A New Family of Generalization Bounds Using Samplewise Evaluated CMI

Anonymous Author(s)
Affiliation
Address
email

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 showing that they 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 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 training data and hypothesis. These bounds have since been extended in several ways [7, 8, 9, 10, 11, 12, 13].

A major step was taken in [14], which considered a setting where the training set is randomly selected from a larger supersample. 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, [14] also derives bounds in terms of the information encoded in these losses rather than in the hypothesis itself, a quantity they refer to as *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

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

exploited by [15], 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 information measures based on the hypothesis space tend to suffer from overfitting as training progresses [11, 16], the bound of [15] 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, 13, 14, 15, 16]. 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, [17, 18, 19, 20] 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 [21, 22] consider any jointly convex function. Finally, [23] 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 has also been shown 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 [15, 24]. Tightening and extending 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 [15]; ii) a linear bound (Theorem 3) that tightens the bound given in [16]; 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 [20]. 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). Numerical experiments conducted on MNIST and CIFAR10 reveal that the binary KL bound results in a tighter characterization of the population loss compared to other two bounds for some deep learning scenarios. 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). Finally, 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).

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, denoted by $P \ll Q$. 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 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 $I^z(X;Y) = D(P_{XY|Z=z} \,||\, P_{X|Z=z}P_{Y|Z=z})$. This quantity is referred to as the *disintegrated* mutual information in [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 i.i.d. according to $\mathcal{D}$. For convenience, we index the columns of $\widetilde{Z}$ starting from 0 and the rows starting from 1. Furthermore, we denote the ith row of $\widetilde{Z}$ as $\widetilde{Z}_i$. Let $S = (S_1, \ldots, S_n)$ denote a membership vector, with entries generated independently

[1] While [15] consider 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.

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 $Z' \sim \mathcal{D}$, and the training loss as $L_{\tilde{z}_s}(\mathcal{A}, \tilde{z}_s, r) = \frac{1}{n}\sum_{i=1}^{n} \ell(\mathcal{A}(\tilde{z}_s, r), [\tilde{z}_s]_i)$. 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 [21] and [23], gives us a generic framework to derive generalization bounds that allows for a wide choice of functions for comparing training loss and population loss. A crucial difference compared to [21] and [23] 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 [25, 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 $\xi_\gamma \leq 0$. 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 [26]). 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 [15, 16]. 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 vector 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 $\mathcal{A}(\tilde{z}_S, R) = W \in \mathcal{W}$ are the parameters of a function $\phi_W : \mathcal{X} \to \mathcal{Y}$. Let $\tilde{x}$ denote the $m \times 2$ matrix obtained by projecting each element of $\tilde{z}$ 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,*

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

We now use our framework to recover some of the results reported in [15]. 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 $\Delta \in [-1,1]$. 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,*

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

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

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

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

However, since this expression is increasing in the size m of the random subset U [15, 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 [16], which are samplewise extensions of the bounds in [14]. 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 and population loss. To this end, similar to [25, 27], 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 [28, Lemma 1], [29, Thm. 3], and [27, Eq. (17)] (which is a generalization of [25, 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 [30, 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' again is an independent copy of S_i. Thus, we have

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

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 [20, 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, 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]. However, as noted after Lemma 2, the concentration inequality needed to establish Theorem 4 is a refinement of the result used in [18]. 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 and interpolation bounds 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 [31, 32, 33]. 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,*

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

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

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

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 [32, 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 [34, 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 [15, 24], 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).

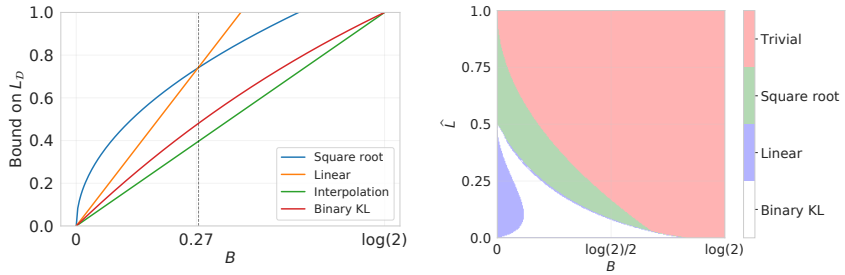
Theorem 8. *Consider a multiclass classification setting, for which $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the instance space and $\mathcal{Y}$ the label space, and assume that $|\mathcal{Y}| = N$. Furthermore, assume that the learning algorithm implements a function $f : \mathcal{X} \to \mathcal{Y}$ where $f \in \mathcal{F}$ belongs to a class of finite Natarajan dimension d_N [35]. Finally, assume that $2n > d_N + 1$. Then,*

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

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

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

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). The resulting bounds essentially recover previously established generalization guarantees in terms of the Natarajan dimension. We discuss this in more detail and present explicit examples in Appendix B.



