegrated mutual information. By Jensen's inequality, this is tighter than (5).

As shown in Appendix A, one can further generalize (4) to obtain an upper bound of the form

$$\mathbb{E}_{\widetilde{Z},U}\left[\sqrt{2I^{\widetilde{Z},U}(\ell(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z}_U); S_U)}\right]. \tag{7}$$

However, since (7) increases with the size m of the random subset U [17, Prop. 1], we focus only on the case $m = 1$, as this leads to the tightest bound. The same holds for (5), as well as for all of the bounds that we present in the remainder of this section, where we also focus only on $m = 1$. In contrast, the choice of $m = 1$ in the squared bound in (6) is suboptimal and actually yields a vacuous bound due to the $2/m$ term.

It is possible to obtain a bound similar to (4), but where an expectation over R is taken outside the square root and the mutual information is also conditioned on R. We present this bound in the following theorem.

Theorem 2 (R-conditioned square-root bound)**.** *Consider the CMI setting. Then,*

$$\left|L_{\mathcal{D}} - \hat{L}\right| \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\widetilde{Z},R}\left[\sqrt{2I^{\widetilde{Z},R}(\ell(\mathcal{A}(\widetilde{Z}_S,R),\widetilde{Z}_i);S_i)}\right] \tag{8}$$

$$\leq \frac{1}{n}\sum_{i=1}^{n} \sqrt{2I(\ell(\mathcal{A}(\widetilde{Z}_S,R),\widetilde{Z}_i);S_i|\widetilde{Z},R)}. \tag{9}$$

A similar bound, but given in terms of mutual information rather than e-CMI, is reported in [11, Thm. 2.4]. The randomness of the learning algorithm can reduce the mutual information between its output and the selection variable S_i. Specifically, one can show that

$$I^{\tilde{z}}(\ell(\mathcal{A}(\tilde{z}_S,R),\tilde{z}_i);S_i) \leq I^{\tilde{z}}(\ell(\mathcal{A}(\tilde{z}_S,R),\tilde{z}_i);S_i|R) \tag{10}$$

which implies that the square-root bound in (5) is tighter than the R-conditioned square-root bound in (9). However, the ordering between the disintegrated square-root bound in (4) and the disintegrated R-conditioned square-root bound in (8) is unclear.

Next, we generalize the linear bounds of [18], which are samplewise extensions of the bounds in [16], to the e-CMI framework. We consider two scenarios. For the first one, we assume that $\gamma = (\gamma_1, \gamma_2)$ are positive constants that satisfy a certain constraint, and let $f_\gamma(\hat{L}, L_{\mathcal{D}}) = \gamma_1 n(L_{\mathcal{D}} - \gamma_2\hat{L})$. For the second one, we assume that $\hat{L} = 0$, the so-called *interpolating* setting, and let $f_\gamma(\hat{L}, L_{\mathcal{D}}) = n\log(2)L_{\mathcal{D}}$. For both of these scenarios, it can be shown that $\xi_\gamma \leq 0$, yielding the following.

Theorem 3 (Linear bound and interpolation bound)**.** *Consider the CMI setting. Let $\Gamma \subset \mathbb{R}_+^2$ denote the set of parameters $\gamma = (\gamma_1, \gamma_2)$ that satisfy $\gamma_1(1-\gamma_2) + (e^{\gamma_1} - 1 - \gamma_1)(1+\gamma_2^2) \leq 0$. Then,*

$$L_{\mathcal{D}} \leq \min_{\gamma\in\Gamma} \gamma_2\hat{L} + \sum_{i=1}^{n} \frac{I(\ell(\mathcal{A}(\widetilde{Z}_S,R),\widetilde{Z}_i);S_i|\widetilde{Z})}{\gamma_1 n}. \tag{11}$$

Furthermore, if $\hat{L} = 0$,

$$L_{\mathcal{D}} \leq \sum_{i=1}^{n} \frac{I(\ell(\mathcal{A}(\widetilde{Z}_S,R),\widetilde{Z}_i);S_i|\widetilde{Z})}{n\log(2)}. \tag{12}$$

The interpolation bound in (12) improves on the linear bound in (11) when $\hat{L} = 0$, as the constraint implies that $\gamma_1^2 - 4(e^{\gamma_1} - 1)(e^{\gamma_1} - 1 - \gamma_1) \geq 0$. This means that $\gamma_1 < 0.37$, whereas $\log 2 > 0.69$.

2.3 Binary KL Bound with Samplewise e-CMI

We now derive bounds in terms of the binary KL divergence between the training loss and the test loss. To this end, similar to [27, 29], we define

$$d_\gamma(q\,||\,p) = \gamma q - \log(1 - p + pe^\gamma). \tag{13}$$

An important property of this function is that

$$\sup_\gamma d_\gamma(q\,||\,p) = d(q\,||\,p). \tag{14}$$

Note that both $d_\gamma(\cdot\,||\,\cdot)$ and $d(\cdot\,||\,\cdot)$ are jointly convex in their arguments. For our next result, we need the following lemma.

Lemma 2. *For $i = 1, \ldots, n$, let $X_i \sim P_{X_i}$, $\mathbb{E}[X_i] = \mu_i$, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$, and $\bar{\mu} = \frac{1}{n}\sum_{i=1}^n \mu_i$. Assume that $X_i \in [0,1]$ almost surely and that all X_i are independent. Then, for every fixed $\gamma > 0$,*

$$\mathbb{E}[\exp(nd_\gamma(\hat{\mu}\,||\,\bar{\mu}))] \leq 1. \tag{15}$$

This inequality, which, to the best of our knowledge, has previously been reported only for identically distributed random variables, follows by combining [30, Lemma 1], [31, Thm. 3], and [29, Eq. (17)] (which is a generalization of [27, Lemma 1.1.1] from binary to bounded random variables).[2]

[2] A similar result, with $d(\cdot\,||\,\cdot)$ instead of $d_\gamma(\cdot\,||\,\cdot)$, is established in [32, Lemma 2].

Consider a fixed $\tilde{z}$. For each i, $\mathbb{E}_{S_i'}\big[L_{[\tilde{z}_{S'}]_i}(\mathcal{A}, \tilde{z}_s, R)\big] = (\ell(\mathcal{A}(\tilde{z}_s, R), \tilde{z}_{i,0}) + \ell(\mathcal{A}(\tilde{z}_s, R), \tilde{z}_{i,1}))/2$, where S_i' is an independent copy of S_i. Thus,

$$\mathbb{E}_{S'}\big[L_{\tilde{z}_{S'}}(\mathcal{A}, \tilde{z}_s, R)\big] = \frac{L_{\tilde{z}_s}(\mathcal{A}, \tilde{z}_s, R) + L_{\tilde{z}_{\bar{s}}}(\mathcal{A}, \tilde{z}_s, R)}{2}. \tag{16}$$

This observation allows us to bound the binary KL divergence between the training loss $\hat{L}$ and the arithmetic mean of the training and population loss, $(\hat{L} + L_{\mathcal{D}})/2$. Combining (15) and (2), with appropriate choices for the variables and functions therein, we obtain the following result.

Theorem 4 (Binary KL bound)**.** *Consider the CMI setting. Then,*

$$d\left(\hat{L} \,||\, \frac{L_{\mathcal{D}} + \hat{L}}{2}\right) \leq \frac{1}{n} \sum_{i=1}^{n} I(\ell(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z}_i); S_i | \widetilde{Z}) \tag{17}$$

which implies that $L_{\mathcal{D}}$ can be bounded as

$$L_{\mathcal{D}} \leq d^{-1}\left(\hat{L}, \frac{1}{n} \sum_{i=1}^{n} I(\ell(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z}_i); S_i | \widetilde{Z})\right) \tag{18}$$

where

$$d^{-1}(q, c) = \sup\left\{p \in [0, 1] : d\left(q \,||\, \frac{q+p}{2}\right) \leq c\right\}. \tag{19}$$

Similar to Theorem 1, we can obtain a disintegrated version of (18) where the expectation over $\widetilde{Z}$ is outside of the inversion of the binary KL divergence. We present this result in the following theorem.

Theorem 5 (Disintegrated binary KL bound)**.** *Consider the CMI setting. Then,*

$$L_{\mathcal{D}} \leq \mathbb{E}_{\widetilde{Z}}\left[d^{-1}\left(\mathbb{E}_{R,S}\left[L_{\widetilde{Z}_S}(\mathcal{A}, \widetilde{Z}_S, R)\right], \frac{1}{n} \sum_{i=1}^{n} I^{\widetilde{Z}}(\ell(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z}_i); S_i)\right)\right]. \tag{20}$$

By using Pinsker's inequality, which implies that $2(q-p)^2 \leq d(q \,||\, p)$, one can weaken (18) to obtain

$$\left|L_{\mathcal{D}} - \hat{L}\right| \leq \sqrt{\frac{2}{n} \sum_{i=1}^{n} I(\ell(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z}_i); S_i | \widetilde{Z})}. \tag{21}$$

However, since the average over i is inside the square root, this bound is weaker than (5) by Jensen's inequality.

To establish Theorem 4, it is crucial that the supremum over γ in Lemma 1 is outside the expectation that defines ξ_γ. Indeed, $\sup_\gamma d_\gamma(q \,||\, p) = d(q \,||\, p)$, and with the notation from Lemma 2, we have [22, Thm. 1]

$$\mathbb{E}[\exp(nd(\hat{\mu} \,||\, \bar{\mu}))] \geq \sqrt{n}. \tag{22}$$

Therefore, having the supremum inside of the expectation would unavoidably lead to an additional term greater than $\log(\sqrt{n})/n$ in the upper bound of (18).

The binary KL bound in (18) is an average, samplewise, e-CMI analogue of the PAC-Bayesian bound with a binary KL divergence on the left-hand side, sometimes referred to as Seeger's bound [4, 20]. However, as noted after Lemma 2, the concentration inequality needed to establish Theorem 4 is a refinement of the result used in [20]. Also, the dependence on $\hat{L}$ and $L_{\mathcal{D}}$ in the binary KL bound is nonstandard. As we show in Section 5, the binary KL bound in (18) is sometimes tighter than the square-root bound in Theorem 1 and the linear bound in Theorem 3.

The binary KL bounds in (18) and (20) can be tightened by considering affine transformations of the arguments of $d(\cdot \,||\, \cdot)$. However, this does not lead to improvements for the low training loss scenarios that we focus on in this paper. Therefore, we relegate these extensions to Appendix B, where we also present an analogue of Theorem 4 in terms of the samplewise mutual information between the learning algorithm's output and the training data, rather than the e-CMI.

While $d^{-1}(\cdot, \cdot)$ does not admit an analytical expression, $d^{-1}(0, \cdot)$ is tractable. This leads to the following simplified form of the disintegrated binary KL bound in (20) when $\hat{L} = 0$.

Theorem 6 (Disintegrated interpolation binary KL bound)**.** *Assume that $\hat{L} = 0$. Then,* (20) *becomes*

$$L_{\mathcal{D}} \leq \mathbb{E}_{\widetilde{Z}}\left[2 - 2e^{-\frac{1}{n}\sum_{i=1}^{n} I^{\widetilde{Z}}(\ell(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z}_i); S_i)}\right]. \tag{23}$$

3 High-Probability Bounds

The techniques presented in Section 2 can be adapted to allow for the derivation of high-probability bounds, in particular through the use of exponential inequalities [33, 34, 35]. To demonstrate this, we now present two such high-probability bounds.

Theorem 7 (High-probability square-root and binary KL bounds)**.** *Let $\lambda = \ell(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z})$. Furthermore, let $P_{\lambda|\widetilde{Z}S}$ denote the conditional distribution of λ given $\widetilde{Z}$ and S, and let $P_{\lambda|\widetilde{Z}}$ denote the conditional distribution of λ given $\widetilde{Z}$. Then, with probability at least $1-\delta$ over the draw of $\widetilde{Z}$ and S,*

$$\mathbb{E}_R\Big[L_{\widetilde{Z}_{\bar{S}}}(\mathcal{A}, \widetilde{Z}_S, R) - L_{\widetilde{Z}_S}(\mathcal{A}, \widetilde{Z}_S, R)\Big] \leq \sqrt{\frac{2}{n-1}\bigg(D(P_{\lambda|\widetilde{Z}S} \,||\, P_{\lambda|\widetilde{Z}}) + \log\frac{\sqrt{n}}{\delta}\bigg)}. \tag{24}$$

Furthermore, also with probability at least $1-\delta$ over the draw of $\widetilde{Z}$ and S,

$$d\bigg(\mathbb{E}_R\Big[L_{\widetilde{Z}_S}(\mathcal{A}, \widetilde{Z}_S, R)\Big] \,||\, \mathbb{E}_R\bigg[\frac{L_{\widetilde{Z}_S}(\mathcal{A}, \widetilde{Z}_S, R) + L_{\widetilde{Z}_{\bar{S}}}(\mathcal{A}, \widetilde{Z}_S, R)}{2}\bigg]\bigg) \leq \frac{D(P_{\lambda|\widetilde{Z}S} \,||\, P_{\lambda|\widetilde{Z}}) + \log\frac{2\sqrt{n}}{\delta}}{n}. \tag{25}$$

These high-probability bounds involve the empirical test loss of the learner, and not the population loss. However, a simple use of the triangle inequality, as detailed in [34, Thm. 3], allows one to make (24) and (25) explicit in the population loss. Note that the information measures that appear in the bounds depend on the full training set, rather than individual samples. While it is possible to obtain tail bounds with samplewise information measures [36, Lemma 8], such bounds have a detrimental linear dependence on $1/\delta$, instead of the benign logarithmic dependence in (24) and (25). Finally, under a stronger technical assumption of absolute continuity, one can obtain high-probability bounds not only with respect to the draw of $\widetilde{Z}$ and S, but also the randomness R of the learning algorithm. We present this result in Appendix B.

