Evaluated CMI Bounds for Meta Learning: Tightness and Expressiveness

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

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

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 $f_i \in \mathcal{F}$ is a task-specific function, and $h \in \mathcal{H}$ 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 is to learn a shared representation for the tasks or to find suitable hyperparameters for a base learner performing image classification.

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. Two current approaches for achieving this goal are: i) to 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 $\mathcal{H}$ on the

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

basis of n samples from $\hat{n}$ different tasks, and that a base learner selects a member f of a function class $\mathcal{F}$, on the basis of 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 output of the learning algorithm and the full training data, recent works provide 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 in [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 meta-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 for each step. By the use of the triangle inequality, the two bounds 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 bounds 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

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

dependencies. 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 meta learning setup that we consider throughout the paper, as well as 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 $\mathcal{D}$ on the task space $\mathcal{T}$. For a given task $\tau \in \mathcal{T}$, there is a corresponding in-task distribution $\mathcal{D}_\tau$ on the sample space $\mathcal{Z}$. The goal 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 the random variable $\hat{R} \in \widehat{\mathcal{R}}$ captures the potential stochasticity of the learner. 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 $\mathcal{A} : \mathcal{Z}^n \times \mathcal{R} \times \mathcal{U} \to \mathcal{W}$. The random vector $R = (R_1, \ldots, R_{\hat{n}}) \in \mathcal{R}^{\hat{n}}$ has entries that capture the potential stochasticity of each base learner. The entries are independent from the data and 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.

Within each task, the training set for the base learner is randomly formed from a supersample according to the conditional mutual information (CMI) framework of [15]. Specifically, for a given task τ, let $Z^\tau \in \mathcal{Z}^{n\times 2}$ denote the supersample, which is an $n \times 2$ matrix with elements generated independently from $\mathcal{D}_\tau$. For convenience, we index the two columns of Z^τ by 0 and 1 and the rows by $1, \ldots, n$. The training set Z^τ_S is formed on the basis of a membership vector $S = (S_1, \ldots, S_n)$, with entries generated independently from a $\mathrm{Bern}(1/2)$ distribution. More precisely, the jth element of Z^τ_S is given by $[Z^\tau_S]_j = Z^\tau_{j,S_j}$, i.e., the S_jth element from the jth row of Z^τ. Furthermore, we let $-S = (1 - S_1, \ldots, 1 - S_n)$ denote the modulo-2 complement of S, which we use to form the test set Z^τ_{-S}, whose jth element is given by $[Z^\tau_{-S}]_j = Z^\tau_{j,-S_j}$. With this construction, we randomly assign each sample in the supersample to either the training set or test set with equal probability.

We now describe the meta-supersample Z, which contains $2n$ samples from $2\hat{n}$ tasks, as in the meta-learning extension of the CMI framework provided in [35]. Throughout, we let $i \in \{1, \ldots, \hat{n}\}$ denote a task index, $j \in \{1, \ldots, n\}$ denote a sample index, and $k, l \in \{0, 1\}$ denote binary indices

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

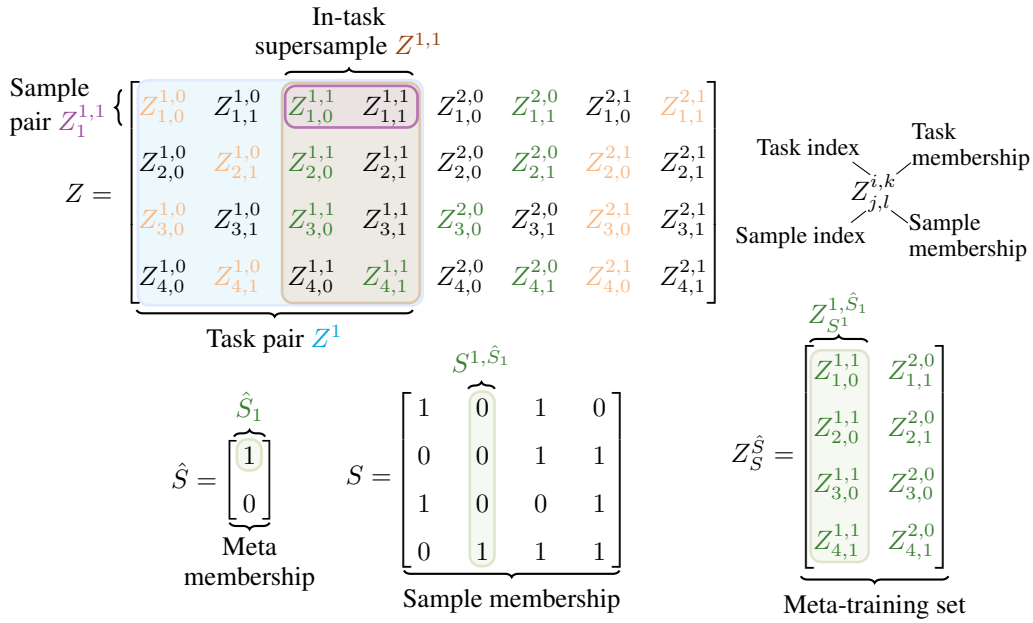

Figure 1: A graphical representation of our notation. In this example, the meta-supersample contains two task pairs: Z^1, which is marked in blue, and Z^2. In turn, Z^1 consists of the two in-task supersamples $Z^{1,0}$ and $Z^{1,1}$, which is marked in brown. Next, $Z^{1,1}$ consists of the four sample pairs $Z_j^{1,1}$, where $j = 1, \ldots, 4$. In the figure, $Z_1^{1,1}$ 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, $\hat{S}_1 = 1$ indicates that we should select task 1. Then, from the first sample pair in this task, $S^{1,\hat{S}_1} = 0$ 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.

indicating task membership and sample membership respectively. Formally, the meta-supersample Z can be viewed as a data structure with $4n\hat{n}$ elements. In Figure 1, we illustrate Z as an $n \times 4\hat{n}$ matrix for the case of $\hat{n} = 2$ task pairs and $n = 4$ sample pairs for each task. We decompose Z as $(Z^1, \ldots, Z^{\hat{n}})$, where each element can be seen as a task pair. Specifically, the pair Z^i can be decomposed as $(Z^{i,0}, Z^{i,1})$, where each element is a task-specific supersample as described above. The task-specific supersamples are $Z^{i,k} = (Z_1^{i,k}, \ldots, Z_n^{i,k})$, where each element is a pair of data samples. Specifically, each sample pair is $Z_j^{i,k} = (Z_{j,0}^{i,k}, Z_{j,1}^{i,k})$, where $Z_{j,l}^{i,k} \in \mathcal{Z}$. The elements of Z are generated as follows. First, we generate $\tau_{i,k} \sim \mathcal{D}$. Then, we independently generate the samples $Z_{j,l}^{i,k} \sim \mathcal{D}_{\tau_{i,k}}$. This is repeated for all indices to form Z.

Finally, we describe how the meta-training data is selected from the meta-supersample 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, while 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. The elements of all these vectors are 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. Using this convention, the training set for the (i,k)th task is constructed as $Z^{i,k}_{S^{i,k}} = (Z^{i,k}_{1,S^{i,k}_1}, \ldots, Z^{i,k}_{n,S^{i,k}_n})$. We will use the shorthands $Z^{i,k}_{S^i} = Z^{i,k}_{S^{i,k}}$ and $Z^{i,k}_{1,S^i_j} = Z^{i,k}_{1,S^{i,k}_j}$. The test set for the (i,k)th task $Z^{i,k}_{-S^i}$ is constructed analogously, but on the basis of $-S^{i,k}$. The full meta-training set is $Z^{\hat{S}}_{S} = (Z^{1,\hat{S}_1}_{S^{1,\hat{S}_1}}, \ldots, Z^{\hat{n},\hat{S}_{\hat{n}}}_{S^{\hat{n},\hat{S}_{\hat{n}}}})$, and the meta-test set $Z^{-\hat{S}}_{-S}$ is defined analogously. With this construction, each task in the meta-supersample is assigned to either the meta-training set or the meta-test set with equal probability. Then, as before, the samples within each task are assigned to an in-task training set or test set with equal probability. The meta-training set consists of training samples within training tasks, while the meta-test set consists of test samples within test tasks.

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

On the basis of the loss matrix λ 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 $L_{\mathcal{D}}$, 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 is used as the auxiliary loss. These four losses are given by

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

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

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 the joint distribution of X and Y 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 conditional 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 $I(X;Y|Z) = \mathbb{E}_Z\left[I^Z(X;Y)\right]$.

