Evaluated CMI Bounds for Meta Learning: Tightness and Expressiveness

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent work has established that the conditional mutual information (CMI) framework of Steinke and Zakynthinou (2020) is expressive enough to capture generalization guarantees in terms of algorithmic stability, VC dimension, and related complexity measures for conventional learning (Harutyunyan et al., 2021, Haghifam et al., 2021). Hence, it provides a unified method for establishing generalization bounds. In meta learning, there has so far been a divide between information-theoretic results and results from classical learning theory. In this work, we take a first step toward bridging this divide. Specifically, we present novel generalization bounds for meta learning in terms of the evaluated CMI (e-CMI). To demonstrate the expressiveness of the e-CMI framework, we apply our bounds to a representation learning setting, with n samples from $\hat{n}$ tasks parameterized by functions of the form $f_i \circ h$. Here, each $f_i(\cdot)$ is a task-specific function, and $h(\cdot)$ is the shared representation. For this setup, we show that the e-CMI framework yields a bound that scales as $\sqrt{\mathcal{C}(\mathcal{H})/(n\hat{n}) + \mathcal{C}(\mathcal{F})/n}$, where $\mathcal{C}(\cdot)$ denotes a complexity measure of the hypothesis class. This scaling behavior coincides with the one reported in Tripuraneni et al. (2020) using Gaussian complexity.

1 Introduction

Meta learning, sometimes referred to as learning to learn, is a process by which performance on a new machine learning task is increased by using knowledge acquired from separate, but related, tasks [1, 2]. Concretely, the meta learner $\hat{\mathcal{A}}$ has access to training data from several different tasks, which are embedded in a common task environment, and aims to extract information from this data. The goal is to use this information to improve the performance of a base learner $\mathcal{A}$ on a new task from the same task environment. For instance, the task environment can consist of different image classification tasks, and the goal of the meta learner can be to learn a shared representation for the tasks or to find suitable hyperparameters for the learning algorithm.

As in conventional learning, a central goal in meta learning is to bound the gap between the loss on the training data and the population risk on unseen data. The two current approaches for achieving this goal are to: i) use techniques from classical learning theory to obtain minimax performance guarantees, or ii) to use information-theoretic methods to obtain algorithm-, data- and distribution-dependent guarantees. So far, these two lines of work have evolved largely separately. In this paper, we take some steps toward unifying them. Specifically, we: i) derive new, tighter information-theoretic generalization bounds for meta learning, and ii) demonstrate that these bounds are expressive enough to recover bounds for meta learning from classical learning theory. To concretize the discussion in this introduction, we assume that the meta learner outputs a member h of a function class $\mathcal{H}$ based

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

on n samples from $\hat{n}$ different tasks, and that a base learner selects a member f of a class $\mathcal{F}$, based on the output of the meta learner and n samples from a given task.

Classical learning theory for meta learning. The theoretical analysis of the benefits of meta learning in terms of loss bounds dates back to [3], where the notion of task environment was formally introduced. More recently, for the setting of representation learning, [4, Thm. 5] derived a risk bound that scales as[1] $\sqrt{\mathcal{C}(\mathcal{H})/\hat{n}} + \sqrt{\mathcal{C}(\mathcal{F})/n}$, where $\mathcal{C}(\cdot)$ denotes a complexity measure of the function class. This demonstrates the benefit of meta learning for tasks that share a common environment. Indeed, in the conventional single-task learning scenario, the n samples from a given task need to be used for learning h and f simultaneously, leading to a $\sqrt{\mathcal{C}(\mathcal{H} \times \mathcal{F})/n}$ bound. The bound provided in [4, Thm. 5] was later improved by [5] to a scaling of $\sqrt{\mathcal{C}(\mathcal{H})/(n\hat{n}) + \mathcal{C}(\mathcal{F})/n}$. This improved scaling, where $\mathcal{C}(\mathcal{H})$ decays with the product $n\hat{n}$, confirms the intuition that all of the $n\hat{n}$ samples that are observed are informative at the environment level. Meta learning has also been extensively studied in several special cases. For instance, [6, 7, 8] study a setting with linear features and task mappings, while [9, 10] consider an online convex optimization setting. In this paper, we will mainly focus on the representation learning setting.

Information-theoretic generalization bounds. For conventional learning, the study of information-theoretic bounds was initiated by [11, 12], where the average generalization gap of a learning algorithm is bounded in terms of the information that the algorithm reveals about the training data. At its heart, this line of work relies on a change of measure technique that relates the training loss to the population loss. While the first information-theoretic bounds were given in terms of the mutual information between the algorithms output and the full training data, recent work provides bounds in terms of the *disintegrated* mutual information between the *loss* that the algorithm incurs on a *single* sample pair and a selection variable indicating which sample is used for training, given a supersample containing both the training and test data. These developments are due to the samplewise approach of [13], the disintegration introduced by [14], the evaluated conditional mutual information (e-CMI) notion from [15], and combinations and extensions of these from [16, 17, 18, 19, 20]. This line of work is also intimately related to PAC-Bayesian generalization bounds [21, 22], where the generalization gap, averaged over the learning algorithm, is bounded with high probability over the data in terms of a KL divergence. This is explored further in [23, 24].