4 Expressiveness of the e-CMI Framework

We now illustrate the unifying nature of the e-CMI framework by demonstrating its expressiveness. Specifically, proceeding similarly to [17, 26], we use the e-CMI framework to rederive known bounds in learning theory. In particular, we consider multiclass classification with finite Natarajan dimension. In the next theorem, we provide a bound on the e-CMI appearing in the results reported in Section 2, as well as a bound on the data-dependent KL divergence appearing in (24) and (25).

Theorem 8. *Consider a multiclass classification setting, for which $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the instance space and $\mathcal{Y}$ the label space, and assume that $|\mathcal{Y}| = N$. Furthermore, assume that the learning algorithm implements a function $f : \mathcal{X} \to \mathcal{Y}$ where $f \in \mathcal{F}$ belongs to a class of finite Natarajan dimension d_N [37]. Finally, assume that $2n > d_N + 1$. Then,*

$$I(\ell(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z}); S|\widetilde{Z}) \leq d_N \log\bigg(\binom{N}{2}\frac{2en}{d_N}\bigg). \tag{26}$$

Furthermore, with probability at least $1-\delta$ under the draw of $\widetilde{Z}$ and S,

$$D(P_{\lambda|\widetilde{Z}S} \,||\, P_{\lambda|\widetilde{Z}}) \leq d_N \log\bigg(\binom{N}{2}\frac{2en}{d_N}\bigg) + \log\frac{1}{\delta}. \tag{27}$$

To obtain generalization bounds from Theorem 8, we need to upper-bound the information terms that appear in the bounds of Sections 2 and 3 by using either (26) or (27). While this can be done for any of the information-theoretic generalization bounds in this paper, we present two concrete examples in the following corollary.

Corollary 1. *Consider a multiclass classification setting, for which $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the instance space and $\mathcal{Y}$ the label space, and assume that $|\mathcal{Y}| = N$. Furthermore, assume that the*

learning algorithm implements a function $f : \mathcal{X} \to \mathcal{Y}$ where $f \in \mathcal{F}$ belongs to a class of finite Natarajan dimension d_N [37]. Finally, assume that $2n \geq d_N + 1$. Then,

$$\left| L_{\mathcal{D}} - \hat{L} \right| \leq \sqrt{\frac{2 d_N \log\left(\binom{N}{2} \frac{2en}{d_N} \right)}{n}}. \tag{28}$$

Furthermore, with probability at least $1 - \delta$ under the draw of $\widetilde{Z}$ and S,

$$d\left(\mathbb{E}_R\left[L_{\widetilde{Z}_S}(\mathcal{A}, \widetilde{Z}_S, R) \right] \,||\, \mathbb{E}_R\left[\frac{L_{\widetilde{Z}_S}(\mathcal{A}, \widetilde{Z}_S, R) + L_{\widetilde{Z}_{\bar{S}}}(\mathcal{A}, \widetilde{Z}_S, R)}{2} \right] \right) \leq \frac{d_N \log\left(\binom{N}{2} \frac{2en}{d_N} \right) + \log \frac{2}{\delta} + \log \frac{4\sqrt{n}}{\delta}}{n}. \tag{29}$$

The scaling behavior of (28) with respect to n recovers the standard rate found in the literature [38]. This is minimax optimal up to a logarithmic term [39, Thm. 29.3].

5 Comparing the Bounds

We now perform some comparisons between the bounds obtained in Section 2. As it turns out, there is no clear ordering between the different bounds in general. For a first comparison, we consider the interpolating setting, where the training loss is zero. We focus on the case where the size m of the random subset U equals 1. Since the squared bound in (6) is vacuous for this choice, it is not considered for this comparison. Moreover, we assume that the learning algorithm is, on average, indifferent to permutations of the training set. Specifically, we assume that the value of $I(\ell(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z}_i); S_i | \widetilde{Z})$ equals a constant B that does not depend on the index i. For a more straightforward comparison, we also exclude the disintegrated bounds in (4) and (20), and postpone their evaluation to Section 6. Thus, in this section, we compare the square-root bound in (5), the linear bound in (11), the interpolation bound in (12), and the binary KL bound in (18). In the following proposition, we establish an ordering between these bounds. The result follows by straightforward arithmetic.

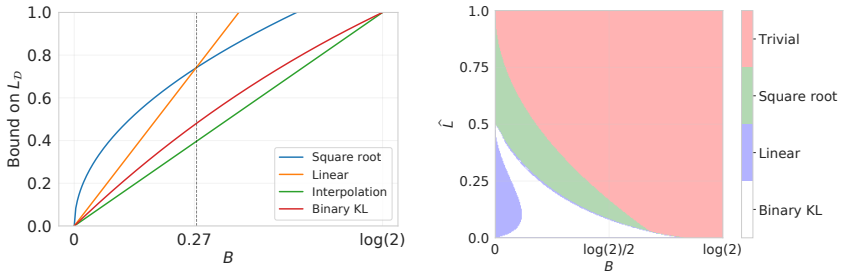
Proposition 1. *Assume that $I(\ell(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z}_i); S_i | \widetilde{Z}) = B$ for all i and that $\hat{L} = 0$. Let $\gamma_{1,\text{opt}}$ be the largest γ_1 such that $\gamma_1^2 - 4(e^{\gamma_1} - 1)(e^{\gamma_1} - 1 - \gamma_1) \geq 0$, and assume that $B < 2\gamma_{1,\text{opt}}^2 \approx 0.27$. Then, the upper bounds on $L_{\mathcal{D}}$ are, in increasing order, the interpolation bound in* (12)*, the binary KL bound in* (18)*, the linear bound in* (11)*, and the square-root bound in* (5)*. If $B > 2\gamma_{1,\text{opt}}^2$, the ordering between the square-root bound in* (11) *and the linear bound in* (5) *is inverted.*

While Proposition 1 gives the ordering between the bounds for the interpolating scenario, the quantitative difference between them is not clear. A numerical illustration is given in Figure 1a, under the same assumptions as in Proposition 1. As established in Proposition 1, the interpolation bound in (12) is superior across the range of values for B, and the binary KL bound in (18) is slightly looser. However, while the constant $\log 2$ in (12) is sharp, the derivation is valid only for interpolating learning algorithms. In contrast, the other bounds are only marginally affected when we allow for a small, non-zero training loss. This added flexibility shows the usefulness of (18) as compared to (12).

To obtain a more complete picture, we numerically evaluate the bounds for a range of values for $\hat{L}$ and B. This is illustrated in Figure 1b. While the binary KL bound in (18) is tightest for most small values of B and $\hat{L}$, a region of this space is dominated by the linear bound in (11). When both B and $\hat{L}$ grow larger, the square-root bound in (5) is the tightest. Thus, the picture that emerges from these comparisons is that, when the training loss is zero, it is best to use (12), while we need (18) for the case of a small, but non-zero, training loss. However, if the training loss is high, one needs to evaluate all three bounds and take the minimum.

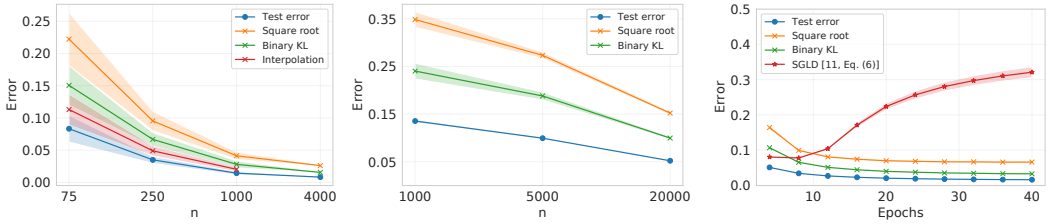
6 Numerical Results

We now consider three deep learning settings and evaluate the bounds derived in Section 2. To the best of our knowledge, the tightest average generalization bounds available in the literature for typical



(a) (b)

Figure 1: (a): A quantitative comparison between the square-root bound in (5), the linear bound in (11), the interpolation bound in (12), and the binary KL bound in (18) under the assumptions of Proposition 1. (b): A comparison between the square-root bound in (5), the linear bound in (11) and the binary KL bound in (18) under the assumptions of Proposition 1, but with non-zero $\hat{L}$. Each point in the figure is color-coded according to which bound gives the tightest characterization of $L_{\mathcal{D}}$ for the given parameters. For the region labeled *Trivial*, no bound performs better than the trivial $L_{\mathcal{D}} \leq 1$.



(a) MNIST, Adam (b) CIFAR10, SGD (c) MNIST, SGLD

Figure 2: Numerical evaluation of the test error in three deep learning scenarios, along with the upper bounds provided by the square-root bound in (4), the binary KL bound in (20) and, when applicable, the interpolation bound in (12) and the SGLD bound in [11, Eq. 6].

deep learning scenarios, such as CNNs trained on MNIST [40] or CIFAR10 [41], are found in [17]. These numerical results are based on [17, Eq. 22], which is a version of (4) where the e-CMI is replaced by the f-CMI. In order to perform a direct comparison, we consider the same experimental setups. The code for our experiments is largely based on the code from [17].[3] Additional numerical results for a setting with randomized labels are given in Appendix C. A detailed description of the training methods, network architectures, and experimental setup is given in Appendix D.

The previous section indicated that the interpolation bound in (12) and the binary KL bound in (20) are the superior bounds when the training loss is zero or low, respectively, which is common for deep learning. Hence, we focus on these bounds, as well as on the square-root bound in (4). When evaluating the bounds, we use the classification error as the loss function. We consider two learning algorithms, namely stochastic gradient descent (SGD) with a fixed random seed and stochastic gradient Langevin dynamics (SGLD). The first is a deterministic learning algorithm with high relevance for current practice. For deterministic learning algorithms, generalization bounds based on the information captured by the weights of a neural network are vacuous, but as noted by [17], bounds based on the information contained in the predictions (or the losses they induce) are typically nonvacuous. The second is a randomized learning algorithm, which allows for comparison to weight-based information-theoretic generalization bounds. For this setting, we also evaluate the SGLD bound of [11, Eq. 6].

First, we consider a CNN trained with Adam (a variant of SGD) on a binarized version of MNIST, where we only consider the digits 4 and 9. The results are reported in Figure 2a. In binary classification, there is a one-to-one mapping between predictions and losses. Hence, for this scenario, the information captured by the matrix of losses is the same as the information captured by the matrix of

[3] Available: `https://github.com/hrayrhar/f-CMI`.

predictions. For this scenario, the binary KL bound in (20) significantly improves on the square-root bound in (4). For $n = 75$, 250, and 1000, all of the neural networks that we trained achieved zero training error, making the interpolation bound in (12) a valid bound. This bound results in the tightest characterization of the test error. However, for $n = 4000$, the training error was non-zero for some neural networks. As a consequence, the interpolation bound in (12) was not applicable.

Next, in Figure 2b, we look at ResNet-50 pretrained on ImageNet and fine-tuned using SGD on CIFAR10. In multi-class classification, the map from predictions to losses is no longer one-to-one. Thus, this is a setting where the e-CMI is potentially lower than the f-CMI. Again, the binary KL bound in (20) significantly improves on the square-root bound in (4).

Finally, as an example of a randomized learning algorithm, we consider a CNN trained with SGLD on the binarized MNIST data set. This is shown in Figure 2c. As training progresses, both (20) and (4) improve on the SGLD bound in [11, Eq. 6]. However, the latter is lower for the initial epochs. While (20) reduces the magnitude of this discrepancy, which was also observed in [17], it does not fully close the gap. This may be due to the fact that the bound in [11, Eq. 6] has an expectation over R outside of the square root, similar to (8), which may be beneficial during early epochs.

7 Discussion and Limitations

In this paper, we presented several generalization bounds in terms of the evaluated CMI. While the results hold for a generically formulated learning problem, they are derived under the assumption of a bounded loss function and independent and identically distributed (i.i.d.) training data. While the boundedness assumption can be alleviated to some degree by using the same type of argument as in [16, Thm. 5.1], it is unclear to what extent the i.i.d. assumption can be relaxed.

Our experiments demonstrate that some of the derived bounds are numerically accurate for several deep learning scenarios. As shown in Appendix C, this remains true also when one considers randomized labels and varies the width, depth, and learning rate of the neural network. These results indicate that this family of bounds is potentially powerful enough to guide the design of deep neural networks. This is an intriguing avenue for future research.

Acknowledgements

This work was partly supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation and the Chalmers AI Research Center (CHAIR).

References

[1] J. Shawe-Taylor and R. C. Williamson. A PAC analysis of a Bayesian estimator. In *Proc. Conf. Learn. Theory (COLT)*, July 1997.

[2] D. A. McAllester. Some PAC-Bayesian theorems. In *Proc. Conf. Learn. Theory (COLT)*, Madison, WI, USA, July 1998.