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 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 $L_{\mathcal{D}}$ in terms of $\widetilde{L}$. Chaining these two bounds, we obtain a bound on $L_{\mathcal{D}}$ 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,*

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

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*

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

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,*

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

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 $W = \{W^i\}_{i=1}^{\hat{n}}$. Then,*

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

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

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]. Furthermore, the bound in (7) compares favorably to [35, Thm. 1], since it decays with the product $n\hat{n}$ rather than with n and $\hat{n}$ separately. This improvement is due to the one-step derivation.

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 they have a more favorable dependence on the number of samples for low training losses. We demonstrate this improved rate for representation learning in Section 4.

Theorem 3 (Binary KL bounds)**.** *For $m \geq 2$, $q \in [0,1]$ and $c > 0$, 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,

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

Furthermore,

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

Interestingly, (11) provides a bound on the sum of the average meta-population loss $L_{\mathcal{D}}$, 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 $L_{\mathcal{D}}$ by using the lower bound $\bar{L} + \widetilde{L} \geq 0$, which is a sensible relaxation when $L_{\mathcal{D}}$ 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 $L_{\mathcal{D}}$ for the case where $\widehat{L} = 0$.

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

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

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 $L_{\mathcal{D}}(Z, \hat{S}, S)$ and $\widehat{L}(Z, \hat{S}, S)$ denote the meta-population loss and training loss given that the meta-training set is constructed from $(Z, \hat{S}, S)$. Specifically,

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

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

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 $\hat{\mathcal{A}}$ and $\mathcal{A}$ are 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 Q_{1,S_1} denote the conditional distribution of λ_{1,S_1} given $(Z, \hat{S}, S)$, and let P_{1,S_1} denote $\mathbb{E}_{\hat{S}}[Q_{1,S_1}]$. Furthermore, let $Q^{1,-\hat{S}_1}$ denote the conditional distribution of $\lambda^{1,-\hat{S}_1}$ given $(Z, \hat{S}_1, S^{1,-\hat{S}_1})$, and let $P^{1,-\hat{S}_1}$ denote $\mathbb{E}_{S^{1,-\hat{S}_1}}\left[Q^{1,-\hat{S}_1}\right]$. Then, there exist constants C_1, C_2 such that, with probability at least $1-\delta$ under the draw of $(Z, \hat{S}, S)$,*

$$\left|L_{\mathcal{D}}(Z, \hat{S}, S) - \widehat{L}(Z, \hat{S}, S)\right| \leq C_1 \sqrt{\frac{D(Q_{1,S_1} \,||\, P_{1,S_1}) + \log(\frac{n\sqrt{\hat{n}}}{\delta})}{\hat{n}}} + C_2 \sqrt{\frac{D(Q^{1,-\hat{S}_1} \,||\, P^{1,-\hat{S}_1}) + \log(\frac{\sqrt{n}}{\delta})}{n}}. \quad (15)$$

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

$$\mathbb{E}_{Z,\hat{S},S}[D(Q_{1,S_1} \,||\, P_{1,S_1})] = I(\lambda_{1,S_1}; \hat{S}|S, Z), \quad (16)$$

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

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 Q denote the conditional distribution of λ given $(Z, \hat{S}, S)$, and let P denote $\mathbb{E}_{\hat{S},S}[Q]$. Then, with probability at least $1-\delta$ under the draw of $(Z, \hat{S}, S)$,*

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

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

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

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 the product of an instance space and label space: $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. 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, we obtain the mapping $f_W \circ h_U : \mathcal{X} \to \mathcal{Y}$.

4.1 Minimax Generalization Bounds

To obtain explicit minimax bounds, we 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. These hold for any learner that outputs hypotheses from the specified classes.

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, let $2n \geq d_{\text{VC}} + 1$ and $2\hat{n} \geq d_N + 1$. Then,*

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

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

Corollary 4 establishes that, for the average setting, we can use the bounds in Theorems 1 and 2 to obtain minimax 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,*

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

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 bounds, 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 $(Z, \hat{S}, S)$,*

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

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

We now see that, suppressing logarithmic factors, the upper bound in (25) 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 these results and the ones in [4, 5]. First, while the complexity measures that we use are related to the Natarajan and VC dimension, the results in [5] are given in terms of Gaussian complexity. Furthermore, while [4, 5] provide excess risk bounds for a fixed target task with m training samples, the bounds in Corollary 6 are generalization bounds for a randomly drawn task. In Section 4.2, we extend our analysis to derive excess risk bounds for a fixed target task.

4.2 Excess Risk Bounds

In order to derive excess risk bounds for a specific target task, as is done in [5], we need to assume that the meta learner $\hat{\mathcal{A}}$ and the base learner $\mathcal{A}$ are empirical risk minimizers. This is in contrast to all previous bounds in this paper, which apply to any learning algorithms. Furthermore, we need a notion of oracle algorithms, which minimize the population loss. Finally, we need to assume that the tasks contained in the meta-supersample satisfy a notion of *task diversity*. Intuitively, this means that, given the output of the empirical risk-minimizing meta learner, the performance of the oracle base

learner on the tasks in the meta-supersample gives a reasonable indication of the performance of the oracle base learner on any possible task. Due to space constraints, we state here an informal version of a high-probability excess risk bound for a specified target task based on the one-step square-root generalization bound in Corollary 6. A precise statement of this result, along with its proof, is given in Appendix B.

Corollary 7 (Informal)**.** *Consider the setting of Corollary 6 and a fixed task τ_0. Let $Z^0_{S^0} \in \mathcal{Z}^m$ be a vector of m samples generated independently according to the data distribution $\mathcal{D}_{\tau_0}$ for task τ_0. Let $\hat{\mathcal{A}}$ and $\mathcal{A}$ be empirical risk minimizers. Let $L_0(Z, \hat{S}, S, Z^0_{S^0})$ denote the population loss on task τ_0 when applying $\hat{\mathcal{A}}$ to $Z^{\hat{S}}_S$ and $\mathcal{A}$ to $(Z^0_{S^0}, \hat{\mathcal{A}}(Z^{\hat{S}}_S))$, and let L^*_0 denote the smallest population loss for task τ_0 that can be obtained using functions from $\mathcal{F}$ and $\mathcal{H}$. Finally, assume that the supersample satisfies a task-diversity assumption with parameters ν, ϵ. Then, there exist constants C_1 and C_2 such that, with probability at least $1-\delta$ under the draw of $(Z, \hat{S}, S, Z^0_{S^0})$,*

$$L_0(Z, \hat{S}, S, Z^0_{S^0}) - L^*_0 \leq C_1 \sqrt{\frac{d_{\text{VC}} \log\left(\frac{\sqrt{m}}{d_{\text{VC}}}\right) + \log\left(\frac{\sqrt{m}}{\delta}\right)}{m}} + C_2 \nu^{-1} \sqrt{\frac{d_N \log\left(\binom{N}{2}\frac{n\hat{n}}{d_N}\right) + \hat{n} d_{\text{VC}} \log\left(\frac{n}{d_{\text{VC}}}\right) + \log\left(\frac{\sqrt{n\hat{n}}}{\delta}\right)}{n\hat{n}}} + \epsilon. \quad (25)$$

The bound in (25) displays the same scaling as the excess risk bound in [5].

It is possible to derive high-probability bounds based on the two-step square-root generalization bound in Corollary 6 by suitably substituting the two-step bound in the proof of Corollary 7. The same can be done using the bounds that are given in terms of information measures, and average excess risk bounds can also be derived by an analogous procedure. Finally, we note that it is possible to derive excess risk bounds for a new, random task, rather than a specified target task, without assuming task diversity. This is done in Corollary 8 in Appendix B.

5 Conclusions

In this paper, we derived new generalization bounds for meta learning using e-CMI, which improve 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]. The algorithm-dependence and expressiveness of our bounds indicate that they can be developed further to guide algorithm design. However, no recipe for this is provided in this paper. 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.

Acknowledgements

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

References

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

[2] S. Thrun and L. Pratt. *Learning to Learn: Introduction and Overview*. Springer, Boston, MA, USA, 1998.

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

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

[5] N. Tripuraneni, M. Jordan, and C. Jin. On the theory of transfer learning: The importance of task diversity. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Canada, Dec. 2020.

[6] K. Lounici, M. Pontil, S. van de Geer, and A. B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4):2164 – 2204, Aug. 2011.

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

[8] M. Pontil and A. Maurer. Excess risk bounds for multitask learning with trace norm regularization. In *Proc. Conf. Learn. Theory (COLT)*, Princeton, NJ, USA, June 2013.