(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 $L_{\mathcal{D}} \leq 1$.

5 Comparing the Bounds

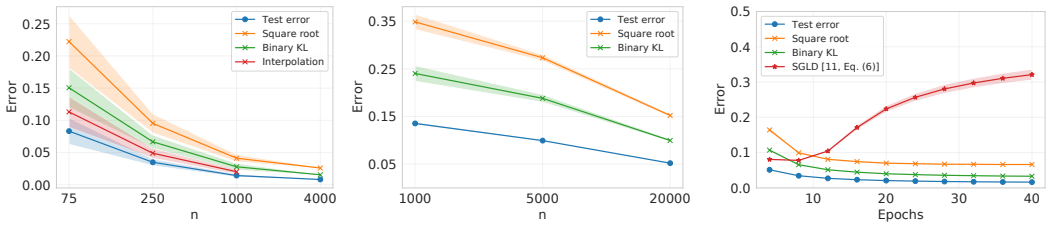
We now perform some comparisons between the bounds obtained in Section 2. As it turns out, there is no clear ordering between the different bounds in general. For a first comparison, we consider the interpolating setting, where the training loss is zero. We focus on the case where the size m of the random 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) from the comparison, 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 a qualitative 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 $B < 2\gamma_{1,\text{opt}}^2$. 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 $B > 2\gamma_{1,\text{opt}}^2 \approx 0.27$, the ordering between the square-root bound in* (11) *and the linear bound in* (5) *is inverted.*

While Proposition 1 gives the order 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 get 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.



(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].

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 deep learning scenarios, such as CNNs trained on MNIST [36] or CIFAR10 [37], are found in [15]. These numerical results are based on [15, 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 [15].[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.

Since the previous section indicated that the interpolation bound in (12) and the binary KL bound in (20) are the superior bounds for low-training loss settings, which is common for deep learning, we focus on these bounds, as well as the square-root bound in (4) as a comparison. 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 [15], the information contained in the predictions (or the losses they induce) can lead to nonvacuous bounds. 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. This is shown in Figure 2a. In binary classification there is a one-to-one mapping between predictions and losses, so for this scenario, the information captured by the vector of losses is the same as that in the vector of 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 that leads to the tightest characterization of the test error. However, for $n = 4000$, the training error was non-zero for some neural networks, albeit very small.

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 potentially is 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 by [15], 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.

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

References

[1] John Shawe-Taylor and Robert C. Williamson. A PAC analysis of a Bayesian estimator. In *Proc. Conf. Learn. Theory (COLT)*, COLT '97, 1997.

[2] David A. McAllester. Some PAC-Bayesian theorems. In *Proc. Conf. Learn. Theory (COLT)*, Madison, WI, USA, 7 1998.

[3] B. Guedj. A primer on PAC-Bayesian learning. *arXiv*, 1 2019.

[4] Pierre Alquier. User-friendly introduction to PAC-Bayes bounds. *arXiv*, 11 2021.

[5] D. Russo and J. Zou. Controlling bias in adaptive data analysis using information theory. In *Proc. Artif. Intell. Statist. (AISTATS)*, Cadiz, Spain, 5 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, 12 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, 12 2018.

[8] Hassan Hafez-Kolahi, Zeinab Golgooni, Shohreh Kasaei, and Mahdieh Soleymani. Conditioning and processing: Techniques to improve information-theoretic generalization bounds. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, volume 33, Vancouver, Canada, 12 2020.

[9] Borja Rodríguez-Gálvez, Germán Bassi, Ragnar Thobaben, and Mikael Skoglund. On random subset generalization error bounds and the stochastic gradient langevin dynamics algorithm. In *Inf. Theory Workshop (ITW)*, Riva del Garda, Italy, 4 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, 5 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, 12 2019.

[12] 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, 7 2021.

[13] Ruida Zhou, Chao Tian, and Tie Liu. Individually conditional individual mutual information bound on generalization error. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Melbourne, Australia, 7 2021.