[3] B. Guedj. A primer on PAC-Bayesian learning. *arXiv*, Jan. 2019.

[4] P. Alquier. User-friendly introduction to PAC-Bayes bounds. *arXiv*, Nov. 2021.

[5] D. Russo and J. Zou. Controlling bias in adaptive data analysis using information theory. In *Proc. Artif. Intell. Statist. (AISTATS)*, Cadiz, Spain, May 2016.

[6] A. Xu and M. Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, USA, Dec. 2017.

[7] A. R. Asadi, E. Abbe, and S. Verdú. Chaining mutual information and tightening generalization bounds. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Montreal, Canada, Dec. 2018.

[8] H. Hafez-Kolahi, Z. Golgooni, S. Kasaei, and M. Soleymani. Conditioning and processing: Techniques to improve information-theoretic generalization bounds. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Canada, Dec. 2020.

[9] B. Rodríguez-Gálvez, G. Bassi, R. Thobaben, and M. Skoglund. On random subset generalization error bounds and the stochastic gradient Langevin dynamics algorithm. In *Proc. Inf. Theory Workshop (ITW)*, Riva del Garda, Italy, Apr. 2020.

[10] Y. Bu, S. Zou, and V. V. Veeravalli. Tightening mutual information-based bounds on generalization error. *IEEE J. Sel. Areas Inf. Theory*, 1(1):121–130, May 2020.

[11] J. Negrea, M. Haghifam, G. K. Dziugaite, A. Khisti, and D. M. Roy. Information-theoretic generalization bounds for SGLD via data-dependent estimates. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Canada, Dec. 2019.

[12] M. Haghifam, J. Negrea, A. Khisti, D.M. Roy, and G.K. Dziugaite. Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Canada, 12 2020.

[13] F. Hellström and G. Durisi. Fast-rate loss bounds via conditional information measures with applications to neural networks. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Melbourne, Australia, July 2021.

[14] R. Zhou, C. Tian, and T. Liu. Individually conditional individual mutual information bound on generalization error. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Melbourne, Australia, July 2021.

[15] G. Neu, G. K. Dziugiate, M. Haghifam, and D. M.. Roy. Information-theoretic generalization bounds for stochastic gradient descent. In *Proc. Conf. Learn. Theory (COLT)*, Boulder, CO, USA, Aug. 2021.

[16] T. Steinke and L. Zakynthinou. Reasoning about generalization via conditional mutual information. In *Proc. Conf. Learn. Theory (COLT)*, Graz, Austria, July 2020.

[17] H. Harutyunyan, M. Raginsky, G. Ver Steeg, and A. Galstyan. Information-theoretic generalization bounds for black-box learning algorithms. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Virtual Conference, Dec. 2021.

[18] F. Hellström and G. Durisi. Data-dependent PAC-Bayesian bounds in the random-subset setting with applications to neural networks. In *Workshop on Inf.-Theoretic Methods Rigorous, Responsible, and Reliable Mach. Learn. (ITR3)*, Virtual conference, July 2021.

[19] J. Langford and M. Seeger. Bounds for averaging classifiers. *CMU Technical report*, CMU-CS-01-102, 2001.

[20] M. Seeger. PAC-Bayesian generalisation error bounds for Gaussian process classification. *J. of Mach. Learn. Res.*, 3:233–269, Oct. 2002.

[21] J. Langford. *Quantitatively Tight Sample Complexity Bounds*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 2002.

[22] A. Maurer. A note on the PAC Bayesian theorem. *arXiv*, Nov. 2004.

[23] P. Germain, A. Lacasse, F. Laviolette, and M. Marchand. PAC-Bayesian learning of linear classifiers. In *Proc. Int. Conf. Mach. Learning (ICML)*, Montreal, Canada, June 2009.

[24] L. Bégin, P. Germain, F. Laviolette, and J. F. Roy. PAC-Bayesian theory for transductive learning. In *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, Reykjavik, Iceland, Apr. 2014.

[25] O. Rivasplata, I. Kuzborskij, C. Szepesvari, and J. Shawe-Taylor. PAC-Bayes analysis beyond the usual bounds. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Canada, Dec. 2020.

[26] M. Haghifam, G. K. Dziugaite, S. Moran, and D. M. Roy. Towards a unified information-theoretic framework for generalization. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Virtual Conference, Dec. 2021.

[27] O. Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56. IMS Lecture Notes Monogr. Ser., 2007.

[28] M. Haddouche, B. Guedj, O. Rivasplata, and J. Shawe-Taylor. PAC-Bayes unleashed: Generalisation bounds with unbounded losses. *Entropy*, 23(10), Oct. 2021.

[29] D. A. McAllester. A PAC-Bayesian tutorial with a dropout bound. *arXiv*, July 2013.

[30] Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer. PAC-Bayesian inequalities for martingales. *IEEE Trans. Inf. Theory*, 58(12):7086–7093, Aug. 2012.

[31] W. Hoeffding. On the distribution of the number of successes in independent trials. *The Annals of Mathematical Statistics*, 27(3):713–721, Sep. 1956.

[32] J. Guan, Z. Lu, and Y. Liu. Improved generalization risk bounds for meta-learning with PAC-Bayes-kl analysis. `https://openreview.net/forum?id=XgS9YPYtdj`, 2021.

[33] Z. Mhammedi, P. Grünwald, and B. Guedj. PAC-Bayes un-expected Bernstein inequality. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Canada, Dec. 2019.

[34] F. Hellström and G. Durisi. Generalization bounds via information density and conditional information density. *IEEE J. Sel. Areas Inf. Theory*, 1(3):824–839, Dec. 2020.

[35] P. Grünwald, T. Steinke, and L. Zakynthinou. PAC-Bayes, MAC-Bayes and conditional mutual information: Fast rate bounds that handle general VC classes. In *Proc. Conf. Learn. Theory (COLT)*, Boulder, CO, USA, Aug. 2021.

[36] H. Wang, Y. Huang, R. Gao, and F. Calmon. Analyzing the generalization capability of SGLD using properties of Gaussian channels. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Virtual Conference, Dec. 2021.

[37] B. K. Natarajan. On learning sets and functions. *Machine Learning*, 4:67–97, 1989.

[38] Y. Guermeur. Large margin multi-category discriminant models and scale-sensitive psi-dimensions. Research report, INRIA, 2004.

[39] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge Univ. Press, Cambridge, U.K., 2014.

[40] L. Deng. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Process. Mag.*, 29(6):141–142, 2012.

[41] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[42] M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time, i. *Comm. Pure Appl. Math*, 28(1):1–47, Jan. 1975.

[43] Y. Polyanskiy and Y. Wu. *Lecture Notes On Information Theory*. 2019.

[44] M. J. Wainwright. *High-Dimensional Statistics: a Non-Asymptotic Viewpoint*. Cambridge Univ. Press, Cambridge, U.K., 2019.

[45] D. Haussler and P. M. Long. A generalization of Sauer's lemma. *J. Combinatorial Theory, Series A*, 71(2):219–240, 1995.

[46] G. K. Dziugaite and D. M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proc. Conf. Uncertainty in Artif. Intell. (UAI)*, Sydney, Australia, Aug. 2017.

[47] I. Csiszar and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge Univ. Press, Cambridge, U.K., 2nd edition, 2011.

[48] G. K. Dziugaite and D. M. Roy. Data-dependent PAC-Bayes priors via differential privacy. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Montreal, Canada, Dec. 2018.

[49] A. Ambroladze, E. Parrado-Hernandez, and J. Shawe-Taylor. Tighter PAC-Bayes bounds. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Canada, Dec. 2006.

[50] G.K. Dziugaite, K. Hsu, W. Gharbieh, and D.M. Roy. On the role of data in PAC-Bayes bounds. In *Proc. Artif. Intell. Statist. (AISTATS)*, San Diego, CA, USA, Apr. 2021.

[51] M. Perez-Ortiz, O. Risvaplata, J. Shawe-Taylor, and C. Szepesvari. Tighter risk certificates for neural networks. *J. of Mach. Learn. Res.*, 22(227), Aug. 2021.

[52] L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, June 2003.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We clarify our contributions by specifically referencing where they can be found in the paper.
   (b) Did you describe the limitations of your work? [Yes] We specify the assumptions and settings for which our results hold.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] The assumptions are stated in Section 1 or in the Theorem statements.
   (b) Did you include complete proofs of all theoretical results? [Yes] Full proofs are provided in Appendix A.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Yes, this is described in the file readme.txt in the supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] This is specified in the supplementary material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] All random results include error bars.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] This is specified in the supplementary material.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] The assets that we use are cited in Section 6 and Appendix D.
   (b) Did you mention the license of the assets? [Yes] The license is discussed in Appendix D.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See supplementary material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] The data used are standard data sets that only include images of objects and animals.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] The data used are standard data sets that only include images of objects and animals.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Deferred Proofs

We now present the proofs of the results from the main paper. In Section A.1, we present the proofs of the average generalization bounds from Section 2. In Section A.2, we prove the high-probability bounds from Section 3. Finally, in Section A.3, we prove the bounds on the information measures from Section 4.

A.1 Proofs for Section 2

We start with Lemma 1.

Proof of Lemma 1. By Jensen's inequality,

$$f_\gamma(\mathbb{E}_{X,Y}[g_1(X,Y)], \mathbb{E}_{X,Y}[g_2(X,Y)]) \leq \mathbb{E}_{X,Y}[f_\gamma(g_1(X,Y), g_2(X,Y))]. \tag{30}$$

By Donsker-Varadhan's variational representation of the KL divergence [42], we have

$$\mathbb{E}_{X,Y}[f_\gamma(g_1(X,Y), g_2(X,Y))] \leq D(P_{XY} \,||\, P_X P_Y) + \log \mathbb{E}_{X,Y'}\left[e^{f_\gamma(g_1(X,Y'), g_2(X,Y'))}\right]. \tag{31}$$

The result follows by noting that $D(P_{XY} \,||\, P_X P_Y) = I(X;Y)$, by re-arranging, and by taking the supremum with respect to γ. □

We now present the proofs of the CMI results given in Section 2. We begin with Remark 1.

Proof of Remark 1. Since each of the random variables is obtained as a function of the previous one, we have the Markov chain

$$W \to \phi_W(\tilde{x}_u) \to \ell(W, \tilde{z}_u). \tag{32}$$

Therefore, the result follows by the data-processing inequality [43, Thm. 2.3.4]. □

Next, we specialize the result of Lemma 1 to the CMI setting in a way that allows for disintegrated CMI results and random subsets.

Corollary 2. *Consider the CMI setting, and fix $\tilde{z}$ and u. As a shorthand, we let $\lambda_u = \ell(\mathcal{A}(\tilde{z}_S, R), \tilde{z}_u)$ denote the matrix of losses incurred on each entry of $\tilde{z}_u$. Let $\lambda_{S_u} = (\lambda_{u_1,S_{u_1}}, \ldots, \lambda_{u_m,S_{u_m}})$. Furthermore, let $\hat{\lambda}_{S_u} = \frac{1}{m}\sum_{i=1}^m \lambda_{u_i,S_{u_i}}$ denote the average loss on the entries of λ selected on the basis of u and S. Let S' be a random variable with the same marginal distribution as S such that λ and S' are independent. Then,*

$$\sup_\gamma \mathbb{E}_{\lambda_u,S_u}\left[f_\gamma(\hat{\lambda}_{S_u}, \hat{\lambda}_{\bar{S}_u})\right] - \log \mathbb{E}_{\lambda_u,S'_u}\left[e^{f_\gamma(\hat{\lambda}_{S'_u}, \hat{\lambda}_{\bar{S}'_u})}\right] \leq I^{\tilde{z},u}(\lambda_u; S_u). \tag{33}$$

Proof. The result follows immediately by applying Lemma 1 with $X = \lambda_u$, $Y = S_u$, $g_1(\lambda_u, S_u) = \hat{\lambda}_{S_u}$, and $g_2(\lambda_u, S_u) = \hat{\lambda}_{\bar{S}_u}$. Since S_u is a discrete random variable, the joint distribution of λ_u, S_u is absolutely continuous with respect to the product of the marginal distributions, that is, the joint distribution of λ_u, S'_u. □

With this, we are ready to derive the results in Theorem 1.