[9] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *Proc. Int. Conf. Mach. Learning (ICML)*, July 2018.

[10] M.F. Balcan, M. Khodak, and A. Talwalkar. Provable guarantees for gradient-based meta-learning. In *Proc. Int. Conf. Mach. Learning (ICML)*, Long Beach, CA, USA, June 2019.

[11] D. Russo and J. Zou. Controlling bias in adaptive data analysis using information theory. In *Proc. Artif. Intell. Statist. (AISTATS)*, Cadiz, Spain, May 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, Dec. 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, May 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, Dec. 2019.

[15] T. Steinke and L. Zakynthinou. Reasoning about generalization via conditional mutual information. In *Proc. Conf. Learn. Theory (COLT)*, Graz, Austria, July 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. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Canada, Dec. 2020.

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

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

[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, July 2021.

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

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

[22] D.A. McAllester. A PAC-Bayesian tutorial with a dropout bound. *arXiv*, July 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, Dec. 2020.

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

[25] S. T. Jose and O. Simeone. Information-theoretic generalization bounds for meta-learning and applications. *Entropy*, 23(1), Jan. 2021.

[26] Q. Chen, C. Shui, and M. Marchand. Generalization bounds for meta-learning: An information-theoretic analysis. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Virtual Conference, Dec. 2021.

[27] S. T. Jose, O. Simeone, and G. Durisi. Transfer meta-learning: Information- theoretic bounds and information meta-risk minimization. *IEEE Trans. Inf. Theor.*, 68(1):474–501, Jan. 2022.

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

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

[30] J. Rothfuss, V. Fortuin, M. Josifoski, and A. Krause. PACOH: Bayes-optimal meta-learning with PAC-guarantees. In *Proc. Int. Conf. Mach. Learning (ICML)*, Virtual conference, July 2021.

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

[32] H. Flynn, D. Reeb, M. Kandemir, and J. Peters. PAC-Bayesian lifelong learning for multi-armed bandits. *Data Min. Knowl. Discov.*, 36(2):841–876, Mar. 2022.

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

[34] J. Yoon, T. Kim, O. Dia, S. Kim, Y. Bengio, and S. Ahn. Bayesian model-agnostic meta-learning. In *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, volume 31, Montreal, Canada, Dec. 2018.

[35] A. Rezazadeh, S. T. Jose, G. Durisi, and O. Simeone. Conditional mutual information-based generalization bound for meta learning. In *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Melbourne, Australia, July 2021.

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

[37] Y. Guermeur. Large Margin Multi-category Discriminant Models and Scale-sensitive Psi-dimensions. Research report, INRIA, Sep. 2006.

[38] M. 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]

A Proofs

In this appendix, we present the proofs of the results in the main paper. First, we give a summary of the notation that is used in this appendix. Then, in Section A.1, we present some lemmas that are useful for proving our main results. In Section A.2, we prove the average generalization bounds from Section 3.1. In Section A.3, we prove the high-probability results from Section 3.2. Finally, in Section A.4, we prove the generalization bounds for multiclass classification from Section 4.1.

Notation summary. For $i \in \{1, \dots, \hat{n}\}$, $k \in \{0, 1\}$, $j \in \{1, \dots, n\}$, and $l \in \{0, 1\}$, we let $Z^{i,k}_{j,l}$ denote the lth sample from the jth sample pair in the kth task of the ith task pair. This is illustrated in Figure 1. Throughout, i denotes a task index, j denotes a sample index, k denotes a selection within the task pair, and l denotes a selection within the sample pair. Furthermore, we let $Z^i_{j,l} = (Z^{i,0}_{j,l}, Z^{i,1}_{j,l})$ and $Z^i_j = \{Z^{i,k}_{j,l}\}^{k=0,1}_{l=0,1}$. The tasks used to form the training set are selected on the basis of the binary vector $\hat{S} = (\hat{S}_1, \dots, \hat{S}_{\hat{n}})$. Within task (i, k), the samples that form the training set are selected on the basis of $S^{i,k} = (S^{i,k}_1, \dots, S^{i,k}_n)$. For convenience, we let $S^i = (S^{i,0}, S^{i,1})$ and $S = (S^1, \dots, S^{\hat{n}})$. The training set for task (i, k) is $Z^{i,k}_{S^{i,k}} = (Z^{i,k}_{1,S^{i,k}_1}, \dots, Z^{i,k}_{n,S^{i,k}_n})$. As a shorthand, $Z^{i,k}_{S^i} = Z^{i,k}_{S^{i,k}}$. The collection of all samples is $Z = \{Z^i_j\}^{i=1:\hat{n}}_{j=1:n}$. The full data set for task (i, k) is $Z^{i,k} = \{Z^{i,k}_{j,l}\}^{l=0,1}_{j=1:n}$. The full data set for all training tasks is $Z^{\hat{S}} = (Z^{1,\hat{S}_1}, \dots, Z^{\hat{n},\hat{S}_{\hat{n}}})$. The jth training sample for task pair i is $Z^i_{j,S_j} = (Z^{i,0}_{j,S_j}, Z^{i,1}_{j,S_j})$. The jth training sample for all tasks is $Z_{j,S_j} = (Z^1_{j,S_j}, \dots, Z^{\hat{n}}_{j,S_j})$. The training sets for all tasks is $Z_S = (Z_{1,S_1}, \dots, Z_{n,S_n})$. The meta-training set is $Z^{\hat{S}}_S = (Z^{1,\hat{S}_1}_{S^1}, \dots, Z^{\hat{n},\hat{S}_1}_{S^{\hat{n}}})$. Finally, the output of the meta learner is U, the output of the base learner for task (i, k) is $W^{i,k}$, and we let $W^i = (W^{i,0}, W^{i,1})$ and $W = (W^1, \dots, W^{\hat{n}})$.

The conventions that we describe for $Z^{i,k}_{j,l}$ apply also for the losses $\lambda^{i,k}_{j,l}$, the instances $X^{i,k}_{j,l}$, the predictions $F^{i,k}_{j,l}$, and the representations $H^{i,k}_{j,l}$ that we consider in this appendix.

A.1 Useful Lemmas

In this section, we present some lemmas that will be useful in the derivations of the main results. We begin with two change of measure inequalities for functions of random variables.

Lemma 1 (Change of measure inequalities)**.** *Let X and Y be two random variables over $\mathcal{X}$ and $\mathcal{Y}$ respectively, 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 $f : \mathcal{X} \times \mathcal{Y} \to [-1, 1]$ and $g : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ be measurable functions. Furthermore, assume that $\mathbb{E}_{X,Y'}[f(X, Y')] = 0$. Then, the following inequalities hold:*

$$|\mathbb{E}_{X,Y}[f(X, Y)]| \leq \sqrt{2I(X; Y)}, \tag{26}$$

$$d(\mathbb{E}_{X,Y}[g(X, Y)] \mid\mid \mathbb{E}_{X,Y'}[g(X, Y')]) \leq I(X; Y). \tag{27}$$

Proof. Donsker-Varadhan's variational representation of the KL divergence implies that

$$\mathbb{E}_{X,Y}[\gamma f(X, Y)] \leq I(X; Y) + \log \mathbb{E}_{X,Y'}\left[e^{\gamma f(X,Y')}\right]. \tag{28}$$

Now, $f(X, Y')$ is bounded to $[-1, 1]$ and $\mathbb{E}_{X,Y'}[f(X, Y')] = 0$. Therefore, $f(X, Y')$ is a sub-Gaussian random variable, which implies that

$$\log \mathbb{E}_{X,Y'}\left[e^{\gamma f(X,Y')}\right] \leq \gamma^2/2. \tag{29}$$

Using this upper bound in (28), we obtain

$$\mathbb{E}_{X,Y}[\gamma f(X, Y)] - \gamma^2/2 \leq I(X; Y). \tag{30}$$

By maximizing the left-hand side over γ, we establish (26).