[14] T. Steinke and L. Zakynthinou. Reasoning about generalization via conditional mutual information. In *Proc. Conf. Learn. Theory (COLT)*, Graz, Austria, 7 2020.

[15] Hrayr Harutyunyan, Maxim Raginsky, Greg Ver Steeg, and Aram Galstyan. Information-theoretic generalization bounds for black-box learning algorithms. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Virtual Conference, 12 2021.

[16] 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, 7 2021.

[17] John Langford and Matthias Seeger. Bounds for averaging classifiers. *CMU Technical report*, CMU-CS-01-102, 2001.

[18] M. Seeger. PAC-Bayesian generalisation error bounds for Gaussian process classification. *J. of Mach. Learn. Res.*, 3:233–269, 10 2002.

[19] John Langford. *Quantitatively Tight Sample Complexity Bounds*. PhD thesis, Carnegie Mellon University, 2002.

[20] Andreas Maurer. A note on the PAC Bayesian theorem, 11 2004.

[21] Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayesian learning of linear classifiers. In *Proc. Int. Conf. Mach. Learning (ICML)*, Montreal, Canada, 06 2009.

[22] Luc Bégin, Pascal Germain, François Laviolette, and Jean-Francis Roy. PAC-Bayesian theory for transductive learning. In *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, volume 33, Reykjavik, Iceland, 04 2014.

[23] Omar Rivasplata, Ilja Kuzborskij, Csaba Szepesvari, and John Shawe-Taylor. PAC-Bayes analysis beyond the usual bounds. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, pages 16833–16845, Vancouver, Canada, 12 2020.

[24] Mahdi Haghifam, Gintare Karolina Dziugaite, Shay Moran, and Daniel M. Roy. Towards a unified information-theoretic framework for generalization, 2021. NeurIPS21.

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

[26] Maxime Haddouche, Benjamin Guedj, Omar Rivasplata, and John Shawe-Taylor. PAC-Bayes unleashed: Generalisation bounds with unbounded losses. *Entropy*, 23(10), 10 2021.

[27] D.A. McAllester. A PAC-Bayesian tutorial with a dropout bound. *arXiv*, 7 2013.

[28] Yevgeny Seldin, François Laviolette, Nicolò Cesa-Bianchi, John Shawe-Taylor, and Peter Auer. PAC-Bayesian inequalities for martingales. *IEEE Trans. Inf. Theory*, 58(12):7086–7093, 8 2012.

[29] Wassily Hoeffding. On the distribution of the number of successes in independent trials. *The Annals of Mathematical Statistics*, 27(3):713–721, 9 1956.

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

[31] Zakaria Mhammedi, Peter Grünwald, and Benjamin Guedj. PAC-Bayes un-expected Bernstein inequality. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, volume 32, Vancouver, Canada, 12 2019.

[32] F. Hellström and G. Durisi. Generalization bounds via information density and conditional information density. *IEEE J. Sel. Areas Inf. Theory*, 1(3):824–839, 12 2020.

[33] 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, 8 2021.

[34] Hao Wang, Yizhe Huang, Rui Gao, and Flavio Calmon. Analyzing the generalization capability of sgld using properties of gaussian channels. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Virtual Conference, 12 2021.

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

[36] Li Deng. The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

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

[38] 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, 1 1975.

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

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

[41] 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, 8 2017.

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

[43] Gintare Karolina Dziugaite and Daniel M Roy. Data-dependent PAC-Bayes priors via differential privacy. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Montreal, Canada, 12 2018.

[44] Amiran Ambroladze, Emilio Parrado-Hernandez, and John Shawe-Taylor. Tighter PAC-Bayes bounds. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, pages 9–16, Vancouver, Canada, 12 2006.

[45] G.K. Dziugaite, K. Hsu, W. Gharbieh, and D.M. Roy. On the role of data in PAC-Bayes bounds, 6 2020.

[46] Maria Perez-Ortiz, Omar Risvaplata, John Shawe-Taylor, and Csaba Szepesvari. Tighter risk certificates for neural networks. *J. of Mach. Learn. Res.*, 22(227), 8 2021.

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

[48] David Haussler and Philip M Long. A generalization of sauer's lemma. *Journal of Combinatorial Theory, Series A*, 71(2):219–240, 1995.

[49] Yann Guermeur. Large Margin Multi-category Discriminant Models and Scale-sensitive Psi-dimensions. Research report, INRIA, 2004.

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 C.
   (b) Did you mention the license of the assets? [Yes] The license is discussed in Appendix C.
   (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]