Proof of Theorem 1. Let $f_\gamma(\hat{\lambda}_{S'_u}, \hat{\lambda}_{\bar{S}'_u}) = \gamma(\hat{\lambda}_{\bar{S}'_u} - \hat{\lambda}_{S'_u}) \equiv \gamma\Delta_{S'_u}$, where we use the same notation as in Corollary 2. Due to the symmetry in the definition of the training and test sets, we have that $\Delta_{S'_u} = -\Delta_{\bar{S}'_u}$. This implies that $\mathbb{E}_{S'_u}\left[\Delta_{S'_u}\right] = 0$. Furthermore, $\Delta_{S'_u}$ is the arithmetic average of m independent terms, each with a range of $[-1, 1]$. Thus, $\Delta_{S'_u}$ is a $1/\sqrt{m}$-sub-Gaussian random variable, from which it follows by definition that

$$\log \mathbb{E}_{S'_u}\left[e^{\gamma\Delta_{S'_u}}\right] \leq \frac{\gamma^2}{2m}. \tag{34}$$

Inserting this into (33), we find that

$$\sup_{\gamma} \gamma\Big(\mathbb{E}_{\lambda_u,S_u}\Big[\hat{\lambda}_{S_u}-\hat{\lambda}_{\bar{S}_u}\Big]\Big)-\frac{\gamma^2}{2m}=\frac{m\Big(\mathbb{E}_{\lambda_u,S_u}\Big[\hat{\lambda}_{S_u}-\hat{\lambda}_{\bar{S}_u}\Big]\Big)^2}{2} \leq I^{\tilde{z},u}(\lambda_u;S_u). \tag{35}$$

Since this result is valid for any fixed $\tilde{z}$ and u, it also holds when we average over them. Thus, by re-arranging and averaging, replacing shorthands by their long forms,

$$\mathbb{E}_{\widetilde{Z},U}\Big[\Big|\mathbb{E}_{S,R}\Big[L_{\widetilde{Z}_{S_U}}(\mathcal{A},\widetilde{Z}_S,R)-L_{\widetilde{Z}_{\bar{S}_U}}(\mathcal{A},\widetilde{Z}_S,R)\Big]\Big|\Big] \leq \mathbb{E}_{\widetilde{Z},U}\left[\sqrt{\frac{2I^{\widetilde{Z},U}(\ell(\mathcal{A}(\widetilde{Z}_S,R),\widetilde{Z}_U);S_U)}{m}}\right]. \tag{36}$$

By applying Jensen's inequality, the expectation and absolute value in the left-hand side can be swapped, and by a further use of Jensen's inequality, we can similarly swap the expectation and square root in the right-hand side. Thus, we have shown that

$$\Big|\mathbb{E}_{\widetilde{Z},S,R}\Big[L_{\mathcal{D}}(\mathcal{A},\widetilde{Z}_S,R)-L_{\widetilde{Z}_S}(\mathcal{A},\widetilde{Z}_S,R)\Big]\Big| \leq \mathbb{E}_{\widetilde{Z},U}\left[\sqrt{\frac{2I^{\widetilde{Z},U}(\ell(\mathcal{A}(\widetilde{Z}_S,R),\widetilde{Z}_U);S_U)}{m}}\right] \tag{37}$$

$$\leq \mathbb{E}_{U}\left[\sqrt{\frac{2I(\ell(\mathcal{A}(\widetilde{Z}_S,R),\widetilde{Z}_U);S_U|\widetilde{Z},U)}{m}}\right]. \tag{38}$$

The right-hand side is an increasing function in the size m of the random subset U [17, Prop. 1]. Hence, we set $m=1$. This establishes the first part of Theorem 1. For the second part of the theorem, we instead consider $f_\gamma(\hat{\lambda}_{S'_u},\hat{\lambda}_{\bar{S}'_u})=\gamma\Delta^2_{\bar{S}'_u}$, with $\gamma\in[0,m/4)$. Since $\Delta_{S'_u}$ is $1/\sqrt{m}$-sub-Gaussian with zero mean, we have [17, Lemma 2]

$$\log \mathbb{E}_{S'_u}\Big[e^{\gamma\Delta^2_{\bar{S}'_u}}\Big] \leq \log\Big(1+\frac{8\gamma}{m}\Big). \tag{39}$$

By using this in (33), we find that

$$\sup_{\gamma}\gamma\,\mathbb{E}_{\lambda_u,S_u}\left[\Big(\hat{\lambda}_{S_u}-\hat{\lambda}_{\bar{S}_u}\Big)^2\right]-\log\Big(1+\frac{8\gamma}{m}\Big) \leq I^{\tilde{z},u}(\lambda_u;S_u). \tag{40}$$

For reasonable values of the involved quantities, the left-hand side is expected to grow as a function of γ. We therefore let $\gamma\to m/4$. Since the bound in (40) holds for the supremum, it also holds for any other choice of γ, including this one. We now have

$$\mathbb{E}_{\lambda_u,S_u}\left[\Big(\hat{\lambda}_{S_u}-\hat{\lambda}_{\bar{S}_u}\Big)^2\right] \leq \frac{4}{m}\left(I^{\tilde{z},u}(\lambda_u;S_u)+\log 3\right). \tag{41}$$

Averaging over $\widetilde{Z}$ and U, replacing shorthands by their long forms, we obtain

$$\mathbb{E}_{\widetilde{Z},R,S,U}\left[\Big(L_{\widetilde{Z}_{S_U}}(\mathcal{A},\widetilde{Z}_S,R)-L_{\widetilde{Z}_{\bar{S}_U}}(\mathcal{A},\widetilde{Z}_S,R)\Big)^2\right] \leq \frac{4}{m}\Big(I(\ell(\mathcal{A}(\widetilde{Z}_S,R),\widetilde{Z}_U);S_U|\widetilde{Z},U)+\log 3\Big). \tag{42}$$

We now want to replace the test loss inside the square with the population loss, while still keeping the expectation outside the square. We can do this by using the results of [17, Eq. (116)-Eq. (123)], from which it follows that

$$\mathbb{E}_{\widetilde{Z},R,S}\left[\Big(L_{\mathcal{D}}(\mathcal{A},\widetilde{Z}_S,R)-L_{\widetilde{Z}_S}(\mathcal{A},\widetilde{Z}_S,R)\Big)^2\right]$$

$$\leq 2\,\mathbb{E}_{\widetilde{Z},R,S,U}\left[\Big(L_{\widetilde{Z}_{\bar{S}_U}}(\mathcal{A},\widetilde{Z}_S,R)-L_{\widetilde{Z}_{S_U}}(\mathcal{A},\widetilde{Z}_S,R)\Big)^2\right]+\frac{1}{2m} \tag{43}$$

$$\leq \frac{8}{m}\Big(I(\ell(\mathcal{A}(\widetilde{Z}_S,R),\widetilde{Z}_U);S_U|\widetilde{Z},U)+\log 3\Big)+\frac{1}{2m} \tag{44}$$

$$\leq \frac{8}{m}\Big(I(\ell(\mathcal{A}(\widetilde{Z}_S,R),\widetilde{Z}_U);S_U|\widetilde{Z},U)+2\Big). \tag{45}$$

□

Proof of Theorem 2. Consider the CMI setting, with a fixed $\tilde{z}$, u and r. As a shorthand, we let $\tau_u = \ell(\mathcal{A}(\tilde{z}_S, r), \tilde{z}_u)$ denote the matrix of losses incurred on each sample of $\tilde{z}_u$. Let $\tau_{S_u} = (\tau_{u_1,S_{u_1}}, \ldots, \tau_{u_m,S_{u_m}})$. Furthermore, let $\hat{\tau}_{S_u} = \frac{1}{m}\sum_{i=1}^m \tau_{u_i,S_i} = L_{\tilde{z}_{S_u}}(\mathcal{A}, \tilde{z}_S, r)$ denote the average loss on the samples of τ indicated by u and S. By following the same steps as in the proof of Theorem 2, we can show that

$$\sup_\gamma \mathbb{E}_{\tau_u,S_u}\left[f_\gamma(\hat{\tau}_{S_u}, \hat{\tau}_{\bar{S}_u})\right] - \log \mathbb{E}_{\tau_u,S'_u}\left[e^{f_\gamma(\hat{\tau}_{S'_u}, \hat{\tau}_{\bar{S}'_u})}\right] \leq I^{\tilde{z},u,r}(\tau_u; S_u). \tag{46}$$

Let $f_\gamma(\hat{\tau}_{S'_u}, \hat{\tau}_{\bar{S}'_u}) = \gamma(\hat{\tau}_{\bar{S}'_u} - \hat{\tau}_{S'_u}) = \gamma\Xi_{S'_u}$. By the same argument as in the proof of Theorem 1, we conclude that $\Xi_{S'_u}$ is $1/\sqrt{m}$-sub-Gaussian. Thus, (34) holds with $\Xi_{S'_u}$ in place of $\Delta_{S'_u}$. Inserting this into (46), we find that

$$\sup_\gamma \gamma\big(\mathbb{E}_{\tau_u,S_u}\left[\hat{\tau}_{S_u} - \hat{\tau}_{\bar{S}_u}\right]\big) - \frac{\gamma^2}{2m} = \frac{m\big(\mathbb{E}_{\tau_u,S_u}\left[\hat{\tau}_{S_u} - \hat{\tau}_{\bar{S}_u}\right]\big)^2}{2} \leq I^{\tilde{z},u,r}(\tau_u; S_u). \tag{47}$$

Finally, by re-arranging, replacing shorthands by their long forms, averaging over $\widetilde{Z}$, S, and R, and using Jensen's inequality to move the expectation inside the absolute value, we find that

$$\left|\mathbb{E}_{\widetilde{Z},S,R}\Big[L_{\mathcal{D}}(\mathcal{A}, \widetilde{Z}_S, R) - L_{\widetilde{Z}_S}(\mathcal{A}, \widetilde{Z}_S, R)\Big]\right| \leq \mathbb{E}_{\widetilde{Z},U,R}\left[\sqrt{\frac{2I^{\widetilde{Z},U,R}(\ell(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z}_U); S_U)}{m}}\right]. \tag{48}$$

The desired bound in (8) follows by setting $m = 1$, and (9) follows by Jensen's inequality. □

Proof of Theorem 3. We need to use the following two concentration results that are implicitly used in [16] and more explicitly stated in [18].

Lemma 3. *Consider a binary random variable X satisfying $P(X = a) = P(X = b) = 1/2$ where $a, b \in [0, 1]$. Let $\bar{X} = a + b - X$ be the complement of X in $\{a, b\}$. Assume that γ_1 and γ_2 satisfy $\gamma_1(1 - \gamma_2) + (e^{\gamma_1} - 1 - \gamma_1)(1 + \gamma_2^2) \leq 0$. Then,*

$$\mathbb{E}_X\left[e^{\gamma_1\left(X - \gamma_2\bar{X}\right)}\right] \leq 1. \tag{49}$$

Furthermore, assume that $a = 0$. Then,

$$\lim_{\gamma\to\infty} \mathbb{E}_X\left[e^{\log 2(\bar{X} - \gamma X)}\right] \leq 1. \tag{50}$$

Proof. The first part follows by [13, Eq. (7)]. For the second part, note that

$$\mathbb{E}_X\left[e^{\log 2(\bar{X} - \gamma X)}\right] = \frac{e^{b\log 2} + e^{-b\gamma\log 2}}{2}. \tag{51}$$

If $b = 0$, the right-hand side of (51) equals 1 for all γ. If $b \neq 0$, we have

$$\lim_{\gamma\to\infty} \frac{e^{b\log 2} + e^{-b\gamma\log 2}}{2} = \frac{e^{b\log 2}}{2} \leq \frac{e^{\log 2}}{2} = 1, \tag{52}$$

where the last inequality follows because $b \in [0, 1]$. □

We now proceed with the proof of Theorem 3. Fix $\tilde{z}$ and u. To apply Corollary 2, we set $f_\gamma(\hat{\lambda}_{S'_u}, \hat{\lambda}_{\bar{S}'_u}) = m\gamma_1(\hat{\lambda}_{S'_u} - \gamma_2\hat{\lambda}_{\bar{S}'_u})$, where we use the same notation as in Corollary 2. Since the elements of S'_u are independent,

$$\mathbb{E}_{S'_u}\left[e^{m\gamma_1(\hat{\lambda}_{S'_u} - \gamma_2\hat{\lambda}_{\bar{S}'_u})}\right] = \prod_{i=1}^m \mathbb{E}_{S'_{u_i}}\left[e^{\gamma_1\left(\lambda_{u_i,S'_{u_i}} - \gamma_2\lambda_{u_i,\bar{S}'_{u_i}}\right)}\right] \leq 1 \tag{53}$$

where the last inequality follows from (49). Hence,

$$\sup_\gamma \mathbb{E}_{\lambda_u,S_u}\left[n\gamma_1(\hat{\lambda}_{S_u} - \gamma_2\hat{\lambda}_{\bar{S}_u})\right] \leq I^{\tilde{z},u}(\lambda_u; S_u). \tag{54}$$

Since (54) holds when we take the supremum over γ, it also holds for any other choice of γ. For now, let $\gamma = (\gamma_1, \gamma_2)$ be any pair of parameters that satisfy the constraint of Theorem 3. Averaging over $\widetilde{Z}$ and U, replacing shorthands by their long forms, we get

$$\mathbb{E}_{\widetilde{Z},S,R}\Big[L_{\mathcal{D}}(\mathcal{A}, \widetilde{Z}_S, R)\Big] \leq \gamma_2\, \mathbb{E}_{\widetilde{Z},S,R}\Big[L_{\widetilde{Z}_S}(\mathcal{A}, \widetilde{Z}_S, R)\Big] + \frac{I(\ell(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z}_U); S_U|\widetilde{Z}, U)}{m\gamma_1}. \quad (55)$$

Since the right-hand side is an increasing function of m, the tightest result is obtained for $m = 1$. Making this choice and optimizing the parameters, we obtain (11).