We now turn to (27). Let $d_\gamma(q \,||\, p) = \gamma q - \log(1 - p + pe^\gamma)$, and note that this function is convex. By Jensen's inequality,

$$d_\gamma(\mathbb{E}_{X,Y}[g(X,Y)] \,||\, \mathbb{E}_{X,Y'}[g(X,Y')]) \leq \mathbb{E}_{X,Y}[d_\gamma(g(X,Y) \,||\, \mathbb{E}_{Y'}[g(X,Y')])]\,. \tag{31}$$

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

$$\mathbb{E}_{X,Y}[d_\gamma(g(X,Y) \,||\, \mathbb{E}_{Y'}[g(X,Y')])] \leq I(X;Y) + \log \mathbb{E}_{X,Y'}\left[e^{d_\gamma\left(g(X,Y') \,||\, \mathbb{E}_{Y'}\left[g(X,Y')\right]\right)}\right]. \tag{32}$$

By [22, Eq. (17)], we have

$$\log \mathbb{E}_{X,Y'}\left[e^{d_\gamma\left(g(X,Y') \,||\, \mathbb{E}_{Y'}\left[g(X,Y')\right]\right)}\right] \leq 0. \tag{33}$$

Thus, by combining (31)-(33),

$$d_\gamma(\mathbb{E}_{X,Y}[g(X,Y)] \,||\, \mathbb{E}_{X,Y'}[g(X,Y')]) \leq I(X;Y). \tag{34}$$

The desired result follows because $\sup_\gamma d_\gamma(\cdot \,||\, \cdot) = d(\cdot \,||\, \cdot)$.

□

Lemma 2 (Conditioning on independent random variables)**.** *Consider the random variables X, Y and Z, where X and Z are independent. Then,*

$$I(X;Y) \leq I(X;Y|Z). \tag{35}$$

Proof. The result follows by using the independence of X and Z (which implies that $I(X;Z) = 0$), the chain rule for mutual information, and the non-negativity of mutual information as follows. Note that

$$I(X;Y,Z) = I(X;Z) + I(X;Y|Z) = I(X;Y|Z). \tag{36}$$

Alternatively,

$$I(X;Y,Z) = I(X;Y) + I(X;Z|Y) \geq I(X;Y). \tag{37}$$

Thus,

$$I(X;Y) \leq I(X;Y,Z) = I(X;Y|Z). \tag{38}$$

□

Lemma 3 (Full-sample relaxation)**.** *Consider n independent random variables $X = \{X_i\}_{i=1}^n$ and a random variable Y. Let ϕ be a convex function. Then,*

$$\frac{1}{n}\sum_{i=1}^n \phi(I(X_i;Y) \leq \phi\left(\frac{I(X;Y)}{n}\right). \tag{39}$$

Proof. Let $X_{<i}$ denote $\{X_1, \ldots, X_{i-1}\}$. By the chain rule of mutual information,

$$I(X;Y) = \sum_{i=1}^n I(X_i;Y|X_{<i}). \tag{40}$$

Due to the independence of the X_i, Lemma 2 implies that $I(X_i;Y|X_{<i}) \geq I(X_i;Y)$. Combined with Jensen's inequality, this implies that

$$\phi\left(\frac{I(X;Y)}{n}\right) \geq \phi\left(\sum_{i=1}^n \frac{I(X_i;Y)}{n}\right) \geq \frac{1}{n}\sum_{i=1}^n \phi(I(X_i;Y))\,. \tag{41}$$

□

Lemma 4 (Sauer-Shelah lemma for the VC and Natarajan dimension)**.** *Let $g_{\mathcal{F}}(\cdot)$ denote the growth function of the function class $\mathcal{F}$. Specifically, $g_{\mathcal{F}}(m)$ is the maximum number of different ways in which a data set of size m can be classified using functions from $\mathcal{F}$. For any function class $\mathcal{F}$ with VC dimension d_{VC},*

$$g_{\mathcal{F}}(m) \leq \sum_{i=0}^{d_{\text{VC}}} \binom{m}{i} \leq \begin{cases} 2^{d_{\text{VC}}+1}, & m < d_{\text{VC}}+1 \\ \left(\frac{em}{d_{\text{VC}}}\right)^{d_{\text{VC}}}, & m \geq d_{\text{VC}}+1 \end{cases} \tag{42}$$

More generally, for any function class $\mathcal{F}$ with range $\{0, \ldots, N-1\}$ and Natarajan dimension d_N,

$$g_{\mathcal{F}}(m) \leq \sum_{i=0}^{d_N} \binom{m}{i} \binom{N}{2}^i \leq \begin{cases} N^{d_N+1}, & m < d_N+1, \\ \left(\binom{N}{2}\frac{em}{d_N}\right)^{d_N}, & m \geq d_N+1. \end{cases} \tag{43}$$

Proof. The first inequality in (43) follows from [36, Cor. 5] and the second follows from [37, Lemma 10]. The result in (42) follows by setting $N = 2$ in (43), for which the Natarajan dimension d_N coincides with the VC dimension [36, p. 222]. □

A.2 Proofs for Section 3.1

Proof of Theorem 1. We start by establishing a task-level generalization bound, i.e., a bound on $\left|L_{\mathcal{D}} - \widetilde{L}\right|$. By Jensen's inequality, the convexity of $|\cdot|$ implies that

$$\left|L_{\mathcal{D}} - \widetilde{L}\right| \leq \frac{1}{n\hat{n}} \sum_{i,j=1}^{\hat{n},n} \mathbb{E}_{Z,\hat{S}_i}\left[\left|\mathbb{E}_{\lambda_j^{i,-\hat{S}_i}, S_j^{i,-\hat{S}_i}}\left[\lambda_{j,-S_j^i}^{i,-\hat{S}_i} - \lambda_{j,S_j^i}^{i,-\hat{S}_i}\right]\right|\right]. \tag{44}$$

Let S' be an independent copy of S. By symmetry, we see that

$$\mathbb{E}_{\lambda_j^{i,-\hat{S}_i}, S_j'^{i,-\hat{S}_i}}\left[\lambda_{j,-S_j'^i}^{i,-\hat{S}_i} - \lambda_{j,S_j'^i}^{i,-\hat{S}_i}\right] = 0. \tag{45}$$

Using (26), we can therefore bound the argument of the expectation in (44) as

$$\left|\mathbb{E}_{\lambda_j^{i,-\hat{S}_i}, S_j^{i,-\hat{S}_i}}\left[\lambda_{j,-S_j^i}^{i,-\hat{S}_i} - \lambda_{j,S_j^i}^{i,-\hat{S}_i}\right]\right| \leq \sqrt{2I^{Z,\hat{S}_i}(\lambda_j^{i,-\hat{S}_i}; S_j^{i,-\hat{S}_i})}. \tag{46}$$

Combining (44) and (46), we obtain the following task-level generalization bound:

$$\left|L_{\mathcal{D}} - \widetilde{L}\right| \leq \frac{1}{n\hat{n}} \sum_{i,j=1}^{\hat{n},n} \mathbb{E}_{Z,\hat{S}_i}\left[\sqrt{2I^{Z,\hat{S}_i}(\lambda_j^{i,-\hat{S}_i}; S_j^{i,-\hat{S}_i})}\right]. \tag{47}$$

This is the first step of the two-step derivation.

Next, we establish an environment-level bound, i.e., a bound on $\left|\widetilde{L} - \widehat{L}\right|$. Again, by Jensen's inequality, the convexity of $|\cdot|$ implies that

$$\left|\widetilde{L} - \widehat{L}\right| \leq \frac{1}{n\hat{n}} \sum_{i,j=1}^{\hat{n},n} \mathbb{E}_{Z,S_j^i}\left[\left|\mathbb{E}_{\lambda_{j,S_j^i}^i, \hat{S}_i}\left[\lambda_{j,S_j^i}^{i,-\hat{S}_i} - \lambda_{j,S_j^i}^{i,\hat{S}_i}\right]\right|\right]. \tag{48}$$

Symmetry implies that

$$\mathbb{E}_{\lambda_{j,S_j^i}^i, \hat{S}_i'}\left[\lambda_{j,S_j^i}^{i,-\hat{S}_i'} - \lambda_{j,S_j^i}^{i,\hat{S}_i'}\right] = 0. \tag{49}$$

We again bound the argument of the expectation, using (26), as

$$\left|\mathbb{E}_{\lambda_{j,S_j^i}^i, \hat{S}_i}\left[\lambda_{j,S_j^i}^{i,-\hat{S}_i} - \lambda_{j,S_j^i}^{i,\hat{S}_i}\right]\right| \leq \sqrt{2I^{Z,S_j^i}(\lambda_{j,S_j^i}^i; \hat{S}_i)}. \tag{50}$$

Combining the two preceding inequalities, we obtain the following environment-level generalization guarantee, which is the second step:

$$\left|\widetilde{L}-\widehat{L}\right| \leq \frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}\mathbb{E}_{Z,S_j^i}\left[\sqrt{2I^{Z,S_j^i}(\lambda_{j,S_j^i}^i;\hat{S}_i)}\right]. \tag{51}$$

We conclude the proof by observing that $\left|L_{\mathcal{D}}-\widehat{L}\right| \leq \left|L_{\mathcal{D}}-\widetilde{L}\right|+\left|\widetilde{L}-\widehat{L}\right|$ by the triangle inequality, and by using (47) and (51) to bound the two terms.