Information-theoretic analysis of meta learning. Recently, information-theoretic generalization bounds have also been applied to meta learning [25, 26, 27]. In parallel, a PAC-Bayesian analysis of meta learning has also been developed [28, 29, 30, 31, 32, 33]. Generalization bounds obtained via information-theoretic methods have also been used as training objectives in order to improve performance [30, 34]. The quantity of interest in meta learning is the meta-population loss, which is the population loss evaluated on a task that was not observed during the meta learning phase. While this quantity is unknown, the meta learner has indirect information about it through the observed meta-training loss, which is the loss that the meta learner incurs on the training samples from each of the observed tasks during the meta learning phase. The standard approach in the information-theoretic and PAC-Bayesian analysis of meta learning consists of two steps. The first step involves bounding the difference between the training loss and a suitably defined auxiliary loss. The second step involves bounding the difference between the meta-population loss and the auxiliary loss. The two natural candidates for this auxiliary loss are the population loss of an observed task, and the training loss for an unobserved task. One of these steps (the first or second, depending on the choice of the auxiliary loss) is purely at the task level, while the other is purely at the environment level. This makes it possible to view each of these steps as a conventional learning problem, so that a standard information-theoretic generalization bound can be applied. By the use of the triangle inequality, these two results are then combined to obtain a bound on the meta-population loss in terms of the meta-training loss. We will refer to this procedure as a *two-step* derivation. An alternative approach was recently used by [26], where a *one-step* procedure was employed. Rather than relying on an auxiliary loss, [26] immediately bound the difference between the meta-population loss and the meta-training loss in terms of a mutual information that captures both task level and environment level dependencies.

[1] In the interest of brevity, we suppress logarithmic factors throughout this section.

The environment and task level dependencies can then be obtained by decomposing this mutual information. The resulting bound turns out to have a better scaling with $\hat{n}$ than the two-step bounds. However, the information-theoretic analyses of meta learning reviewed so far do not provide any rigorous characterization of the scaling behavior of the bounds. In particular, the dependence of the information measures on the sample size is typically ignored. This precludes a direct comparison between these information-theoretic bounds and classical learning theory results.

Contributions. Focusing on the meta learning setup, we present novel information-theoretic bounds based on the e-CMI framework, and demonstrate how to recover minimax results from classical learning theory via these bounds. Our specific contributions are as follows. In Section 3.1, we derive bounds for the average generalization error in terms of the disintegrated, samplewise e-CMI of the meta learner and base learner: in Theorem 1 and 2, we provide square-root bounds, which are shown to be tighter than results in the literature; in Theorem 3, we derive novel bounds in terms of the binary KL divergence. For low values of the training loss, the binary KL bounds display a more favorable dependence on the number of data samples than the square-root bounds. Next, in Section 3.2, we extend these average bounds to obtain high-probability generalization guarantees. This is necessary to perform comparisons with high-probability bounds from classical learning theory. Finally, in Section 4, we demonstrate the expressiveness of our bounds by applying them to a representation learning setting. Under certain assumptions about the hypothesis classes, we provide upper bounds on the information measures that appear in our bound in terms of complexity measures. The results that we obtain via this procedure display a scaling behavior that coincides with the one reported in [5]. This demonstrates that the e-CMI framework is expressive enough to recover the scaling behavior of generalization guarantees for meta learning obtained via classical learning theory.

2 Problem Setup and Notation

We now introduce the CMI setup for meta learning and the necessary notation for stating our results. Similar to [3], we consider a task environment formulation that includes the representation learning setting of [4, 5] as a special case.

Our meta learning setup involves the following quantities. We consider a task distribution $\widehat{\mathcal{D}}$ on the task space $\widehat{T}$. For a given task $\tau \in \widehat{T}$, there is a corresponding in-task distribution $\mathcal{D}_\tau$ on the sample space $\mathcal{Z}$. The objective of the meta learner is to output a meta hypothesis $U \in \mathcal{U}$. This is done on the basis of n samples from $\hat{n}$ tasks. Formally, the meta learner is a mapping $\hat{\mathcal{A}} : \mathcal{Z}^{n \times \hat{n}} \times \widehat{\mathcal{R}} \to \mathcal{U}$, where a random variable $\hat{R} \in \widehat{\mathcal{R}}$ captures the potential stochasticity of the algorithm. The goal of the base learner is to output a hypothesis $W \in \mathcal{W}$, given the output of the meta learner and n samples from a specific task. Formally, the base learner is a mapping $\mathcal{A} : \mathcal{Z}^n \times \mathcal{R} \times \mathcal{U} \to \mathcal{W}$, where the vector of randomnesses for each task is $R = (R_1, \ldots, R_{\hat{n}}) \in \mathcal{R}^{\hat{n}}$. These randomnesses are independent from the data, and are assumed to be identically distributed.[2] Here, the spaces $\mathcal{U}$ and $\mathcal{W}$ may be function spaces or parameter spaces, depending on the learning algorithms.

We now describe how the training data is obtained from a meta-supersample, in accordance with the meta-learning extension in [35] of the conditional mutual information (CMI) framework of [15]. In this setting, the meta-training set is formed as follows. First, $\hat{n}$ tasks are randomly selected from a set of $2\hat{n}$ tasks. Then, for each task, n samples are randomly selected from a set of $2n$ samples. This setup, as well as our notation, is illustrated in Figure 1 for an example with $\hat{n} = 2$ task pairs and $n = 4$ sample pairs for each task. Formally, the meta-supersample $\hat{Z}$ can be viewed as a data structure with $4n\hat{n}$ elements. In Figure 1, we illustrate this as an $n \times 4\hat{n}$ matrix. We decompose $\hat{Z}$ as $(Z^1, \ldots, Z^{\hat{n}})$, where each element can be seen as a task pair. Specifically, for $i = 1, \ldots, \hat{n}$, the pair Z^i can be decomposed as $(Z^{i,0}, Z^{i,1})$, where each element is a task-specific supersample. For $k = 0, 1$, the task-specific super sample is $Z^{i,k} = (Z_1^{i,k}, \ldots, Z_n^{i,k})$, where each element now is a pair of data samples. Specifically, for $j = 1, \ldots, n$, each sample pair is $Z_j^{i,k} = (Z_{j,0}^{i,k}, Z_{j,1}^{i,k})$, where $Z_{j,l}^{i,k} \in \mathcal{Z}$ for $l = 0, 1$. The elements of $\hat{Z}$ are generated as follows. First, for $i = 1, \ldots, \hat{n}$

[2] While identical distributions are not strictly necessary for our results, it simplifies the presentation.