We now turn to (12). Fix $\tilde{z}$ and u. Now, set $f_\gamma(\hat{\lambda}_{S'_u}, \hat{\lambda}_{\bar{S}'_u}) = m \log 2(\hat{\lambda}_{\bar{S}'_u} - \gamma\hat{\lambda}_{S'_u})$. Then, (33) implies that

$$\mathbb{E}_{\lambda_u,S_u}\Big[m \log 2(\hat{\lambda}_{\bar{S}_u} - \gamma\hat{\lambda}_{S_u})\Big] - \log \mathbb{E}_{\lambda_u,S'_u}\Big[e^{m \log 2(\hat{\lambda}_{\bar{S}'_u} - \gamma\hat{\lambda}_{S'_u})}\Big] \leq I^{\tilde{z},u}(\lambda_u; S_u). \quad (56)$$

In view of (50), we let $\gamma \to \infty$. Note that the assumption that $\hat{L} = 0$ implies that $\lambda_{i,S_i} = 0$ with probability 1. Thus, by the independence of the elements of S'_u, (50) implies that

$$\lim_{\gamma\to\infty} \mathbb{E}_{\lambda_u,S'_u}\left[e^{m \log 2(\hat{\lambda}_{\bar{S}'_U} - \gamma\hat{\lambda}_{S'_U})}\right] = \lim_{\gamma\to\infty} \prod_{i=1}^{m} \mathbb{E}_{\lambda_u,S'_{U_i}}\left[e^{\log 2\left(\lambda_{U_i,\bar{S}'_{U_i}} - \gamma\lambda_{U_i,S'_{U_i}}\right)}\right] \leq 1. \quad (57)$$

We now average over $\widetilde{Z}$ and U. By the assumption that $\hat{L} = 0$, we have

$$\mathbb{E}_{\widetilde{Z},U,\lambda_U,S_U}\Big[m \log 2(\hat{\lambda}_{\bar{S}_u} - \gamma\hat{\lambda}_{S_u})\Big] = \mathbb{E}_{\widetilde{Z},U,\lambda_U,S_U}\Big[m \log 2(\hat{\lambda}_{\bar{S}_u})\Big]. \quad (58)$$

This means that we can discard the second term in (56) when $\gamma \to \infty$. We finally establish the desired result by setting $m = 1$. □

Proof of Lemma 2. Let $Y_1, \ldots, Y_n$ be independent Bernoulli random variables satisfying $\mathbb{E}[Y_i] = \mu_i$. It follows from [30, Lemma 1] that, for every convex function f,

$$\mathbb{E}[f(X_1, \ldots, X_n)] \leq \mathbb{E}[f(Y_1, \ldots, Y_n)]. \quad (59)$$

Next, let $Y = \sum_{i=1}^n Y_i$ and let $\mu = \sum_{i=1}^n \mu_i/n$. Also, let $\bar{Y} \sim \mathrm{Bin}(n, \mu)$ be a Binomial random variable. Then, for any strictly convex function g, we have [31, Thm. 3]

$$\mathbb{E}[g(Y)] \leq \mathbb{E}\big[g(\bar{Y})\big]. \quad (60)$$

Now, let h be the linear function $h(X_1, \ldots, X_n) = \sum_{i=1}^n X_i = X$. Notice that the strict convexity of g implies that $g \circ h$ is convex. It then follows from (59) with $f = g \circ h$ that

$$\mathbb{E}[g(X)] = \mathbb{E}[g \circ h(X_1, \ldots, X_n)] \leq \mathbb{E}[g \circ h(Y_1, \ldots, Y_n)] = \mathbb{E}[g(Y)]. \quad (61)$$

Combining (60) and (61), we conclude that

$$\mathbb{E}[g(X)] \leq \mathbb{E}\big[g(\bar{Y})\big]. \quad (62)$$

Now, we set $g(x) = \exp(nd_\gamma(x/n \,||\, \mu))$, which is strictly convex. Then, it follows from (62) that

$$\mathbb{E}[\exp(nd_\gamma(X/n \,||\, \mu))] \leq \mathbb{E}\big[\exp(nd_\gamma(\bar{Y}/n \,||\, \mu))\big]. \quad (63)$$

Note that $X/n = \hat{\mu}$. Next, by [29, Eq. (17)], we have

$$\mathbb{E}\big[\exp(nd_\gamma(\bar{Y}/n \,||\, \mu))\big] \leq 1. \quad (64)$$

Combining (63) and (64), we obtain the desired result. □

Proof of Theorem 4 and Theorem 5. The results in Theorem 4 and Theorem 5 follow as a special case of Theorem 10, which is stated and proven in Appendix B. □

Proof of Theorem 6. With the standard convention that $0 \log 0 = 0$, which is motivated by continuity, we get

$$d\Big(0 \,||\, \frac{p}{2}\Big) = \log \frac{1}{1 - \frac{p}{2}}. \quad (65)$$

Then, it follows from (19) that $d^{-1}(0, c) = 2 - 2e^{-c}$. Combining this with (20), we obtain (23). □

A.2 Proofs for Section 3

We now turn to the proof of the high-probability bound in Theorem 7.

Proof of Theorem 7. Let the size of the subset U be n. Let $f_\gamma(\cdot,\cdot)$ be a function that is jointly convex in its arguments. Now, consider the random variables $\lambda \sim P_{\lambda|\widetilde{Z}S}$ and $\lambda' \sim P_{\lambda|\widetilde{Z}}$. We use the notation from Corollary 2, and set $\lambda_S = (\lambda_{1,S_1},\ldots,\lambda_{n,S_n})$, $\hat{\lambda}_S = \sum_{i=1}^n \lambda_{i,S_i}/n$ and $\hat{\lambda}_{\bar{S}} = \sum_{i=1}^n \lambda_{i,\bar{S}_i}/n$. Then, by Jensen's inequality,

$$f_\gamma\Big(\mathbb{E}_R\Big[L_{\widetilde{Z}_S}(\mathcal{A},\widetilde{Z}_S,R)\Big],\mathbb{E}_R\Big[L_{\widetilde{Z}_{\bar{S}}}(\mathcal{A},\widetilde{Z}_S,R)\Big]\Big) = f_\gamma\Big(\mathbb{E}_\lambda\Big[\hat{\lambda}_S\Big],\mathbb{E}_\lambda\Big[\hat{\lambda}_{\bar{S}}\Big]\Big) \tag{66}$$

$$\leq \mathbb{E}_\lambda\Big[f_\gamma(\hat{\lambda}_S,\hat{\lambda}_{\bar{S}})\Big]. \tag{67}$$

By Donsker-Varadhan's variational representation of the KL divergence [42],

$$\mathbb{E}_\lambda\Big[f_\gamma(\hat{\lambda}_S,\hat{\lambda}_{\bar{S}})\Big] \leq D(P_{\lambda|\widetilde{Z}S}\,||\,P_{\lambda|\widetilde{Z}}) + \log\mathbb{E}_{\lambda'}\Big[\exp(f_\gamma(\hat{\lambda}'_S,\hat{\lambda}'_{\bar{S}}))\Big]. \tag{68}$$

By Markov's inequality, we conclude that with probability at least $1-\delta$ under the draw of $\widetilde{Z}$ and S,

$$D(P_{\lambda|\widetilde{Z}S}\,||\,P_{\lambda|\widetilde{Z}}) + \log\mathbb{E}_{\lambda'}\Big[\exp(f_\gamma(\hat{\lambda}'_S,\hat{\lambda}'_{\bar{S}}))\Big] \tag{69}$$

$$\leq D(P_{\lambda|\widetilde{Z}S}\,||\,P_{\lambda|\widetilde{Z}}) + \log\frac{1}{\delta}\,\mathbb{E}_{\lambda',\widetilde{Z},S}\Big[\exp(f_\gamma(\hat{\lambda}'_S,\hat{\lambda}'_{\bar{S}}))\Big] \tag{70}$$

$$= D(P_{\lambda|\widetilde{Z}S}\,||\,P_{\lambda|\widetilde{Z}}) + \log\frac{1}{\delta}\,\mathbb{E}_{\lambda,\widetilde{Z},S'}\Big[\exp(f_\gamma(\hat{\lambda}_{S'},\hat{\lambda}_{\bar{S}'}))\Big]. \tag{71}$$

In (71), S' is a copy of S that is independent from λ. The step from (70) to (71) relies on the observation that λ' is independent of S but has the same joint distribution with $\widetilde{Z}$ as λ does. Now, let $f_\gamma(\hat{\lambda}_{S'},\hat{\lambda}_{\bar{S}'}) = \frac{(n-1)}{2}(\hat{\lambda}_{\bar{S}'}-\hat{\lambda}_{S'})^2 \equiv \frac{(n-1)}{2}\Delta_{S'}^2$. Due to the symmetry in the definition of the training and test sets, we have that $\Delta_{S'} = -\Delta_{\bar{S}'}$. This implies that $\mathbb{E}_{S'}[\Delta_{S'}] = 0$. Furthermore, $\Delta_{S'}$ is the arithmetic average of n independent terms, each with a range of $[-1,1]$. Thus, $\Delta_{S'}$ is a $1/\sqrt{n}$-sub-Gaussian random variable, from which it follows that [44, Thm. 2.6.(IV)]

$$\log\mathbb{E}_{S'}\Big[e^{\frac{(n-1)}{2}\Delta_{S'}^2}\Big] \leq \log\big(\sqrt{n}\big). \tag{72}$$

Inserting this into (71) and combining the result with (67), we obtain

$$\frac{n-1}{2}\,\mathbb{E}_R\Big[(L_{\widetilde{Z}_{\bar{S}}}(\mathcal{A},\widetilde{Z}_S,R)-L_{\widetilde{Z}_S}(\mathcal{A},\widetilde{Z}_S,R))^2\Big] \leq D(P_{\lambda|\widetilde{Z}S}\,||\,P_{\lambda|\widetilde{Z}}) + \log\frac{\sqrt{n}}{\delta}. \tag{73}$$

By Jensen's inequality, this implies that

$$\frac{n-1}{2}\Big(\mathbb{E}_R\Big[L_{\widetilde{Z}_{\bar{S}}}(\mathcal{A},\widetilde{Z}_S,R)-L_{\widetilde{Z}_S}(\mathcal{A},\widetilde{Z}_S,R)\Big]\Big)^2 \leq D(P_{\lambda|\widetilde{Z}S}\,||\,P_{\lambda|\widetilde{Z}}) + \log\frac{\sqrt{n}}{\delta} \tag{74}$$

from which the result in (24) follows.

To prove (25), let $f_\gamma(\hat{\lambda}_{S'},\hat{\lambda}_{\bar{S}'}) = nd(\hat{\lambda}_{S'}\,||\,(\hat{\lambda}_{S'}+\hat{\lambda}_{\bar{S}'})/2)$. We now note that $\mathbb{E}_{S'}\Big[\hat{\lambda}_{S'}\Big] = \hat{\lambda}$, where $\hat{\lambda} = \sum_{i=1}^n(\lambda_{i,0}+\lambda_{i,1})/2n$. For any s', this can be written as $\hat{\lambda} = (\hat{\lambda}_{s'}+\hat{\lambda}_{\bar{s}'})/2$ due to symmetry. Therefore, by [22, Thm. 1],

$$\log\mathbb{E}_{S'}\Big[e^{nd(\hat{\lambda}_{S'}\,||\,(\hat{\lambda}_{S'}+\hat{\lambda}_{\bar{S}'})/2)}\Big] \leq \log\Big(\sqrt{2n}\Big). \tag{75}$$

Again, inserting this into (71) and combining the result with (67), we obtain

$$d\Bigg(\mathbb{E}_R\Big[L_{\widetilde{Z}_S}(\mathcal{A},\widetilde{Z}_S,R)\Big]\;||\;\mathbb{E}_R\Bigg[\frac{L_{\widetilde{Z}_S}(\mathcal{A},\widetilde{Z}_S,R)+L_{\widetilde{Z}_{\bar{S}}}(\mathcal{A},\widetilde{Z}_S,R)}{2}\Bigg]\Bigg)$$

$$\leq \frac{D(P_{\lambda|\widetilde{Z}S}\,||\,P_{\lambda|\widetilde{Z}}) + \log\frac{2\sqrt{n}}{\delta}}{n}, \tag{76}$$

which gives the desired result. □

A.3 Proofs for Section 4

Before proceeding with the proof of Theorem 8, we need the following lemma.