□

Proof of Corollary 1. We begin with the first sum on the right-hand side of (4). By Jensen's inequality,

$$\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}\mathbb{E}_{Z,S_j^i}\left[\sqrt{2I^{Z,S_j^i}(\lambda_{j,S_j^i}^i;\hat{S}_i)}\right] \leq \sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}2I(\lambda_{j,S_j^i}^i;\hat{S}_i|Z,S_j^i)}. \tag{52}$$

By Lemma 2 and the independence of $\hat{S}_i$ and (S,R), we conclude that conditioning on S and R does not increase the mutual information. Hence,

$$\sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}2I(\lambda_{j,S_j^i}^i;\hat{S}_i|Z,S_j^i)} \leq \sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}2I(\lambda_{j,S_j^i}^i;\hat{S}_i|Z,S,R)}. \tag{53}$$

Since adding more random variables to the argument of the mutual information cannot decrease it, we have

$$\sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}2I(\lambda_{j,S_j^i}^i;\hat{S}_i|Z,S,R)} \leq \sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}2I(\lambda_{j,S_j^i};\hat{S}_i|Z,S,R)}, \tag{54}$$

where $\lambda_{j,S_j^i}=\{\lambda_{j,S_j^i}^i\}_{i=1}^{\hat{n}}$. Since the $\hat{S}_i$ are independent, it follows from Lemma 3 that

$$\sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}2I(\lambda_{j,S_j^i};\hat{S}_i|Z,S,R)} \leq \sqrt{\frac{1}{n}\sum_{j=1}^{n}\frac{2I(\lambda_{j,S_j^i};\hat{S}|Z,S,R)}{\hat{n}}}. \tag{55}$$

Now, note that given Z, S and R, the losses λ_{j,S_j^i} are a function of the output of the meta learner U. Therefore,

$$\sqrt{\frac{1}{n}\sum_{j=1}^{n}\frac{2I(\lambda_{j,S_j^i};\hat{S}|Z,S,R)}{\hat{n}}} \leq \sqrt{\frac{2I(U;\hat{S}|Z,S,R)}{\hat{n}}} \leq \sqrt{\frac{2I(U;\hat{S}|Z,S)}{\hat{n}}}, \tag{56}$$

where the last step follows from the independence of U and R. By combining (52)-(56), we can bound the first sum in the right-hand side of (4) as

$$\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}\mathbb{E}_{Z,S_j^i}\left[\sqrt{2I^{Z,S_j^i}(\lambda_{j,S_j^i}^i;\hat{S}_i)}\right] \leq \sqrt{\frac{2I(U;\hat{S}|Z,S)}{\hat{n}}}. \tag{57}$$

For the second sum on the right-hand side of (4), we again use Jensen's inequality to conclude that

$$\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}\mathbb{E}_{Z,\hat{S}_i}\left[\sqrt{2I^{Z,\hat{S}_i}(\lambda_j^{i,-\hat{S}_i};S_j^{i,-\hat{S}_i})}\right] \leq \sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}2I(\lambda_j^{i,-\hat{S}_i};S_j^{i,-\hat{S}_i}|Z,\hat{S}_i)}. \tag{58}$$

Since adding more random variables does not decrease the mutual information,

$$\sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} 2I(\lambda_j^{i,-\hat{S}_i}; S_j^{i,-\hat{S}_i}|Z,\hat{S}_i)} \leq \sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} 2I(\lambda^{i,-\hat{S}_i}; S_j^{i,-\hat{S}_i}|Z,\hat{S}_i)}. \tag{59}$$

By Lemma 2 and the independence of the $S_j^{i,-\hat{S}_i}$,

$$\sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} 2I(\lambda^{i,-\hat{S}_i}; S_j^{i,-\hat{S}_i}|Z,\hat{S}_i)} \leq \sqrt{\frac{1}{\hat{n}}\sum_{i=1}^{\hat{n}} \frac{2I(\lambda^{i,-\hat{S}_i}; S^{i,-\hat{S}_i}|Z,\hat{S}_i)}{n}}. \tag{60}$$

Since R_i, $\hat{S}_i$ and $S^{i,-\hat{S}_i}$ have the same distribution for all $i = 1, \ldots, \hat{n}$,

$$\sqrt{\frac{1}{\hat{n}}\sum_{i=1}^{\hat{n}} \frac{2I(\lambda^{i,-\hat{S}_i}; S^{i,-\hat{S}_i}|Z,\hat{S}_i)}{n}} = \sqrt{\frac{2I(\lambda^{1,-\hat{S}_1}; S^{1,-\hat{S}_1}|Z,\hat{S}_i)}{n}}. \tag{61}$$

Finally, given Z and $\hat{S}_i$, $\lambda^{1,-\hat{S}_1}$ is a function of $W^{1,-\hat{S}_1}$. Hence,

$$\sqrt{\frac{2I(\lambda^{1,-\hat{S}_1}; S^{1,-\hat{S}_1}|Z,\hat{S}_i)}{n}} \leq \sqrt{\frac{2I(W^{1,-\hat{S}_1}; S^{1,-\hat{S}_1}|Z,\hat{S}_i)}{n}}. \tag{62}$$

By combining (58)-(62), we can bound the second term in the right-hand side of (4) as

$$\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} \mathbb{E}_{Z,\hat{S}_i}\left[\sqrt{2I^{Z,\hat{S}_i}(\lambda_j^{i,-\hat{S}_i}; S_j^{i,-\hat{S}_i})}\right] \leq \sqrt{\frac{2I(W^{1,-\hat{S}_1}; S^{1,-\hat{S}_1}|Z,\hat{S}_i)}{n}}. \tag{63}$$

The result follows by combining (57) and (63). □

Proof of Theorem 2. By Jensen's inequality, we have

$$\left|L_{\mathcal{D}} - \widehat{L}\right| \leq \frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} \mathbb{E}_Z\left[\left|\mathbb{E}_{\lambda_j^i,\hat{S}_i,S_j^i}\left[\lambda_{j,-S_j^i}^{i,-\hat{S}_i} - \lambda_{j,S_j^i}^{i,\hat{S}_i}\right]\right|\right]. \tag{64}$$

Now, let $\hat{S}'$ and S' be independent copies of $\hat{S}$ and S. Note that

$$\mathbb{E}_{\lambda_j^i,\hat{S}_i',{S'}_j^i}\left[\lambda_{j,-{S'}_j^i}^{i,-\hat{S}_i'} - \lambda_{j,{S'}_j^i}^{i,\hat{S}_i'}\right] = 0. \tag{65}$$

We can therefore apply (26), with $X = \lambda_j^i$ and Y being the pair of random variables $(\hat{S}_i', {S'}_j^i)$, to bound the argument of the expectation as

$$\left|\mathbb{E}_{\lambda_j^i,\hat{S}_i,S_j^i}\left[\lambda_{j,-S_j^i}^{i,-\hat{S}_i} - \lambda_{j,S_j^i}^{i,\hat{S}_i}\right]\right| \leq \sqrt{2I^Z(\lambda_j^i; \hat{S}_i, S_j^i)}. \tag{66}$$

Combining (64) and (66), we establish the desired result.