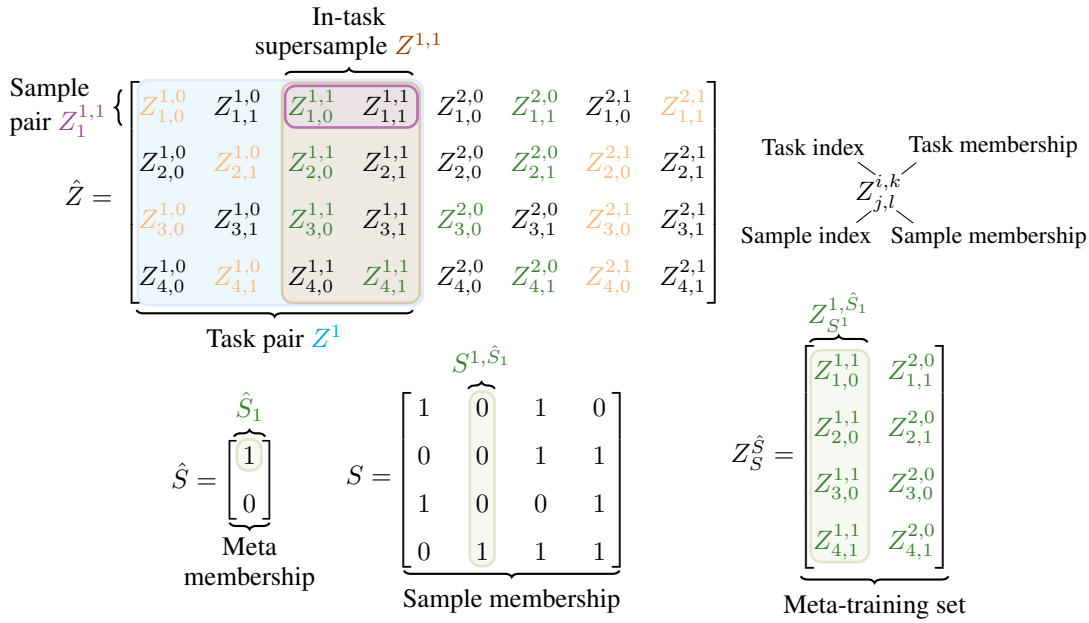

Figure 1: A graphical representation of our notation and the CMI setup for meta learning. In this example, the meta-supersample contains the two task pairs Z^1, which is marked in blue, and Z^2. In turn, Z^1 consists of the two in-task supersamples $Z^{1,0}$ and $Z^{1,1}$, which is marked in brown. Next, $Z^{1,1}$ consists of the four sample pairs $Z_j^{1,1}$, where $j = 1, \ldots, 4$. In the image, $Z_1^{1,1}$ is marked in purple. Finally, $Z_1^{1,1}$ consists of a pair of samples, $Z_{1,0}^{1,1}$ and $Z_{1,1}^{1,1}$. When a binary vector is used as subscript or superscript, this indicates that we should enumerate the meta-supersample according to that vector. To illustrate this, consider the construction of the meta-training set $Z_S^{\hat{S}}$. The meta-training set is formed on the basis of the meta-subset choice $\hat{S}$ and the observed-task subset choice $S^{\hat{S}}$. For instance, from the first task-pair, $\hat{S}_1$ indicates that we should select task 1. Then, from the first sample pair in this task, $S^{1,\hat{S}_1}$ indicates that we should select sample 0, which is $Z_{1,0}^{1,1}$, marked in green. Repeating this for each sample pair in the in-task supersample $Z^{1,1}$, we can identify the remaining elements of $Z_{S^1}^{1,\hat{S}_1}$. This procedure is performed for each task index until all samples of $Z_S^{\hat{S}}$ are identified. The meta-test set $Z_{-S}^{-\hat{S}}$ is formed by an analogous procedure, but is now based on $-\hat{S}$ and $-S^{-\hat{S}}$. The entries from the meta-supersample that are selected for $Z_{-S}^{-\hat{S}}$ are marked in orange. For $\hat{\lambda}$, which contains the losses induced on the meta-supersample, we use an analogous notation.

and $k = 0, 1$, we generate $\tau_{i,k} \sim \widehat{\mathcal{D}}$. Then, for each $j = 1, \ldots, n$ and $k, l = 0, 1$, we generate the samples $Z_{j,l}^{i,k} \sim \mathcal{D}_{\tau_{i,k}}$ independently of each other.

Next, we describe how the training data is selected from the meta-supersample $\hat{Z}$. This is done on the basis of the meta-membership vector $\hat{S}$ and the in-task membership vector S. Specifically, the meta-membership vector $\hat{S} = (\hat{S}_1, \ldots, \hat{S}_{\hat{n}})$ is an $\hat{n}$-dimensional vector, with each element drawn independently from a $\mathrm{Bern}(1/2)$ distribution. Furthermore, the sample membership vector $S = (S^{1,0}, S^{1,1}, \ldots, S^{\hat{n},0}, S^{\hat{n},1})$ is a collection of $2\hat{n}$ vectors, where each $S^{i,k} = (S_1^{i,k}, \ldots, S_n^{i,k})$ is an n-dimensional vector with elements generated independently from a $\mathrm{Bern}(1/2)$ distribution. For any Bernoulli matrix X, we let $-X$ denote its elementwise complement modulo 2, i.e., $\mathbf{1} - X$, where $\mathbf{1}$ is the all-one matrix.

These membership vectors are used to form the meta-training set as follows. We use the convention that, when a binary vector is used as a subscript or superscript of Z, this indicates that we should enumerate over this vector. Thus, for $i = 1, \ldots, \hat{n}$ and $k = 0, 1$, we let $Z_{S^{i,k}}^{i,k} = (Z_{1,S_1^{i,k}}^{i,k}, \ldots, Z_{n,S_n^{i,k}}^{i,k})$.