Lemma 4. *Let $g_{\mathcal{F}}(\cdot)$ denote the growth function of the function class $\mathcal{F}$, i.e., $g_{\mathcal{F}}(m)$ is the maximum number of different ways in which a data set of size m can be classified using functions from $\mathcal{F}$. For any function class $\mathcal{F}$ with range $\{0, \ldots, N-1\}$ and Natarajan dimension d_N,*

$$g_{\mathcal{F}}(m) \leq \sum_{i=0}^{d_N} \binom{m}{i} \binom{N}{2}^i \leq \begin{cases} N^{d_N+1}, & m < d_N + 1, \\ \left(\binom{N}{2} \frac{em}{d_N} \right)^{d_N}, & m \geq d_N + 1. \end{cases} \tag{77}$$

Proof. The first inequality follows from [45, Cor. 5] and the second follows from [38, Lemma 10]. □

Proof of Theorem 8. We begin with (26), the proof of which is similar to that of [17, Thm. 4.1], which, however, focuses on the VC dimension. Let $f(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z})$ denote the predictions of the algorithm on the supersample. Given $\widetilde{Z}$, the losses induced by the algorithm are a function of the predictions. Therefore, by the data-processing inequality, [43, Thm. 2.3.4]

$$I(\ell(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z}); S|\widetilde{Z}) \leq I(f(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z}); S|\widetilde{Z}). \tag{78}$$

Let $F(\widetilde{Z})$ denote the set of all possible values taken by $f(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z})$ by varying S and R but keeping $\widetilde{Z}$ fixed. Then,

$$I(f(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z}); S|\widetilde{Z}) \leq H(f(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z})|\widetilde{Z}) \tag{79}$$

$$\leq \sup_{\widetilde{Z}} \log \left| F(\widetilde{Z}) \right|. \tag{80}$$

Here, $H(f(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z})|\widetilde{Z})$ denotes the conditional entropy of $f(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z})$ given $\widetilde{Z}$. Since the learning algorithm implements a function from $\mathcal{F}$, and $\widetilde{Z}$ consists of $2n$ samples, we conclude that $\left|F(\widetilde{Z})\right|$ can be no larger than $g_{\mathcal{F}}(2n)$. Using Lemma 4, we note that, if $2n \geq d_N + 1$,

$$\sup_{\widetilde{Z}} \log \left| F(\widetilde{Z}) \right| \leq \log g_{\mathcal{F}}(2n) \leq d_N \log \left(\binom{N}{2} \frac{2en}{d_N} \right). \tag{81}$$

This concludes the proof of (26).

We now turn to (27). First, by Jensen's inequality,

$$D(P_{\lambda|\widetilde{Z}S} \,||\, P_{\lambda|\widetilde{Z}}) = \mathbb{E}_{P_{\lambda|\widetilde{Z}S}} \left[\log \frac{P_{\lambda|\widetilde{Z}S}}{P_{\lambda|\widetilde{Z}}} \right] \leq \log \mathbb{E}_{P_{\lambda|\widetilde{Z}S}} \left[\frac{P_{\lambda|\widetilde{Z}S}}{P_{\lambda|\widetilde{Z}}} \right]. \tag{82}$$

By Markov's inequality we conclude that, with probability at least $1 - \delta$ under the draw of $(\widetilde{Z}, S)$,

$$\log \mathbb{E}_{P_{\lambda|\widetilde{Z}S}} \left[\frac{P_{\lambda|\widetilde{Z}S}}{P_{\lambda|\widetilde{Z}}} \right] \leq \log \left(\frac{1}{\delta} \mathbb{E}_{P_{\lambda\widetilde{Z}S}} \left[\frac{P_{\lambda|\widetilde{Z}S}}{P_{\lambda|\widetilde{Z}}} \right] \right). \tag{83}$$

Since λ is a discrete random variable, its probability mass function is bounded by 1. Hence,

$$\log \left(\frac{1}{\delta} \mathbb{E}_{P_{\lambda\widetilde{Z}S}} \left[\frac{P_{\lambda|\widetilde{Z}S}}{P_{\lambda|\widetilde{Z}}} \right] \right) \leq \log \left(\frac{1}{\delta} \mathbb{E}_{P_{\lambda\widetilde{Z}}} \left[\frac{1}{P_{\lambda|\widetilde{Z}}} \right] \right). \tag{84}$$

By upper-bounding the average over $\widetilde{Z}$ by its supremum, we find that

$$\log \left(\frac{1}{\delta} \mathbb{E}_{P_{\lambda\widetilde{Z}}} \left[\frac{1}{P_{\lambda|\widetilde{Z}}} \right] \right) \leq \log \left(\frac{1}{\delta} \sup_{\widetilde{Z}} \mathbb{E}_{P_{\lambda|\widetilde{Z}}} \left[\frac{1}{P_{\lambda|\widetilde{Z}}} \right] \right). \tag{85}$$

Let $L(\widetilde{Z})$ denote the set of all possible values that λ can take given $\widetilde{Z}$. Then,

$$\log\left(\frac{1}{\delta}\sup_{\widetilde{Z}}\mathbb{E}_{P_{\lambda|\widetilde{Z}}}\left[\frac{1}{P_{\lambda|\widetilde{Z}}}\right]\right) = \log\left(\frac{1}{\delta}\sup_{\widetilde{Z}}\sum_{\lambda\in L(\widetilde{Z})} P_{\lambda|\widetilde{Z}}(\lambda)\frac{1}{P_{\lambda|\widetilde{Z}}(\lambda)}\right) \tag{86}$$

$$= \sup_{\widetilde{Z}}\log\frac{\left|L(\widetilde{Z})\right|}{\delta}. \tag{87}$$

Finally, since the map from $F(\widetilde{Z})$ to $L(\widetilde{Z})$ induced by $\ell(\cdot,\cdot)$ is surjective,

$$\sup_{\widetilde{Z}}\log\frac{\left|L(\widetilde{Z})\right|}{\delta} \leq \sup_{\widetilde{Z}}\log\frac{\left|F(\widetilde{Z})\right|}{\delta}. \tag{88}$$

Again, note that $\left|F(\widetilde{Z})\right|$ can be no larger than the growth function of $\mathcal{F}$ evaluated at $2n$, i.e., $g_{\mathcal{F}}(2n)$. Therefore, if $2n \geq d_N + 1$, it follows from Lemma 4 that

$$\sup_{\widetilde{Z}}\log\left|F(\widetilde{Z})\right| \leq \log g_{\mathcal{F}}(2n) \leq d_N \log\left(\binom{N}{2}\frac{2en}{d_N}\right). \tag{89}$$

The desired result now follows by combining (82)-(89). □

Proof of Corollary 1. By Jensen's inequality,

$$\frac{1}{n}\sum_{i=1}^{n}\sqrt{2I(\ell(\mathcal{A}(\widetilde{Z}_S,R),\widetilde{Z}_i);S_i|\widetilde{Z})} \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}2I(\ell(\mathcal{A}(\widetilde{Z}_S,R),\widetilde{Z}_i);S_i|\widetilde{Z})}. \tag{90}$$

From the independence of the S_i, it follows that

$$\frac{1}{n}\sum_{i=1}^{n}I(\ell(\mathcal{A}(\widetilde{Z}_S,R),\widetilde{Z}_i);S_i|\widetilde{Z}) \leq \frac{I(\ell(\mathcal{A}(\widetilde{Z}_S,R),\widetilde{Z});S|\widetilde{Z})}{n}. \tag{91}$$

We first combine (5), (90), and (91) to upper-bound the samplewise information term in (5) by its full-sample counterpart. Then, to establish the result in (28), we use (26) to upper-bound the full-sample e-CMI in terms of the Natarajan dimension.

To establish (29), we first use (25) and (27). Since both of these inequalities are probabilistic and hold with probability at least $1-\delta$, by the union bound, they both hold with probability at least $1-2\delta$. To obtain the statement in Corollary 1, we replace δ with $\delta/2$. □

B Additional Theoretical Results

In this section, we present some additional theoretical results. In Section B.1, we present an analogue of Theorem 4 given in terms of the regular mutual information rather than the e-CMI. In Section B.2, we show how to tighten the bounds in Theorem 4 and 5 by introducing affine transformations. In Section B.3, we present a version of Theorem 7 that holds with high probability also over the draw of R. Finally, in Section B.4, we performed a more detailed comparison of our bounds to results found in the literature.

B.1 Binary KL Bound with Samplewise Mutual Information

In this section, we derive a binary KL bound in terms of the samplewise mutual information between the learning algorithm's output and the training data. This can be seen as an average, samplewise version of what in the PAC-Bayesian literature is sometimes referred to as Seeger's bound [4, Sec. 3.2.3]. To the best of our knowledge, neither this samplewise bound nor its full-sample analogue have been explicitly stated in the literature, although the necessary ingredients for deriving the full-sample bound are already present in [29]. Note that, since we do not consider the CMI setting in Theorem 9, the definitions of $\hat{L}$ and $L_{\mathcal{D}}$ differ from the ones used in the rest of the paper.

Theorem 9 (Binary KL mutual information bound)**.** *Let Z be an n-dimensional vector with entries generated independently according to $\mathcal{D}$. Furthermore, let $\hat{L} = \mathbb{E}_{Z,R}[L_Z(\mathcal{A}, Z, R)]$ and $L_{\mathcal{D}} = \mathbb{E}_{Z,R}[L_{\mathcal{D}}(\mathcal{A}, Z, R)]$. Then,*

$$d(\hat{L} \,||\, L_{\mathcal{D}}) \leq \frac{1}{n} \sum_{i=1}^{n} I(\mathcal{A}(Z, R); Z_i) \tag{92}$$

where $d(\hat{L} \,||\, L_{\mathcal{D}})$ is the binary KL divergence, i.e., the KL divergence between two Bernoulli distributions with parameters $\hat{L}$ and $L_{\mathcal{D}}$, respectively.

Proof of Theorem 9. Let $F = \mathcal{A}(Z, R)$ denote the output of the learning algorithm, and let $L_{\mathcal{D}}(f) = \mathbb{E}_Z[\ell(f, Z)]$ with $Z \sim \mathcal{D}$. Let $Z' \sim \mathcal{D}$ be independent of F. For any fixed f, we have $\mathbb{E}_{Z'}[\ell(f, Z')] = L_{\mathcal{D}}(f)$. Therefore, by Lemma 2,

$$\mathbb{E}_{F,Z'}\left[e^{d_\gamma(\ell(F,Z') \,||\, L_{\mathcal{D}}(F))}\right] \leq 1. \tag{93}$$

We now set $X = F$, $Y = Z_i$, $g_1(F, Z_i) = \ell(F, Z_i)$, $g_2(F, Z_i) = L_{\mathcal{D}}(F)$, and $f_\gamma(\cdot, \cdot) = d_\gamma(\cdot \,||\, \cdot)$, and note that (93) implies that $\xi_\gamma \leq 0$ in Lemma 1. Hence, we conclude that

$$\sup_\gamma \mathbb{E}_{F,Z_i}[d_\gamma(\ell(F, Z_i) \,||\, L_{\mathcal{D}}(F))] \leq I(F; Z_i). \tag{94}$$

By decomposing the training loss as $L_Z(\mathcal{A}, Z, R) = \frac{1}{n}\sum_{i=1}^{n} \ell(F, Z_i)$ and using (14), we have

$$d(\mathbb{E}_{F,Z}[L_Z(\mathcal{A},Z,R)] \,||\, \mathbb{E}_{F,Z}[L_{\mathcal{D}}(F)]) = \sup_\gamma d_\gamma\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{F,Z_i}[\ell(F, Z_i)], \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{F,Z_i}[L_{\mathcal{D}}(F)]\right). \tag{95}$$

Since $d_\gamma(\cdot \,||\, \cdot)$ is jointly convex in its arguments, Jensen's inequality implies that

$$\sup_\gamma d_\gamma\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{F,Z_i}[\ell(F, Z_i]\,, \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{F,Z_i}[L_{\mathcal{D}}(F)]\right) \leq \sup_\gamma \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{F,Z_i}[d_\gamma(\ell(F, Z_i), L_{\mathcal{D}}(F))]. \tag{96}$$

Combining this with (94), we obtain

$$d(\mathbb{E}_{F,Z}[L_Z(\mathcal{A}, Z, R)] \;||\; \mathbb{E}_{F,Z}[L_{\mathcal{D}}(F)]) \leq \frac{1}{n}\sum_{i=1}^{n} I(F; Z_i) \tag{97}$$

from which the result follows. □

To convert this result into an upper bound on the population loss, one needs to numerically invert the binary KL divergence. This can be done by evaluating [46]

$$L_{\mathcal{D}} \leq \sup\left\{L'_{\mathcal{D}} \in [0, 1] : d(\hat{L} \,||\, L'_{\mathcal{D}}) \leq \frac{1}{n}\sum_{i=1}^{n} I(\mathcal{A}(Z, R); Z_i)\right\}. \tag{98}$$

We now rewrite (92) to make the connection to standard PAC-Bayesian results [27, 29] more apparent. Let $F = \mathcal{A}(Z, R)$, and let $P_{F|Z}$ denote the stochastic kernel that represents the randomized learning algorithm, that is, the PAC-Bayesian posterior. Furthermore, let P_F denote the corresponding marginal distribution, and let $P_Z = \mathcal{D}^n$ denote the data distribution. Finally, let $P_{F|Z_i}$ denote the resulting distribution when we marginalize $P_{F|Z}$ over all training examples except the ith. By the golden formula [47, Eq. (8.7)], we can upper-bound the mutual information by replacing the true marginal distribution in the KL divergence with some auxiliary distribution. Specifically,

$$I(F; Z_i) \leq D(P_{F|Z_i} P_{Z_i} \,||\, Q_F P_{Z_i}), \tag{99}$$

where Q_F is an arbitrary distribution on $\mathcal{F}$ satisfying $P_{F|Z_i}$ is absolutely continuous with respect to Q_F. Here, Q_F corresponds to the PAC-Bayesian prior. Thus, Theorem 9 implies that

$$d(\hat{L} \,||\, L_{\mathcal{D}}) \leq \frac{1}{n}\sum_{i=1}^{n} D(P_{F|Z_i} P_{Z_i} \,||\, Q_F P_{Z_i}). \tag{100}$$

Note that in all bounds reported in the rest of the paper, one can replace the true marginal with an auxiliary distribution. In some situations, this leads to more tractable bounds.