□

Proof of Corollary 2. By Jensen's inequality,

$$\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} \mathbb{E}_Z\left[\sqrt{2I^Z(\lambda_j^i; \hat{S}_i, S_j^i)}\right] \leq \sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} 2I(\lambda_j^i; \hat{S}_i, S_j^i|Z)}. \tag{67}$$

Since adding more random variables does not decrease the mutual information,

$$\sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} 2I(\lambda_j^i;\hat{S}_i,S_j^i|Z)} \leq \sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} 2I(\lambda;\hat{S}_i,S_j^i|Z)}, \tag{68}$$

where $\lambda = \{\lambda_j^i\}_{j=1:n}^{i=1:\hat{n}}$. By the independence of the S_j^i for different j, and $(\hat{S}_i, S^i)$ for different i,

$$\sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} 2I(\lambda;\hat{S}_i,S_j^i|Z)} \leq \sqrt{\frac{2I(\lambda;\hat{S},S|Z)}{n\hat{n}}}. \tag{69}$$

Given Z, the losses λ are a function of $W = \{W^i\}_{i=1}^{\hat{n}}$. Thus,

$$\sqrt{\frac{2I(\lambda;\hat{S},S|Z)}{n\hat{n}}} \leq \sqrt{\frac{2I(W;\hat{S},S|Z)}{n\hat{n}}}. \tag{70}$$

Combining (67)-(70), we establish the first inequality in (7). Next, since adding random variables does not decrease mutual information,

$$\sqrt{\frac{2I(W;\hat{S},S|Z)}{n\hat{n}}} \leq \sqrt{\frac{2I(U,W;\hat{S},S|Z)}{n\hat{n}}} \tag{71}$$

$$\leq \sqrt{\frac{2I(U;\hat{S},S|Z)+2I(W;\hat{S},S|Z,U)}{n\hat{n}}} \tag{72}$$

where the second step follows from the chain rule. Since the conditional distribution $P_{W\hat{S}S|ZU}$ factorizes as $P_{\hat{S}|ZU}\prod_{i=1}^{\hat{n}} P_{W^iS^i|ZU}$, we have that

$$I(W;\hat{S},S|Z,U) = \sum_{i=1}^{\hat{n}} I(W^i;S^i|Z,U). \tag{73}$$

Furthermore, since $(R_i, \hat{S}_i, S^i)$ are identically distributed for all i,

$$\sum_{i=1}^{\hat{n}} I(W^i;S^i|Z,U) = \hat{n}I(W^1;S^1|Z,U). \tag{74}$$

By combining (70)-(74), we get

$$\sqrt{\frac{2I(W;\hat{S},S|Z)}{n\hat{n}}} \leq \sqrt{\frac{2I(U;\hat{S},S|Z)+2\hat{n}I(W^1;S^1|Z,U)}{n\hat{n}}}. \tag{75}$$

This establishes the second inequality in (7).

Finally, by the chain rule,

$$I(U;\hat{S},S|Z) \leq I(U;\hat{S},S|Z)+I(U;Z) = I(U;Z,\hat{S},S) = I(U;Z_S^{\hat{S}}). \tag{76}$$

Similarly,

$$I(W^1;S^1|Z,U) \leq I(W^1;S^1|Z,U)+I(W^1;Z|U) = I(W^1;S^1,Z|U) = I(W^1;Z_{S^1}^1|U). \tag{77}$$

By combining (75)-(77), we establish (8). Thus, to summarize, we have shown that

$$\left|L_{\mathcal{D}} - \widehat{L}\right| \leq \sqrt{\frac{2I(W;\hat{S},S|Z)}{n\hat{n}}} \leq \sqrt{\frac{2I(U;\hat{S},S|Z) + 2\hat{n}I(W^1;S^1|Z,U)}{n\hat{n}}} \tag{78}$$

$$\leq \sqrt{\frac{2I(U;Z_S^{\hat{S}}) + 2\hat{n}I(W^1;Z_{S^1}^1|U)}{n\hat{n}}}. \tag{79}$$

□

Proof of Theorem 3. We begin by proving (10). First, we derive a task-level generalization bound. By Jensen's inequality, we have

$$d\left(\widetilde{L}\,||\,\frac{\widetilde{L}+L_{\mathcal{D}}}{2}\right) \leq \frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}\mathbb{E}_{Z,\hat{S}_i}\left[\mathbb{E}_{\lambda_j^{i,-\hat{S}_i},S_j^{i,-\hat{S}_i}}\left[d\left(\lambda_{j,S_j^i}^{i,-\hat{S}_i}\,||\,\frac{\lambda_{j,-S_j^i}^{i,-\hat{S}_i}+\lambda_{j,S_j^i}^{i,-\hat{S}_i}}{2}\right)\right]\right]. \tag{80}$$

Since $S_j^{i,-\hat{S}_i} \in \{0,1\}$, $\lambda_{j,-S_j^i}^{i,-\hat{S}_i} + \lambda_{j,S_j^i}^{i,-\hat{S}_i} = \lambda_{j,0}^{i,-\hat{S}_i} + \lambda_{j,1}^{i,-\hat{S}_i}$ does not actually depend on $S_j^{i,-\hat{S}_i}$. Now, let S' be an independent copy of S. It follows that

$$\mathbb{E}_{\lambda_j^{i,-\hat{S}_i},S_j'^{i,-\hat{S}_i}}\left[\lambda_{j,S_j'^i}^{i,-\hat{S}_i}\right] = \mathbb{E}_{\lambda_j^{i,-\hat{S}_i}}\left[\frac{\lambda_{j,0}^{i,-\hat{S}_i}+\lambda_{j,1}^{i,-\hat{S}_i}}{2}\right]. \tag{81}$$

We can thus use (26) to bound the argument of the expectation as

$$\mathbb{E}_{\lambda_j^{i,-\hat{S}_i},S_j^{i,-\hat{S}_i}}\left[d\left(\lambda_{j,S_j^i}^{i,-\hat{S}_i}\,||\,\frac{\lambda_{j,-S_j^i}^{i,-\hat{S}_i}+\lambda_{j,S_j^i}^{i,-\hat{S}_i}}{2}\right)\right] \leq I^{Z,\hat{S}_i}(\lambda_j^{i,-\hat{S}_i};S_j^{i,-\hat{S}_i}). \tag{82}$$

Combining the two inequalities, we obtain

$$d\left(\widetilde{L}\,||\,\frac{\widetilde{L}+L_{\mathcal{D}}}{2}\right) \leq \frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} I(\lambda_j^{i,-\hat{S}_i};S_j^{i,-\hat{S}_i}|Z,\hat{S}_i). \tag{83}$$

Recall that

$$d_2^{-1}(q,c) = \sup\left\{p\in[0,1] : d\left(q\,||\,\frac{q+p}{2}\right) \leq c\right\}. \tag{84}$$

Using $d_2^{-1}(\cdot)$ to invert (83), we get

$$L_{\mathcal{D}} \leq d_2^{-1}\left(\widetilde{L}, \frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} I(\lambda_j^{i,-\hat{S}_i};S_j^{i,-\hat{S}_i}|Z,\hat{S}_i)\right). \tag{85}$$

Next, we perform similar steps at the environment level. First, by Jensen's inequality,

$$d\left(\widehat{L}\,||\,\frac{\widehat{L}+\widetilde{L}}{2}\right) \leq \frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}\mathbb{E}_{Z,S_j^i}\left[\mathbb{E}_{\lambda_{j,S_j^i}^i,\hat{S}_i}\left[d\left(\lambda_{j,S_j^i}^{i,\hat{S}_i}\,||\,\frac{\lambda_{j,S_j^i}^{i,\hat{S}_i}+\lambda_{j,S_j^i}^{i,-\hat{S}_i}}{2}\right)\right]\right]. \tag{86}$$

Let $\hat{S}'$ be an independent copy of $\hat{S}$. By a similar argument as in the proof of the task-level bound,

$$\mathbb{E}_{\lambda_{j,S_j^i}^i,\hat{S}_i'}\left[\lambda_{j,S_j^i}^{i,\hat{S}_i'}\right] = \mathbb{E}_{\lambda_{j,S_j^i}^i}\left[\frac{\lambda_{j,S_j^i}^{i,0}+\lambda_{j,S_j^i}^{i,1}}{2}\right]. \tag{87}$$

We can therefore again bound the argument of the expectation with (27) to obtain

$$\mathbb{E}_{\lambda^i_{j,S^i_j},\hat{S}_i}\left[d\left(\lambda^{i,\hat{S}_i}_{j,S^i_j}\,||\,\frac{\lambda^{i,\hat{S}_i}_{j,S^i_j}+\lambda^{i,-\hat{S}_i}_{j,S^i_j}}{2}\right)\right] \leq I^{Z,S^i_j}(\lambda^i_{j,S^i_j};\hat{S}_i). \tag{88}$$

By combining the two inequalities, we find that

$$d\left(\widehat{L}\,||\,\frac{\widehat{L}+\widetilde{L}}{2}\right) \leq \frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} I(\lambda^i_{j,S^i_j};\hat{S}_i|Z,S^i_j) \tag{89}$$

which, through the use of $d_2^{-1}(\cdot)$, implies that

$$\widetilde{L} \leq d_2^{-1}\left(\widehat{L},\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} I(\lambda^i_{j,S^i_j};\hat{S}_i|Z,S^i_j)\right). \tag{90}$$

To complete the proof, we use the following observation. Assume that $L_{\mathcal{D}} \leq B(\widetilde{L})$, where $B(\cdot)$ is a non-decreasing function. Then, if $\widetilde{L} \leq \widehat{B}(\widehat{L})$, we have $L_{\mathcal{D}} \leq B(\widehat{B}(\widehat{L}))$. To apply this observation, we note that $d_2^{-1}(\cdot,c)$ is non-decreasing for $c>0$. Chaining the two bounds, we obtain

$$L_{\mathcal{D}} \leq d_2^{-1}\left(\widetilde{L},\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} I(\lambda^{i,-\hat{S}_i}_j;S^{i,-\hat{S}_i}_j|Z,\hat{S}_i)\right) \tag{91}$$

$$\leq d_2^{-1}\left(d_2^{-1}\left(\widehat{L},\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} I(\lambda^i_{j,S^i_j};\hat{S}_i|Z,S^i_j)\right),\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} I(\lambda^{i,-\hat{S}_i}_j;S^{i,-\hat{S}_i}_j|Z,\hat{S}_i)\right). \tag{92}$$

This establishes (10).