Here, $Z^{i,k}_{S^{i,k}}$ is the training set for the (i,k)th task. To reduce clutter, we will often use the shorthands $Z^{i,k}_{S^i} = Z^{i,k}_{S^{i,k}}$ and $Z^{i,k}_{1,S^i_j} = Z^{i,k}_{1,S^{i,k}_j}$. In an analogous way, the test set for the (i,k)th task $Z^{i,k}_{-S^i}$ is constructed similarly, but on the basis of $-S^{i,k}$. The full meta-training set is given by $Z^{\hat{S}}_S = (Z^{1,\hat{S}_1}_{S^{1,\hat{S}_i}}, \ldots, Z^{\hat{n},\hat{S}_{\hat{n}}}_{S^{\hat{n},\hat{S}_{\hat{n}}}})$, where each in-task training set is constructed as indicated above. Again, the meta-test set $Z^{-\hat{S}}_{-S}$ is formed by an analogous procedure.

We now describe how the performance of the meta learner and base learner is evaluated. We denote the output of the meta learner as $U = \hat{\mathcal{A}}(Z^{\hat{S}}_S, \hat{R})$, and the output of the base learner for task (i,k) as $W^{i,k} = \mathcal{A}(Z^{i,k}_{S^i}, R_i, U) \in \mathcal{F}$, where $\mathcal{F}$ is a function space. The performance of the learners is evaluated through a loss function $\ell : \mathcal{F} \times \mathcal{Z} \to [0,1]$. We denote the losses that the meta learner and base learner induce on the meta-supersample $\hat{Z}$ by $\hat{\lambda}$, which inherits the subscript and superscript notation that we described for $\hat{Z}$. Thus, for $i = 1, \ldots, \hat{n}$, $j = 1, \ldots, n$, and $k, l = 0, 1$, we have $\lambda^{i,k}_{j,l} = \ell(W^{i,k}, z^{i,k}_{j,l})$. In other words, $\lambda^{i,k}_{j,l}$ is the loss on the (j,l)th sample in the (i,k)th task.

On the basis of the loss matrix $\hat{\lambda}$ and the membership vectors $\hat{S}$ and S, we can compute four different losses. The main quantity that we are interested in bounding is the average meta-population loss $L_{\widehat{\mathcal{D}}}$, which is the loss on test data for unobserved tasks. The quantity that the meta learner has access to is the average meta-training loss, $\widehat{L}$, which is the training loss for observed tasks. The other two losses are the average auxiliary test loss, $\bar{L}$, which is the loss on test data for observed tasks, and the average auxiliary training loss $\widetilde{L}$, which is the loss on training data for unobserved tasks. In the two-step derivations, one of these two quantities are used as the auxiliary loss. These four losses are given by

$$L_{\widehat{\mathcal{D}}} = \frac{1}{n\hat{n}} \sum_{i,j=1}^{\hat{n},n} \mathbb{E}_{\lambda^i_j, \hat{Z}} \left[\lambda^{i,-\hat{S}_i}_{j,-S^i_j} \right], \qquad \widehat{L} = \frac{1}{n\hat{n}} \sum_{i,j=1}^{\hat{n},n} \mathbb{E}_{\lambda^i_j, \hat{Z}} \left[\lambda^{i,\hat{S}_i}_{j,S^i_j} \right] \tag{1}$$

$$\bar{L} = \frac{1}{n\hat{n}} \sum_{i,j=1}^{\hat{n},n} \mathbb{E}_{\lambda^i_j, \hat{Z}} \left[\lambda^{i,\hat{S}_i}_{j,-S^i_j} \right], \qquad \widetilde{L} = \frac{1}{n\hat{n}} \sum_{i,j=1}^{\hat{n},n} \mathbb{E}_{\lambda^i_j, \hat{Z}} \left[\lambda^{i,-\hat{S}_i}_{j,S^i_j} \right]. \tag{2}$$

Finally, we end this section by introducing some information-theoretic quantities that appear in our bounds. First, let P and Q be two probability measures such that P is absolutely continuous with respect to Q. The KL divergence between P and Q is denoted by $D(P \,||\, Q)$. For the special case where P and Q are Bernoulli distributions with parameters p and q, we let

$$d(p \,||\, q) = D(P \,||\, Q) = p \log\left(\frac{q}{p}\right) + (1-p) \log\left(\frac{1-p}{1-q}\right). \tag{3}$$

We refer to $d(p \,||\, q)$ as the binary KL divergence. The mutual information between the random variables X and Y is given by $I(X;Y) = D(P_{XY} \,||\, P_X P_Y)$, where P_{XY} is their joint distribution and P_X and P_Y are the corresponding marginals. The disintegrated mutual information between X and Y given a third random variable Z is given by $I^Z(X;Y) = D(P_{XY|Z} \,||\, P_{X|Z} P_{Y|Z})$, where $P_{XY|Z}$ is the joint distribution of X and Y given Z and $P_{X|Z} P_{Y|Z}$ is the product distribution formed from the corresponding marginals. The expectation over Z of the disintegrated mutual information is the conditional mutual information $I(X;Y|Z) = \mathbb{E}_Z\big[I^Z(X;Y)\big]$.

3 Generalization Bounds for Meta Learning with e-CMI

In this section, we present generalization bounds in terms of the e-CMI of the meta learner and base learner. In Section 3.1, we derive average square-root bounds that tighten results from [26, 35], as well as novel binary KL bounds. In Section 3.2, we extend these results to obtain novel, high-probability information-theoretic bounds for meta learning. In Section 4, we will demonstrate the expressiveness of the e-CMI framework by using the bounds from this section to recover generalization guarantees from classical learning theory for representation learning.