One commonly used approach to tighten PAC-Bayesian bounds is to consider data-dependent priors [25]. This can be achieved through techniques such as differential privacy [48] or data splitting, where the training samples are divided into one part used for evaluating the bound and one part for constructing the prior [49, 50, 51]. As noted by [18], the CMI setting can be seen as a way to automatically obtain data-dependent priors.

B.2 Affine Transformations of the Arguments in the Binary KL Bound

As mentioned in Section 2.3, the binary KL bounds in Theorem 4 and Theorem 5 can be tightened by considering affine transformations of the arguments of the arguments of $d(\cdot \,||\, \cdot)$. We present this in the following theorem.

Theorem 10 (Affine binary KL bounds)**.** *Let $g_{ab} : [0,1]^2 \to [0,1]$ be given by*

$$g_{ab}(x,y) = \frac{ax+by - \min(a,b,a+b,0)}{|b|+|a|}. \tag{101}$$

Furthermore, let

$$d_{ab}^{-1}(q,c) = \sup\left\{p \in [0,1] : d\Big(g_{ab}(q,p), g_{ab}\Big(\frac{q+p}{2}, \frac{q+p}{2}\Big)\Big) \leq c\right\}. \tag{102}$$

Consider the CMI setting. Then, for any a and b,

$$L_{\mathcal{D}} \leq d_{ab}^{-1}\Big(\hat{L}, \frac{1}{n}\sum_{i=1}^{n} I(\ell(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z}_i); S_i|\widetilde{Z})\Big). \tag{103}$$

Furthermore,

$$L_{\mathcal{D}} \leq \mathbb{E}_{\widetilde{Z}}\left[d_{ab}^{-1}\Big(\mathbb{E}_{R,S}\Big[L_{\widetilde{Z}_S}(\mathcal{A}, \widetilde{Z}_S, R)\Big], \frac{1}{n}\sum_{i=1}^{n} I^{\widetilde{Z}}(\ell(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z}_i); S_i)\Big)\right]. \tag{104}$$

Proof of Theorem 10. Fix $\tilde{z}$ and u. Let S' be a random variable with the same marginal distribution as S such that λ and S' are independent. We will use the same notation as in Corollary 2, and set $\lambda_u = \ell(\mathcal{A}(\tilde{z}_S, R), \tilde{z}_u)$, $\lambda_{S_u} = (\lambda_{u_1,S_{u_1}}, \ldots, \lambda_{u_m,S_{u_m}})$, and $\hat{\lambda}_{S_u} = \frac{1}{m}\sum_{i=1}^{m} \lambda_{u_i,S_{u_i}}$. Then,

$$\mathbb{E}_{S'}\Big[\hat{\lambda}_{S'_u}\Big] = \mathbb{E}_{S'}\Big[\hat{\lambda}_{\bar{S}'_u}\Big] = \frac{1}{m}\sum_{i=1}^{m} \frac{\lambda_{u_i,0} + \lambda_{u_i,1}}{2} = \hat{\lambda}_u. \tag{105}$$

Notice that, for any s, we have $\hat{\lambda}_u = (\hat{\lambda}_{s_u} + \hat{\lambda}_{\bar{s}_u})/2$. Since g_{ab} is linear in both of its arguments, it follows that

$$\mathbb{E}_{S'}\Big[g_{ab}\Big(\hat{\lambda}_{S'_u}, \hat{\lambda}_{\bar{S}'_u}\Big)\Big] = g_{ab}\Big(\hat{\lambda}_u, \hat{\lambda}_u\Big). \tag{106}$$

Thus, we can apply Lemma 2 with $\hat{\mu} = g_{ab}(\hat{\lambda}_{S'_u}, \hat{\lambda}_{\bar{S}'_u})$ and $\bar{\mu} = g_{ab}(\hat{\lambda}_u, \hat{\lambda}_u)$. Note that, since $\tilde{z}$ is fixed, the summands in $\hat{\mu}$ are independent but not identically distributed, and in particular, they do not have the same mean. This implies that [29, Eq. (17)] does not suffice and we need Lemma 2. It then follows from Corollary 2, applied with $f_\gamma(\hat{\lambda}_{S_u}, \hat{\lambda}_{\bar{S}_u}) = m d_\gamma\Big(g_{ab}\Big(\hat{\lambda}_{S_u}\Big), g_{ab}\Big(\hat{\lambda}_{\bar{S}_u}\Big)\Big)$, that

$$\sup_\gamma \mathbb{E}_{\lambda_u,S_u}\Big[m d_\gamma\Big(g_{ab}\Big(\hat{\lambda}_{S_u}, \hat{\lambda}_{\bar{S}_u}\Big), g_{ab}\Big(\hat{\lambda}_u, \hat{\lambda}_u\Big)\Big)\Big] \leq I^{\tilde{z},u}(\lambda_u; S_u). \tag{107}$$

By Jensen's inequality, we can move the expectation inside the jointly convex function $d_\gamma(\cdot \,||\, \cdot)$ and linear function g_{ab}, and then perform the optimization over γ, to get

$$d\Big(g_{ab}\Big(\mathbb{E}_{\lambda_u,S_u}\Big[\hat{\lambda}_{S_u}\Big], \mathbb{E}_{\lambda_u,S_u}\Big[\hat{\lambda}_{\bar{S}_u}\Big]\Big), g_{ab}\Big(\mathbb{E}_{\lambda_u,S_u}\Big[\hat{\lambda}_u\Big], \mathbb{E}_{\lambda_u,S_u}\Big[\hat{\lambda}_u\Big]\Big)\Big) \leq \frac{I^{\tilde{z},u}(\lambda_u; S_u)}{m}. \tag{108}$$

Finally, averaging over $\widetilde{Z}$ and U, replacing shorthands by their long forms, and again using Jensen's inequality to move the expectations inside $d(\cdot \,||\, \cdot)$ and g_{ab}, we get

$$d\left(g_{ab}\left(\hat{L}, L_{\mathcal{D}}\right) \,||\, g_{ab}\left(\frac{\hat{L}+L_{\mathcal{D}}}{2}, \frac{\hat{L}+L_{\mathcal{D}}}{2}\right)\right) \leq \frac{I(\ell(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z}_U); S_U|\widetilde{Z}, U)}{m}. \tag{109}$$

Since the right-hand side is an increasing function of the size m of the random subset U, the tightest bound is obtained by setting $m = 1$, from which the result in (103) follows.

To prove (104), we return to (108). By averaging over U and using Jensen's inequality to move this average inside the convex function $d(\cdot \,||\, \cdot)$, we find that

$$d\Big(g_{ab}\Big(\mathbb{E}_{U,\lambda_U,S_U}\left[\hat{\lambda}_{S_u}\right] \,||\, \mathbb{E}_{U,\lambda_U,S_U}\left[\hat{\lambda}_{\bar{S}_u}\right]\Big), g_{ab}\Big(\mathbb{E}_{U,\lambda_U,S_U}\left[\hat{\lambda}_u\right], \mathbb{E}_{U,\lambda_U,S_U}\left[\hat{\lambda}_u\right]\Big)\Big) \leq \frac{I^{\tilde{z}}(\lambda_U; S_U|U)}{m}. \tag{110}$$

By the definition of d_{ab}^{-1}, this implies that

$$\mathbb{E}_{U,\lambda_U,S_U}\left[\hat{\lambda}_{\bar{S}_u}\right] \leq d_{ab}^{-1}\left(\mathbb{E}_{U,\lambda_U,S_U}\left[\hat{\lambda}_{S_u}\right], \frac{I^{\tilde{z}}(\lambda_U; S_U|U)}{m}\right). \tag{111}$$

By averaging over $\widetilde{Z}$ and replacing shorthands with their long forms, we find that

$$\mathbb{E}_{\widetilde{Z},S,R}\Big[L_{\mathcal{D}}(\mathcal{A}, \widetilde{Z}_S, R)\Big] \leq \mathbb{E}_{\widetilde{Z}}\left[d_{ab}^{-1}\left(L_{\widetilde{Z}_S}(\mathcal{A}, \widetilde{Z}_S, R), \frac{I^{\widetilde{Z}}(\ell(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z}_U); S_U|U)}{m}\right]\right). \tag{112}$$

Since d_{ab}^{-1} is an increasing function of its second argument and $I^{\widetilde{Z}}(\ell(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z}_U); S_U|U)/m$ is increasing in m, the tightest bound is obtained with $m = 1$. With this choice, the result in (104) follows. □

Note that we can establish (18) and (20), hence proving Theorem 4 and 5, by setting $a = 1$ and $b = 0$ in (103) and (104) respectively. However, we can also optimize over a and b, which sometimes leads to tighter bounds. As an example, $d_{0,1}^{-1}(0.3, 0.125) < d_{1,0}^{-1}(0.3, 0.125)$. This shows that $a = 1$, $b = 0$ does not always give the tightest result.

B.3 Single-draw bound

As mentioned in Section 3, under a technical assumption of absolute continuity, it is possible to obtain a version of Theorem 7 that holds with high probability also over the draw of R. We present this result in the following theorem.

Theorem 11 (High-probability bounds with respect to R)**.** *Let $\lambda = \ell(\mathcal{A}(\widetilde{Z}_S, R), \widetilde{Z})$. Furthermore, let $P_{\lambda|\widetilde{Z}S}$ denote the conditional distribution of λ given $\widetilde{Z}$ and S, and let $P_{\lambda|\widetilde{Z}}$ denote the conditional distribution of λ given $\widetilde{Z}$. Let $\imath(\lambda, S|\widetilde{Z}) = \log P_{\lambda|\widetilde{Z}S}/P_{\lambda|\widetilde{Z}}$ denote the conditional information density between λ and S given $\widetilde{Z}$. Assume that $P_{\lambda|\widetilde{Z}}$ is absolutely continuous with respect to $P_{\lambda|\widetilde{Z}S}$. Then, with probability at least $1-\delta$ over the draw of $\widetilde{Z}$, S, and R,*

$$L_{\widetilde{Z}_{\bar{S}}}(\mathcal{A}, \widetilde{Z}_S, R) - L_{\widetilde{Z}_S}(\mathcal{A}, \widetilde{Z}_S, R) \leq \sqrt{\frac{2}{n-1}\left(\imath(\lambda, S|\widetilde{Z}) + \log\frac{\sqrt{n}}{\delta}\right)}. \tag{113}$$

Furthermore, also with probability at least $1-\delta$ over the draw of $\widetilde{Z}$, S and R,

$$d\left(L_{\widetilde{Z}_S}(\mathcal{A}, \widetilde{Z}_S, R) \,||\, \frac{L_{\widetilde{Z}_S}(\mathcal{A}, \widetilde{Z}_S, R) + L_{\widetilde{Z}_{\bar{S}}}(\mathcal{A}, \widetilde{Z}_S, R)}{2}\right) \leq \frac{\imath(\lambda, S|\widetilde{Z}) + \log\frac{2\sqrt{n}}{\delta}}{n}. \tag{114}$$

Proof. For any function $f_\gamma(\cdot,\cdot)$, define

$$\xi_\gamma = \log \mathbb{E}_{\lambda,\widetilde{Z},S'}\left[e^{f_\gamma(\hat{\lambda}_{S'}, \hat{\lambda}_{\bar{S}'})}\right] = \log \mathbb{E}_{\lambda',\widetilde{Z},S}\left[e^{f_\gamma(\hat{\lambda}_{S'}, \hat{\lambda}_{\bar{S}'})}\right], \tag{115}$$

where we used the observation that $\lambda, \widetilde{Z}, S'$ has the same distribution as $\lambda', \widetilde{Z}, S$. By our absolute continuity assumption, [43, Prop. 17.1] implies that

$$1 = \mathbb{E}_{\lambda,\widetilde{Z},S}\left[e^{f_\gamma(\hat{\lambda}_{S'},\hat{\lambda}_{\bar{S}'})-\imath(\lambda,S|\widetilde{Z})-\xi_\gamma}\right]. \tag{116}$$

Note that if $\widetilde{Z}$ and S are fixed, the randomness of λ is fully captured by R. By Markov's inequality, we conclude that, with probability at least $1-\delta$ under the draw of $\widetilde{Z}$, S and R,

$$e^{f_\gamma(\hat{\lambda}_{S'},\hat{\lambda}_{\bar{S}'})-\imath(\lambda,S|\widetilde{Z})-\xi_\gamma} \leq \log\frac{1}{\delta} \tag{117}$$

from which it follows that

$$f_\gamma(\hat{\lambda}_{S'},\hat{\lambda}_{\bar{S}'}) \leq \log\frac{1}{\delta} + \imath(\lambda,S|\widetilde{Z}) + \xi_\gamma. \tag{118}$$

We establish the result in (113) by setting $f_\gamma(\hat{\lambda}_{S'},\hat{\lambda}_{\bar{S}'}) = \frac{(n-1)}{2}(\hat{\lambda}_{\bar{S}'} - \hat{\lambda}_{S'})^2$ and using (72). Similarly, we establish the result in (114) by setting $f_\gamma(\hat{\lambda}_{S'},\hat{\lambda}_{\bar{S}'}) = nd(\hat{\lambda}_{S'} \,||\, (\hat{\lambda}_{S'} + \hat{\lambda}_{\bar{S}'})/2)$ and using (75).