Next, we turn to (11). By Jensen's inequality, we have

$$d\left(\widehat{L}\,||\,\frac{\widehat{L}+\bar{L}+\widetilde{L}+L_{\mathcal{D}}}{4}\right)$$
$$\leq \frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}\mathbb{E}_Z\left[\mathbb{E}_{\lambda^i_j,\hat{S}_i,S^i_j}\left[d\left(\lambda^{i,\hat{S}_i}_{j,S^i_j}\,||\,\frac{\lambda^{i,\hat{S}_i}_{j,S^i_j}+\lambda^{i,\hat{S}_i}_{j,-S^i_j}+\lambda^{i,-\hat{S}_i}_{j,S^i_j}+\lambda^{i,-\hat{S}_i}_{j,-S^i_j}}{4}\right)\right]\right]. \tag{93}$$

Let $\hat{S}'$ and S' be independent copies of $\hat{S}$ and S respectively. We note that

$$\mathbb{E}_{\lambda^i_j,\hat{S}'_i,S'^i_j}\left[\lambda^{i,\hat{S}_i}_{j,S^i_j}\right] = \mathbb{E}_{\lambda^i_j}\left[\frac{\lambda^{i,0}_{j,0}+\lambda^{i,0}_{j,1}+\lambda^{i,1}_{j,0}+\lambda^{i,1}_{j,1}}{4}\right]. \tag{94}$$

This means that we can apply (27) to the argument of the expectation to get

$$\mathbb{E}_{\lambda^i_j,\hat{S}_i,S^i_j}\left[d\left(\lambda^{i,\hat{S}_i}_{j,S^i_j}\,||\,\frac{\lambda^{i,\hat{S}_i}_{j,S^i_j}+\lambda^{i,\hat{S}_i}_{j,-S^i_j}+\lambda^{i,-\hat{S}_i}_{j,S^i_j}+\lambda^{i,-\hat{S}_i}_{j,-S^i_j}}{4}\right)\right] \leq I^Z(\lambda^i_j;\hat{S}_i,S^i_j). \tag{95}$$

The result in (11) now follows by combining (93) and (95).

□

Proof of Corollary 3. Since $\lim_{x\to 0^+} x\log x = 0$, we use the convention that $0\log 0 = 0$. From the definition of $d(\cdot\,||\,\cdot)$, we thus get

$$d\Big(0\,||\,\frac{p}{4}\Big) = \log\frac{1}{1-\frac{p}{4}} \tag{96}$$

from which it follows that

$$d_4^{-1}(0,c) = 4 - 4e^{-c}. \tag{97}$$

By the non-negativity of the loss function, $L_{\mathcal{D}} \le L_{\mathcal{D}} + \bar{L} + \widetilde{L}$. Thus, combining (11) with (97), we obtain the result in (12).

□

A.3 Proofs for Section 3.2

To derive the simplified result stated in Theorem 4, we assume that the meta learner and base learner are invariant to the order of the data samples. However, this assumption is only necessary to simplify the expression, and a similar bound holds more generally without this assumption. Therefore, we first state and prove this more general result in Theorem 6. Then, we describe how to simplify the result to obtain Theorem 4. Later, when proving Corollary 6, we will use the more general Theorem 6 as the basis of the derivation.

Theorem 6. *Consider the setting introduced in Section 2. For each j, let Q_{j,S_j} denote the conditional distribution of λ_{j,S_j} given $(Z,\hat{S},S)$, and let P_{j,S_j} denote $\mathbb{E}_{\hat{S}}\big[Q_{j,S_j}\big]$. Furthermore, let $Q^{i,-\hat{S}_i}$ denote the conditional distribution of $\lambda^{i,-\hat{S}_i}$ given $(Z,\hat{S},S)$, and let $P^{i,-\hat{S}_i}$ denote $\mathbb{E}_{S^{-\hat{S}}}\Big[Q^{i,-\hat{S}_i}\Big]$. Then, with probability at least $1-\delta$ under the draw of $(Z,\hat{S},S)$,*

$$\Big|L_{\mathcal{D}} - \widehat{L}\Big| \le \sqrt{\frac{\frac{1}{n}\sum_{j=1}^{n} 2D(Q_{j,S_j}\,||\,P_{j,S_j}) + 2\log(\frac{2n\sqrt{\hat{n}}}{\delta})}{\hat{n}-1}} + \sqrt{\frac{\sum_{i=1}^{\hat{n}} 2D(Q^{i,-\hat{S}_i}\,||\,P^{i,-\hat{S}_i}) + 2\log(\frac{2\sqrt{n\hat{n}}}{\delta})}{n\hat{n}-1}}. \tag{98}$$

Proof of Theorem 6. First, let $\widetilde{L}(Z,\hat{S},S)$ denote the training loss on unobserved tasks,

$$\widetilde{L}(Z,\hat{S},S) = \frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} \mathbb{E}_{\hat{R},R_i}\Big[\ell(\mathcal{A}(Z_{S^i}^{i,-\hat{S}_i}, R_i, \hat{\mathcal{A}}(Z_S^{\hat{S}},\hat{R})), Z_{j,-S_j^i}^{i,-\hat{S}_i})\Big]. \tag{99}$$

We begin by establishing an environment-level bound. Let λ_{j,S_j} be distributed according to Q_{j,S_j}. By Jensen's inequality,

$$\frac{\hat{n}-1}{2}\Big(\widetilde{L}(Z,\hat{S},S) - \widehat{L}(Z,\hat{S},S)\Big)^2 \le \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}_{\lambda_{j,S_j}}\left[\frac{\hat{n}-1}{2}\left(\frac{1}{\hat{n}}\sum_{i=1}^{\hat{n}}\lambda_{j,S_j^i}^{i,-\hat{S}_i} - \lambda_{j,S_j^i}^{i,\hat{S}_i}\right)^2\right]. \tag{100}$$

Now, let λ'_{j,S_j} be distributed according to P_{j,S_j}. By Donsker-Varadhan's variational representation of the KL divergence,

$$\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}_{\lambda_{j,S_j}}\left[\frac{\hat{n}-1}{2}\left(\frac{1}{\hat{n}}\sum_{i=1}^{\hat{n}}\lambda_{j,S_j^i}^{i,-\hat{S}_i} - \lambda_{j,S_j^i}^{i,\hat{S}_i}\right)^2\right]$$
$$\le \frac{1}{n}\sum_{j=1}^{n}\left(D(Q_{j,S_j}\,||\,P_{j,S_j}) + \log\mathbb{E}_{\lambda'_{j,S_j}}\left[\exp\left(\frac{\hat{n}-1}{2}\left(\frac{1}{\hat{n}}\sum_{i=1}^{\hat{n}}\lambda'^{i,-\hat{S}_i}_{j,S_j^i} - \lambda'^{i,\hat{S}_i}_{j,S_j^i}\right)^2\right)\right]\right). \tag{101}$$

For each j, Markov's inequality implies that, with probability at least $1-\delta$ under the draw of $(Z, \hat{S}, S)$,

$$D(Q_{j,S_j} \,||\, P_{j,S_j}) + \log \mathbb{E}_{\lambda'_{j,S_j}} \left[\exp\left(\frac{\hat{n}-1}{2} \left(\frac{1}{\hat{n}} \sum_{i=1}^{\hat{n}} \lambda'^{i,-\hat{S}_i}_{j,S^i_j} - \lambda'^{i,\hat{S}_i}_{j,S^i_j} \right)^2 \right) \right]$$
$$\leq D(Q_{j,S_j} \,||\, P_{j,S_j}) + \log \mathbb{E}_{\lambda'_{j,S_j},Z,\hat{S},S} \left[\frac{1}{\delta} \exp\left(\frac{\hat{n}-1}{2} \left(\frac{1}{\hat{n}} \sum_{i=1}^{\hat{n}} \lambda'^{i,-\hat{S}_i}_{j,S^i_j} - \lambda'^{i,\hat{S}_i}_{j,S^i_j} \right)^2 \right) \right]. \quad (102)$$