3.1 Average Bounds

In Theorem 1, we present a square-root bound for the average generalization error obtained through a two-step derivation. Specifically, one step consists of bounding the unobserved training loss $\widetilde{L}$ in terms of the observed training loss $\widehat{L}$, and the second step bounds the meta-population loss $L_{\widehat{\mathcal{D}}}$ in terms of $\widetilde{L}$. Chaining these two bounds, we obtain a bound on $L_{\widehat{\mathcal{D}}}$ in terms of $\widehat{L}$. The bound depends on the information captured by two random variables: the task-level variable $\lambda_j^{i,-\hat{S}_i}$, which contains the training loss and test loss for task $(i, -\hat{S}_i)$, as well as the environment-level variable $\lambda_{j,S_j^i}^i$, which contains the training losses for both the observed task $(i, \hat{S}_i)$ and the unobserved task $(i, -\hat{S}_i)$. We provide the proof of this result in Appendix A, along with the proofs of all other results in this paper.

Theorem 1 (Two-step square-root bound)**.** *Consider the setup described in Section 2. Then,*

$$\left|L_{\widehat{\mathcal{D}}} - \widehat{L}\right| \leq \frac{1}{n\hat{n}} \sum_{i,j=1}^{\hat{n},n} \mathbb{E}_{\hat{Z},S_j^i}\left[\sqrt{2I^{\hat{Z},S_j^i}(\lambda_{j,S_j^i}^i; \hat{S}_i)}\right] + \frac{1}{n\hat{n}} \sum_{i,j=1}^{\hat{n},n} \mathbb{E}_{\hat{Z},\hat{S}_i}\left[\sqrt{2I^{\hat{Z},\hat{S}_i}(\lambda_j^{i,-\hat{S}_i}; S_j^{i,-\hat{S}_i})}\right]. \quad (4)$$

The first term captures the environment-level generalization error while the second term captures the task-level generalization error. In order to clarify the relation between Theorem 1 and results from the literature, we relax it by upper-bounding the disintegrated individual-sample e-CMI terms by their integrated, full-sample, parametric CMI counterparts.

Corollary 1. *Theorem 1 implies that*

$$\left|L_{\widehat{\mathcal{D}}} - \widehat{L}\right| \leq \sqrt{\frac{2I(U; \hat{S}|\hat{Z}, S)}{\hat{n}}} + \sqrt{\frac{2I(W^{1,-\hat{S}_1}; S^{1,-S_1}|\hat{Z}, \hat{S}_1)}{n}}. \quad (5)$$

This recovers the result of [35, Thm. 1], demonstrating that Theorem 4 is tighter.

Next, we present an alternative square-root bound that is obtained through a one-step derivation. This bound depends on the information captured by the random variable λ_j^i, which contains the training and test loss for both the observed and unobserved tasks.

Theorem 2 (One-step square-root bound)**.** *Consider the setup described in Section 2. Then,*

$$\left|L_{\widehat{\mathcal{D}}} - \widehat{L}\right| \leq \frac{1}{n\hat{n}} \sum_{i,j=1}^{\hat{n},n} \mathbb{E}_{\hat{Z}}\left[\sqrt{2I^{\hat{Z}}(\lambda_j^i; \hat{S}_i, S_j^i)}\right]. \quad (6)$$

Again, to compare this bound to results in the literature, we relax it by upper-bounding the disintegrated individual-sample e-CMI terms by their integrated, full-sample, parametric counterparts.

Corollary 2. *Let $W^i = (W^{i,0}, W^{i,1})$ and $\hat{W} = \{W^i\}_{i=1}^{\hat{n}}$. Then,*

$$\left|L_{\widehat{\mathcal{D}}} - \widehat{L}\right| \leq \sqrt{\frac{2I(\hat{W}; \hat{S}, S|\hat{Z})}{n\hat{n}}} \leq \sqrt{\frac{2I(U; \hat{S}, S|\hat{Z}) + 2\hat{n}I(W^1; S^1|\hat{Z}, U)}{n\hat{n}}} \quad (7)$$

$$\leq \sqrt{\frac{2I(U; Z_S^{\hat{S}}) + 2\hat{n}I(W^1; Z_{S^1}^1|U)}{n\hat{n}}}. \quad (8)$$

Up to some constant factors, this recovers the result in [26, Thm. 5.1]. Note that, if $\hat{\mathcal{A}}$ or $\mathcal{A}$ are deterministic learning algorithms with continuous outputs, the mutual information terms in (8) are unbounded. In contrast, the CMI terms in (7) are always finite. This is discussed in more detail in [15].

Finally, in Theorem 3, we present two novel bounds in terms of the binary KL divergence. The advantage of these bounds, as compared to the square-root bounds in Theorem 1 and 2, is that for low training losses, they have a more favorable dependence on the number of samples. We demonstrate this improved rate for representation learning in Section 4.

Theorem 3 (Binary KL bounds)**.** *For $m > 1$, we let*

$$d_m^{-1}(q, c) = \sup\left\{p \in [0,1] : d\left(q, \frac{q+p}{m}\right) \leq c\right\}. \tag{9}$$

Then,

$$L_{\widehat{\mathcal{D}}} \leq d_2^{-1}\left(d_2^{-1}\left(\widehat{L}, \frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} I(\lambda^i_{j,S^i_j}; \hat{S}_i|\hat{Z}, S^i_j)\right), \frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} I(\lambda^{i,-\hat{S}_i}_j; S^{i,-\hat{S}_i}_j|\hat{Z}, \hat{S}_i)\right). \tag{10}$$

Furthermore,

$$L_{\widehat{\mathcal{D}}} + \bar{L} + \widetilde{L} \leq d_4^{-1}\left(\widehat{L}, \frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} I(\lambda^i_j; \hat{S}_i, S^i_j|\hat{Z})\right). \tag{11}$$

Interestingly, (11) provides a bound on the sum of the average meta-population loss $L_{\widehat{\mathcal{D}}}$, the test loss on observed tasks $\bar{L}$, and the training loss on unobserved tasks $\widetilde{L}$. Due to the nonnegativity of the loss, we can obtain an explicit bound on $L_{\widehat{\mathcal{D}}}$ by using the lower bound $\bar{L}, \widetilde{L} \geq 0$, which is a sensible relaxation when $L_{\widehat{\mathcal{D}}}$ is the dominant term. By this relaxation, we weaken the bound at most by a constant factor. As previously mentioned, the bounds in Theorem 3 can have a more favorable dependence on the number of samples than the square-root bounds in Theorem 1 and 2 when the training loss is low. In the following corollary, we present a bound on $L_{\widehat{\mathcal{D}}}$ for the case where $\widehat{L} = 0$.