□

The bounds in Theorem 11 are given in terms of the conditional information density [34]. Assuming that the learning algorithm implements a function from a class of bounded Natarajan dimension, the conditional information density can be bounded in a similar way as was done for the e-CMI and the KL divergence in Theorem 8. We present the resulting bound in the following theorem.

Theorem 12. *Consider a multiclass classification setting, for which $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the instance space and $\mathcal{Y}$ the label space, and assume that $|\mathcal{Y}| = N$. Furthermore, assume that the learning algorithm implements a function $f : \mathcal{X} \to \mathcal{Y}$ where $f \in \mathcal{F}$ belongs to a class of finite Natarajan dimension d_N [37]. Finally, assume that $2n \geq d_N + 1$. Then, with probability at least $1-\delta$ under the draw of $\widetilde{Z}$, S, and R,*

$$\imath(\lambda,S|\widetilde{Z}) \leq d_N \log\left(\binom{N}{2}\frac{2en}{d_N}\right) + \log\frac{1}{\delta}. \tag{119}$$

Proof. Note that if $\widetilde{Z}$ and S are fixed, the randomness of λ is fully captured by R. Thus, by Markov's inequality, we conclude that with probability at least $1-\delta$ under the draw of $\widetilde{Z}$, S, and R,

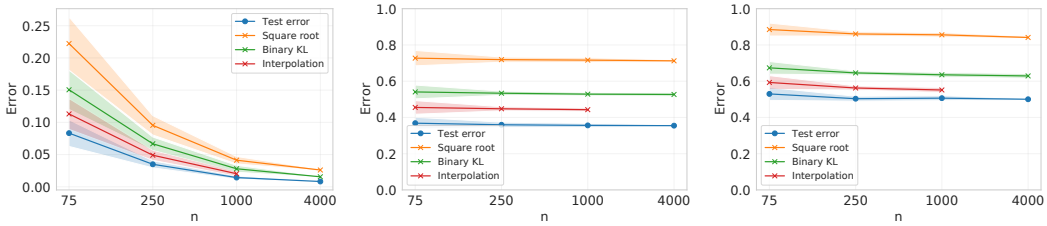
$$\imath(\lambda,S|\widetilde{Z}) = \log\frac{P_{\lambda|\widetilde{Z}S}}{P_{\lambda|\widetilde{Z}}} \leq \log\left(\frac{1}{\delta}\,\mathbb{E}_{P_{\lambda\widetilde{Z}S}}\left[\frac{P_{\lambda|\widetilde{Z}S}}{P_{\lambda|\widetilde{Z}}}\right]\right). \tag{120}$$

The right-hand side of (120) coincides with the right-hand side of (83). The desired result now follows by combining (120) and (84)-(89). □

B.4 Comparison to previous bounds

In this section, we compare the bounds in this paper to comparable results in from previous work. First, we perform a comparison with the results reported in [13], where a number of bounds that are functionally similar to ours are derived. The bounds that we derive in this paper are tighter due to the use of evaluated CMI. By the data-processing inequality for KL divergence [43, Thm. 2.2.6], our bounds can be relaxed to obtain bounds with the ordinary CMI in place of the evaluated CMI, demonstrating that they compare favorably to those of [13]. The same argument holds for the high-probability bounds.

Next, we compare (24) to the bound in [35, Corollary 1]. In terms of convergence rates, the bound in [35, Corollary 1] interpolates between our square-root and linear bounds, where the specific rate depends on the parameter β^* in the Bernstein condition in [35]. For the case of $0/1$-loss, we have that $B = 4$ and $\beta^* = 0$ in the Bernstein condition. Thus, [35, Corollary 1] has the same $1/\sqrt{n}$ rate as our square-root bound. Quantitatively, one expects our square-root bound to be tighter for several reasons: i) the bound in [35, Corollary 1] includes a constant $1/\eta_{\max} > 28.8$ that multiplies the KL divergence and $\log 1/\delta$ terms, ii) the bound in [35, Corollary 1] includes an extra $\eta_{\max}/(4n)$ term, and



(a) No label corruption (b) 50% label corruption (c) 100% label corruption

Figure 3: Numerical evaluation of the test error for the binarized MNIST setting considered in Figure 2a, but with varying degrees of label corruption, along with the upper bounds provided by the square-root bound in (4), the binary KL bound in (20) and, when applicable, the interpolation bound in (12).

(a) CIFAR10, SGLD (b) CIFAR10, SGLD (zoomed)

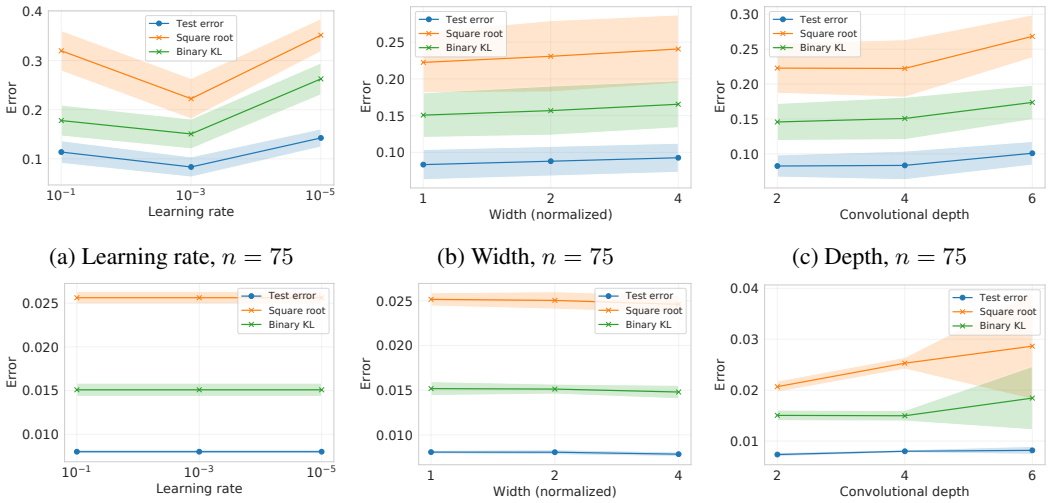
Figure 4: Numerical evaluation of the test error for a pre-trained ResNet-50 fine-tuned with SGLD on CIFAR-10, along with the upper bounds provided by the square-root bound in (4), the binary KL bound in (20), and the SGLD bound in [11, Eq. 6].

iii) the $\log 1/\delta$ term in the bound in [35, Corollary 1] appears outside of the square-root containing the KL divergence. To perform a direct quantitative comparison, some simplifying assumptions are needed. First, while the data-processing inequality implies that the KL divergence in (24) is always smaller than or equal to the KL divergence in the bound of [35, Corollary 1], we shall assume that both KL divergences equal 1. Furthermore, we set $\delta = 0.01$ and $n = 1000$. Under these assumptions, the bound in [35, Corollary 1] gives a generalization gap of approximately 2.89, which is vacuous, whereas the bound in (24) gives a generalization gap of approximately 0.13. This discrepancy arises mainly due to the large constants described above, and holds for other reasonable values of the parameters involved.

C Additional Numerical Results

In this section, we present some additional numerical results for deep learning settings.

To study a scenario with heavy overfitting, we repeat the experiment with SGD on the binarized MNIST data set, as presented in Figure 2a, but with varying degrees of label randomization. Specifically, for each sample in the binarized MNIST data set, we consider a random variable $C \sim \mathrm{Bern}(a)$, where a is the probability of corruption. If $C = 1$, the label is set to either 4 or 9, picked uniformly at random. If $C = 0$, we leave the label unchanged. For all levels of label corruption, the networks reached training errors of zero or near zero. In Figure 3a, we consider $a = 0$, so no labels are corrupted. This is the same as Figure 2a, and is included only for reference. In Figure 3b, we consider $a = 0.5$. Finally, in Figure 3c, we set $a = 1$, so that all labels are corrupted. Thus, for this scenario, the training set carries no relevant information, and the classifier completely overfits to the data set. For all levels of corruptions, the interpolation bound in (12) gives the tightest bound when applicable. Furthermore, the binary KL bound in (20) is tighter than the square-root bound in (4). The bounds give somewhat reasonable estimates of the test error. In contrast, the generalization bounds for randomized labels reported in previous works are vacuous [46, 13].



(a) Learning rate, $n = 75$ (b) Width, $n = 75$ (c) Depth, $n = 75$

(d) Learning rate, $n = 4000$ (e) Width, $n = 4000$ (f) Depth, $n = 4000$

Figure 5: Numerical evaluation of the test error for the binarized MNIST setting considered in Figure 2a, but with varying learning rate, width, and depth of the network, along with the upper bounds provided by the square-root bound in (4) and the binary KL bound in (20). For the figures in the first row, we fix $n = 75$, and for the second row, we fix $n = 4000$.

In Figure 4, we consider SGLD, as in Figure 2c, but for CIFAR10 instead of the binarized MNIST data set. Again, the binary KL bound in (20) gives the tightest result. Unlike for the binarized MNIST scenario, the SGLD bound in [11, Eq. 6] is not the tightest during early epochs.

Finally, to study the robustness of our bounds to various hyperparameter changes, we again repeat the experiment with SGD on the binarized MNIST data set, but with varying learning rate, network width, and network depth. First, we consider the case of $n = 75$, since this leads to higher test errors, making variations in it more noticeable. Second, we consider $n = 4000$ to examine whether the same qualitative conclusions hold for a larger sample size. This is shown in Figure 5. When varying the width in Figure 5b, the number of units in each of the convolutional layers of the original network, described in Table 1, is multiplied by the factor given on the horizontal axis. When varying the depth in Figure 5c, the convolutional depth of 4 corresponds to the original network, described in Table 1, the depth of 2 is the same network but with the two first convolutional layers removed, and the depth of 6 is the same network but with two additional convolutional layers with 64 units, 3×3 size, a stride of 1, and padding 1. The bounds seem to correlate well with the test error across these scenarios, and they correctly predict the values for each hyperparameter that lead to the lowest and highest test errors. This effect is noticeable for $n = 75$, where there is significant variation between the different hyperparameter values. For $n = 4000$, this variation is heavily reduced, but the overall behavior of our bounds is still consistent with the behavior of the test error.

D Experimental Details

In this section, we describe the training procedures, network architectures, and experimental details used in Section 6. Note that the setups are the same as those considered in [17]. The experiments were run on Google Colab Pro GPUs.

In all experiments, we estimate the test error of the networks and the upper bounds provided by the square-root bound in (4) and the binary KL bound in (20) (and [11, Eq. (6)] for Figure 2c) by drawing k_1 samples of $\widetilde{Z}$, which consists of n pairs of samples drawn randomly from the corresponding dataset, and k_2 samples of S (and R for Figure 2c). We run the training algorithm for each of these samples, compute the training loss, and estimate the population loss using $\tilde{z}_{\bar{s}}$. For each $\tilde{z}$, we estimate the mutual information $I^{\tilde{z}}(\ell(\mathcal{A}(\tilde{z}_S, R), \tilde{z}_i); S_i)$ using a plug-in estimator [52]. Based on these estimates, the bounds are computed. The results in Figure 2 illustrate the estimated averages, with shaded areas indicating one standard deviation.

Table 1: The neural network used for the binarized MNIST experiments in Figure 2a and Figure 2c.

CONV. LAYER, 32 UNITS, 4×4 SIZE, STRIDE 2, PADDING 1, BATCH NORM., RELU ACTIVATION
CONV. LAYER, 32 UNITS, 4×4 SIZE, STRIDE 2, PADDING 1, BATCH NORM., RELU ACTIVATION
CONV. LAYER, 64 UNITS, 3×3 SIZE, STRIDE 2, PADDING 0, BATCH NORM., RELU ACTIVATION
CONV. LAYER, 256 UNITS, 3×3 SIZE, STRIDE 1, PADDING 0, BATCH NORM., RELU ACTIVATION
FULLY CONNECTED LAYER, 128 UNITS, RELU ACTIVATION
FULLY CONNECTED LAYER, 2 UNITS, LINEAR ACTIVATION

D.1 Binarized MNIST

For Figure 2a and 2c, we consider the network described in Table 1. The dataset that we use consist of the parts of MNIST that corresponds to the digits 4 and 9. Both Figure 2a and Figure 2c are based on 5 samples of $\widetilde{Z}$, with 30 samples of S for each $\tilde{z}$.

D.1.1 Adam

For Figure 2a, we optimize the network using the Adam algorithm, with a 0.001 learning rate and $\beta_1 = 0.9$, for 200 epochs with a batch size of 128.

D.1.2 SGLD

For Figure 2c, we optimize the network using SGLD for 40 epochs with a batch size of 100. The learning rate starts at 0.01 and decays by a factor of 0.9 after each 100 iterations. The inverse temperature schedule is given by $\min(4000, \max(100, 10e^{t/100}))$ where t is the iteration.

D.2 CIFAR10 and SGD

For Figure 2b, we consider the network ResNet-50, which is pretrained on Imagenet. Then, we fine-tune it on CIFAR10 using SGD with a 0.01 learning rate and 0.9 momentum for 40 epochs with a batch size of 64. During training, we use random horizontal flips and random 28×28 cropping as data augmentations. The values in the figure are based on 2 samples of $\widetilde{Z}$, with 40 samples of S for each $\tilde{z}$.