Corollary 3. *Assume that $\widehat{L} = 0$. Then, Theorem 3 implies that*

$$L_{\widehat{\mathcal{D}}} \leq 4 - 4\exp\left(-\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} I(\lambda^i_j; \hat{S}_i, S^i_j|\hat{Z})\right) \leq \frac{4}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} I(\lambda^i_j; \hat{S}_i, S^i_j|\hat{Z}). \tag{12}$$

Compared to the bound in Theorem 2, there is no square root in Corollary 3. As we show in Section 4, this can lead to a faster rate of decay with the number of samples.

3.2 High-probability Bounds

In the previous section, we provided bounds on the average generalization error. However, meta learning bounds obtained via classical learning theory are typically high-probability bounds [4, 5]. In order to assess the expressiveness of the e-CMI framework in terms of its ability to recover these results, we now extend the bounds from Section 3.1 to the high-probability setting. For this, we need some additional notation. We let $L_{\widehat{\mathcal{D}}}(\hat{Z}, \hat{S}, S)$ and $\widehat{L}(\hat{Z}, \hat{S}, S)$ denote the meta-population loss and training loss given that the meta-training set is constructed from $(\hat{Z}, \hat{S}, S)$. Specifically,

$$L_{\widehat{\mathcal{D}}}(\hat{Z}, \hat{S}, S) = \frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} \mathbb{E}_{\hat{R},R_i}\left[\ell(\mathcal{A}(Z^{i,-\hat{S}_i}_{S^i}, R_i, \hat{\mathcal{A}}(Z^{\hat{S}}_S, \hat{R})), Z^{i,-\hat{S}_i}_{j,-S^i_j})\right], \tag{13}$$

$$\widehat{L}(\hat{Z}, \hat{S}, S) = \frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} \mathbb{E}_{\hat{R},R_i}\left[\ell(\mathcal{A}(Z^{i,\hat{S}_i}_{S^i}, R_i, \hat{\mathcal{A}}(Z^{\hat{S}}_S, \hat{R})), Z^{i,\hat{S}_i}_{j,S^i_j})\right]. \tag{14}$$

We now present a high-probability version of the two-step square root bound in Theorem 1. To simplify the presentation, we omit explicit constants and assume that $\hat{\mathcal{A}}$ is indifferent to the order of the data samples. The theorem statement is provided in more general form in Appendix A.

Theorem 4 (High-probability two-step square-root bound)**.** *Let $\hat{Q}^1_\lambda$ denote the conditional distribution of λ_{1,S_1} given $(\hat{Z}, \hat{S}, S)$, and let $\hat{P}^1_\lambda$ denote $\mathbb{E}_{\hat{S}}\left[\hat{Q}^1_\lambda\right]$. Furthermore, let Q^1_λ denote the conditional distribution of $\lambda^{1,-\hat{S}_1}$ given $(\hat{Z}, \hat{S}_1, S^{1,-S_1})$, and let P^1_λ denote $\mathbb{E}_{S^{1,-S_1}}\left[Q^1_\lambda\right]$. Then, there exist*

constants C_1, C_2 such that, with probability at least $1-\delta$ under the draw of $(\hat{Z}, \hat{S}, S)$,

$$\left|L_{\widehat{\mathcal{D}}}(\hat{Z},\hat{S},S)-\widehat{L}(\hat{Z},\hat{S},S)\right| \leq C_1\sqrt{\frac{D(\hat{Q}^1_\lambda \,||\, \hat{P}^1_\lambda)+\log(\frac{n\sqrt{\hat{n}}}{\delta})}{\hat{n}}}+C_2\sqrt{\frac{D(Q^1_\lambda \,||\, P^1_\lambda)+\log(\frac{\sqrt{n}}{\delta})}{n}}. \quad (15)$$

The KL divergences in (15) can be interpreted as pointwise e-CMIs. Indeed,

$$\mathbb{E}_{\hat{Z},\hat{S},S}\left[D(\hat{Q}^1_\lambda \,||\, \hat{P}^1_\lambda)\right] = I(\lambda_{1,S_1}; \hat{S}|S,\hat{Z}), \quad (16)$$

$$\mathbb{E}_{\hat{Z},\hat{S},S}\left[D(Q^1_\lambda \,||\, P^1_\lambda))\right] = I(\lambda^{1,-\hat{S}_1}; S^{1,-S_1}|\hat{Z},\hat{S}_1). \quad (17)$$

Finally, we present a high-probability version of the one-step square root bound in Theorem 2.

Theorem 5 (High-probability one-step square-root bound)**.** *Let $\hat{Q}_\lambda$ denote the conditional distribution of $\hat{\lambda}$ given $(\hat{Z}, \hat{S}, S)$, and let $\hat{P}_\lambda$ denote $\mathbb{E}_{\hat{S},S}\left[\hat{Q}_\lambda\right]$. Then, with probability at least $1-\delta$ under the draw of $(\hat{Z}, \hat{S}, S)$,*

$$\left|L_{\widehat{\mathcal{D}}}(\hat{Z},\hat{S},S)-\widehat{L}(\hat{Z},\hat{S},S)\right| \leq \sqrt{\frac{2\left(D(\hat{Q}_\lambda \,||\, \hat{P}_\lambda)+\log(\frac{\sqrt{n\hat{n}}}{\delta})\right)}{n\hat{n}-1}}. \quad (18)$$

Again, the KL divergence can be interpreted as a pointwise e-CMI, since

$$\mathbb{E}_{\hat{Z},\hat{S},S}\left[D(\hat{Q}_\lambda \,||\, \hat{P}_\lambda)\right] = I(\hat{\lambda}; \hat{S}, S|\hat{Z}). \quad (19)$$

4 Expressiveness of the Bounds

In Section 3, we presented several new information-theoretic generalization bounds, and demonstrated that they improve upon known bounds from the literature. We now turn our focus to the expressiveness of the e-CMI framework. In particular, we show that the bounds from Section 3 can be used to recover generalization guarantees for meta learning from classical learning theory.

Specifically, we consider the representation learning setting that is analyzed in [5]. We use the following notation. First, the sample space is split into an instance space and label space as $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. The aim of the meta learner is to find a representation $h_U \in \mathcal{H}$, while the base learner outputs a task-specific function $f_W \in \mathcal{F}$. Composing these functions leads to the mapping $f_W \circ h_U : \mathcal{X} \to \mathcal{Y}$.

To obtain explicit minimax bounds, we will assume that $\mathcal{H}$ has finite Natarajan dimension d_N and that $\mathcal{F}$ has finite VC dimension d_{VC}. This allows us to derive bounds on the entropy of the representations and predictions that the meta learner and base learner induce on the meta-supersample. This, in turn, leads to bounds on the e-CMI terms that appear in the bounds in Section 3. In the following corollary, we present the bounds that are obtained by bounding the e-CMI terms in Theorem 1 and 2.

Corollary 4. *Assume that the range of $\mathcal{H}$ has cardinality N, that the Natarajan dimension of $\mathcal{H}$ is d_N, and that the VC dimension of $\mathcal{F}$ is d_{VC}. Also, assume that $2n \geq d_{\text{VC}}+1$ and $2\hat{n} \geq d_N+1$. Then,*

$$\left|L_{\widehat{\mathcal{D}}}-\widehat{L}\right| \leq \sqrt{\frac{2d_N \log\left(\binom{N}{2}\frac{2e\hat{n}}{d_N}\right)}{\hat{n}}}+\sqrt{\frac{2d_{\text{VC}}\log\left(\frac{2en}{d_{\text{VC}}}\right)}{n}}, \quad (20)$$

$$\left|L_{\widehat{\mathcal{D}}}-\widehat{L}\right| \leq \sqrt{\frac{2d_N \log\left(\binom{N}{2}\frac{4en\hat{n}}{d_N}\right)+4\hat{n}d_{\text{VC}}\log\left(\frac{2en}{d_{\text{VC}}}\right)}{n\hat{n}}}. \quad (21)$$

Corollary 4 establishes that, for the average setting, we can use the bounds in Theorems 1 and 2 to recover known bounds for function classes with bounded Natarajan and VC dimensions. Note that, in the upper-bound of (20), we have fully decoupled the complexity of the two function classes. This is

made possible by the fact that we used $\widetilde{L}$ as the auxiliary loss in the derivation of Theorem 1, rather than $\bar{L}$. We discuss this in more detail in Appendix A.

Next, we consider the interpolating setting, where $\widehat{L} = 0$. Under this assumption, we demonstrate that we can achieve a better rate of convergence with respect to the number of training samples. The result, presented in the following corollary, relies on similarly bounding the e-CMI term in Corollary 3.

Corollary 5. *Consider the setting of Corollary 4. Furthermore, assume that $\widehat{L} = 0$. Then,*

$$L_{\widehat{\mathcal{D}}} \leq \frac{4d_N \log\left(\binom{N}{2}\frac{4en\hat{n}}{d_N}\right) + 8\hat{n}d_{\text{VC}}\log\left(\frac{2en}{d_{\text{VC}}}\right)}{n\hat{n}}. \tag{22}$$

The result in Corollary 5 demonstrates that, for the interpolating setting, the e-CMI framework is expressive enough to yield a bound that, ignoring logarithmic factors, decays as $1/(n\hat{n})$, often referred to as a fast rate.

Finally, noting that the bounds in [4] and [5] are high-probability rather than average, we also derive high-probability generalization bounds. In order to achieve this, we need probabilistic upper-bounds on the KL divergences that appear in Theorem 4 and 5, similar to how the e-CMI terms were bounded for Corollary 4 and 5. The resulting bounds are presented in the following corollary.

Corollary 6. *Consider the setting of Corollary 4. Then, there exist constants C_1, C_2, C_3 such that, with probability at least $1-\delta$ under the draw of $(\hat{Z}, \hat{S}, S)$,*

$$\left|L_{\widehat{\mathcal{D}}}(\hat{Z},\hat{S},S) - \widehat{L}(\hat{Z},\hat{S},S)\right| \leq C_1\sqrt{\frac{d_N \log\left(\binom{N}{2}\frac{\hat{n}}{d_N}\right) + \log\left(\frac{n\sqrt{\hat{n}}}{\delta}\right)}{\hat{n}}} + C_2\sqrt{\frac{d_{\text{VC}}\log\left(\frac{n}{d_{\text{VC}}}\right) + \log\left(\frac{\sqrt{n}}{\delta}\right)}{n}}, \tag{23}$$

$$\left|L_{\widehat{\mathcal{D}}}(\hat{Z},\hat{S},S) - \widehat{L}(\hat{Z},\hat{S},S)\right| \leq C_3\sqrt{\frac{d_N \log\left(\binom{N}{2}\frac{n\hat{n}}{d_N}\right) + \hat{n}d_{\text{VC}}\log\left(\frac{n}{d_{\text{VC}}}\right) + \log\left(\frac{\sqrt{n\hat{n}}}{\delta}\right)}{n\hat{n}}}. \tag{24}$$

We now see that, suppressing logarithmic factors, the upper bound in (23) scales as $\sqrt{\mathcal{C}(\mathcal{H})/\hat{n}} + \sqrt{\mathcal{C}(\mathcal{F})/n}$, whereas the upper bound in (24) scales as $\sqrt{\mathcal{C}(\mathcal{H})/(n\hat{n}) + \mathcal{C}(\mathcal{F})/n}$. This matches the rates obtained by [4] and [5], respectively, demonstrating that the e-CMI framework combined with the one-step approach is expressive enough to recover the scaling of these results.

Note that there are some differences between our results and the ones in [4, 5]. First, while the complexity measures that we use are related to the Natarajan and VC dimension, [5] consider Gaussian complexity. Furthermore, while we obtain generalization bounds for a generic new task, [5] consider excess risk bounds for a fixed target task with m training samples. To handle this, [5] introduces the notion of task diversity, which is not needed in our analysis. In Appendix B, we extend our analysis to derive excess risk bounds for empirical risk minimizers.

5 Conclusions

In this paper, we derived new generalization bounds for meta learning using e-CMI, improving upon information-theoretic bounds found in the literature. By considering a representation learning setting, we demonstrated that e-CMI bounds obtained via a conventional two-step approach lead to rates that coincide with those found in [4]. In contrast, we showed that by combining the e-CMI framework with a one-step approach, we recover the more favourable scaling found in [5]. Note that, while the bounds in [5] are uniform over the hypothesis class, the information-theoretic bounds that we derive are inherently algorithm- and data-dependent. As a consequence, they are nonvacuous when applied to settings such as classification with deep neural networks [20]. It should also be noted that the complexity measures that we consider differ from the Gaussian complexity in [5]. An intriguing topic for further study is to clarify the connection between e-CMI and Gaussian complexity.

References

[1] Rich Caruana. Multitask learning. *Mach. Learn.*, 28(1):41–75, 7 1997.

[2] Sebastian Thrun and Lorien Pratt. *Learning to Learn: Introduction and Overview*. Springer US, Boston, MA, 1998.

[3] Jonathan Baxter. A model of inductive bias learning. *J. Artif. Int. Res.*, 12(1):149–198, 3 2000.

[4] Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *J. Mach. Learn. Res.*, 17(1):2853–2884, 1 2016.

[5] Nilesh Tripuraneni, Michael Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, volume 33, Vancouver, Canada, 12 2020.

[6] Karim Lounici, Massimiliano Pontil, Sara van de Geer, and Alexandre B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4):2164 – 2204, 8 2011.

[7] Giovanni Cavallanti, Nicoló Cesa-Bianchi, and Claudio Gentile. Linear algorithms for online multitask classification. *Journal of Machine Learning Research*, 11(97):2901–2934, 10 2010.

[8] Massimiliano Pontil and Andreas Maurer. Excess risk bounds for multitask learning with trace norm regularization. In *Proc. Conf. Learn. Theory (COLT)*, pages 55–76, Princeton, NJ, USA, 6 2013.

[9] Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *Proc. Int. Conf. Mach. Learning (ICML)*, Jul 2018.

[10] Maria-Florina Balcan, Mikhail Khodak, and Ameet Talwalkar. Provable guarantees for gradient-based meta-learning. In *Proc. Int. Conf. Mach. Learning (ICML)*, Long Beach, CA, USA, Jun 2019.

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

[12] 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.

[13] 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.

[14] 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.

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

[16] 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. *arXiv*, 4 2020.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

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

[23] 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.

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

[25] Sharu Theresa Jose and Osvaldo Simeone. Information-theoretic generalization bounds for meta-learning and applications. *Entropy*, 23(1), 1 2021.

[26] Qi CHEN, Changjian Shui, and Mario Marchand. Generalization bounds for meta-learning: An information-theoretic analysis. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, volume 34, Virtual Conference, 12 2021.

[27] Sharu Theresa Jose, Osvaldo Simeone, and Giuseppe Durisi. Transfer meta-learning: Information- theoretic bounds and information meta-risk minimization. *IEEE Trans. Inf. Theor.*, 68(1):474–501, 1 2022.

[28] Anastasia Pentina and Christoph Lampert. A PAC-Bayesian bound for lifelong learning. In *Proc. Int. Conf. Mach. Learning (ICML)*, Bejing, China, 6 2014.

[29] Ron Amit and Ron Meir. Meta-learning by adjusting priors based on extended PAC-Bayes theory. In *Proc. Int. Conf. Mach. Learning (ICML)*, Stockholm, Sweden, 7 2018.

[30] Jonas Rothfuss, Vincent Fortuin, Martin Josifoski, and Andreas Krause. PACOH: Bayes-optimal meta-learning with PAC-guarantees. In *Proc. Int. Conf. Mach. Learning (ICML)*, Virtual conference, 7 2021.

[31] 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.

[32] Hamish Flynn, David Reeb, Melih Kandemir, and Jan Peters. PAC-Bayesian lifelong learning for multi-armed bandits. *Data Min. Knowl. Discov.*, 36(2):841–876, 3 2022.

[33] Alec Farid and Anirudha Majumdar. Generalization bounds for meta-learning via PAC-Bayes and uniform stability. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Virtual Conference, 12 2021.

[34] Jaesik Yoon, Taesup Kim, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, volume 31, Montreal, Canada, 12 2018.

[35] Arezou Rezazadeh, Sharu Theresa Jose, Giuseppe Durisi, and Osvaldo Simeone. Conditional mutual information-based generalization bound for meta learning. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Melbourne, Australia, 7 2021.

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

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

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

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] The contributions section specifies where in the paper each contribution can be found.
   (b) Did you describe the limitations of your work? [Yes] We state the assumptions required for our results in Section 2 and the statements of the results.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Foundational work, unrelated to a specific application.
   (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] We state the assumptions required for our results in Section 2 and the statements of the results.
   (b) Did you include complete proofs of all theoretical results? [Yes] Full proofs are included in the supplementary material.
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)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (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)? [N/A]
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? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
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]