By the union bound, this implies that, with $\delta \to \delta/n$, (102) holds for all j simultaneously with probability at least $1-\delta$. Thus, with probability at least $1-\delta$ under the draw of $(Z, \hat{S}, S)$,

$$\frac{1}{n}\sum_{j=1}^{n} \left(D(Q_{j,S_j} \,||\, P_{j,S_j}) + \log \mathbb{E}_{\lambda'_{j,S_j}} \left[\exp\left(\frac{\hat{n}-1}{2} \left(\frac{1}{\hat{n}} \sum_{i=1}^{\hat{n}} \lambda'^{i,-\hat{S}_i}_{j,S^i_j} - \lambda'^{i,\hat{S}_i}_{j,S^i_j} \right)^2 \right) \right] \right)$$
$$\leq \frac{1}{n}\sum_{j=1}^{n} \left(D(Q_{j,S_j} \,||\, P_{j,S_j}) + \log \mathbb{E}_{\lambda'_{j,S_j},Z,\hat{S},S} \left[\frac{n}{\delta} \exp\left(\frac{\hat{n}-1}{2} \left(\frac{1}{\hat{n}} \sum_{i=1}^{\hat{n}} \lambda'^{i,-\hat{S}_i}_{j,S^i_j} - \lambda'^{i,\hat{S}_i}_{j,S^i_j} \right)^2 \right) \right] \right). \quad (103)$$

Note that, on the right-hand side of (103), $\hat{S}$ is independent from (λ'_{j,S_j}, Z, S). Furthermore, for each (i,j), $\lambda'^{i,-\hat{S}_i}_{j,S^i_j} - \lambda'^{i,\hat{S}_i}_{j,S^i_j}$ is bounded to $[-1,1]$ and $\mathbb{E}_{\hat{S}_i}\left[\lambda'^{i,-\hat{S}_i}_{j,S^i_j} - \lambda'^{i,\hat{S}_i}_{j,S^i_j}\right] = 0$. This implies that $\frac{1}{\hat{n}}\sum_{i=1}^{\hat{n}} \lambda'^{i,-\hat{S}_i}_{j,S^i_j} - \lambda'^{i,\hat{S}_i}_{j,S^i_j}$ is a $1/\sqrt{\hat{n}}$-sub-Gaussian random variable, from which it follows that [38, Thm. 2.6.(IV)]

$$\log \mathbb{E}_{\lambda'_{j,S_j},Z,\hat{S},S} \left[\exp\left(\frac{\hat{n}-1}{2} \left(\frac{1}{\hat{n}} \sum_{i=1}^{\hat{n}} \lambda'^{i,-\hat{S}_i}_{j,S^i_j} - \lambda'^{i,\hat{S}_i}_{j,S^i_j} \right)^2 \right) \right] \leq \log(\sqrt{\hat{n}}). \quad (104)$$

By substituting (104) into (103), we obtain

$$\frac{1}{n}\sum_{j=1}^{n} \left(D(Q_{j,S_j} \,||\, P_{j,S_j}) + \log \mathbb{E}_{\lambda'_{j,S_j}} \left[\exp\left(\frac{\hat{n}-1}{2} \left(\frac{1}{\hat{n}} \sum_{i=1}^{\hat{n}} \lambda'^{i,-\hat{S}_i}_{j,S^i_j} - \lambda'^{i,\hat{S}_i}_{j,S^i_j} \right)^2 \right) \right] \right)$$
$$\leq \frac{1}{n}\sum_{j=1}^{n} D(Q_{j,S_j} \,||\, P_{j,S_j}) + \log\left(\frac{n\sqrt{\hat{n}}}{\delta} \right). \quad (105)$$

By combining (100)-(105), we get, after some arithmetic,

$$\left| \widetilde{L}(Z,\hat{S},S) - \widehat{L}(Z,\hat{S},S) \right| \leq \sqrt{\frac{\frac{1}{n}\sum_{j=1}^{n} 2D(Q_{j,S_j} \,||\, P_{j,S_j}) + 2\log\left(\frac{n\sqrt{\hat{n}}}{\delta}\right)}{\hat{n}-1}}. \quad (106)$$

We now turn to the task level. Let $Q^{-\hat{S}}$ denote the conditional distribution of $\lambda^{-\hat{S}}$ given $(Z, \hat{S}, S)$, and let $P^{-\hat{S}}$ denote $\mathbb{E}_{S^{-\hat{S}}}\left[Q^{-\hat{S}}\right]$. Let $\lambda^{-\hat{S}}$ be distributed according to $Q^{-\hat{S}}$. By Jensen's inequality,

$$\frac{n\hat{n}-1}{2}\left(L_{\mathcal{D}}(Z,\hat{S},S) - \widetilde{L}(Z,\hat{S},S) \right)^2 \leq \mathbb{E}_{\lambda^{-\hat{S}}} \left[\frac{n\hat{n}-1}{2} \left(\frac{1}{n\hat{n}} \sum_{i,j=1}^{\hat{n},n} \lambda^{i,-\hat{S}_i}_{j,-S^i_j} - \lambda^{i,-\hat{S}_i}_{j,S^i_j} \right)^2 \right]. \quad (107)$$

Now, let $\lambda'^{-\hat{S}}$ be distributed according to $P^{-\hat{S}}$. By Donsker-Varadhan's variational representation of the KL divergence,

$$\mathbb{E}_{\lambda^{-\hat{S}}}\left[\frac{n\hat{n}-1}{2}\left(\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}\lambda^{i,-\hat{S}_i}_{j,-S^i_j}-\lambda^{i,-\hat{S}_i}_{j,S^i_j}\right)^2\right]$$
$$\leq D(Q^{-\hat{S}}\,||\,P^{-\hat{S}})+\log\mathbb{E}_{\lambda'^{-\hat{S}}}\left[\exp\left(\frac{n\hat{n}-1}{2}\left(\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}\lambda'^{i,-\hat{S}_i}_{j,-S^i_j}-\lambda'^{i,-\hat{S}_i}_{j,S^i_j}\right)^2\right)\right]. \quad (108)$$

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

$$D(Q^{-\hat{S}}\,||\,P^{-\hat{S}})+\log\mathbb{E}_{\lambda'^{-\hat{S}}}\left[\exp\left(\frac{n\hat{n}-1}{2}\left(\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}\lambda'^{i,-\hat{S}_i}_{j,-S^i_j}-\lambda'^{i,-\hat{S}_i}_{j,S^i_j}\right)^2\right)\right]$$
$$\leq D(Q^{-\hat{S}}\,||\,P^{-\hat{S}})+\log\mathbb{E}_{\lambda'^{-\hat{S}},Z,\hat{S},S}\left[\frac{1}{\delta}\exp\left(\frac{n\hat{n}-1}{2}\left(\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}\lambda'^{i,-\hat{S}_i}_{j,-S^i_j}-\lambda'^{i,-\hat{S}_i}_{j,S^i_j}\right)^2\right)\right]. \quad (109)$$

Note that $S^{-\hat{S}}$ is independent from $(\lambda'^{-\hat{S}},Z,\hat{S},S^{\hat{S}})$, that for each (i,j), $\lambda'^{i,-\hat{S}_i}_{j,-S^i_j}-\lambda'^{i,-\hat{S}_i}_{j,S^i_j}$ is bounded to $[-1,1]$, and that $\mathbb{E}_{S^{i,-\hat{S}_i}_j}\left[\lambda'^{i,-\hat{S}_i}_{j,-S^i_j}-\lambda'^{i,-\hat{S}_i}_{j,S^i_j}\right]=0$. Thus, it follows that [38, Thm. 2.6.(IV)]

$$\log\mathbb{E}_{\lambda'^{-\hat{S}},Z,\hat{S},S}\left[\exp\left(\frac{n\hat{n}-1}{2}\left(\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}\lambda'^{i,-\hat{S}_i}_{j,-S^i_j}-\lambda'^{i,-\hat{S}_i}_{j,S^i_j}\right)^2\right)\right]\leq\log\left(\sqrt{n\hat{n}}\right). \quad (110)$$

By substituting (110) into (109), we obtain

$$D(Q^{-\hat{S}}\,||\,P^{-\hat{S}})+\log\mathbb{E}_{\lambda'^{-\hat{S}}}\left[\exp\left(\frac{n\hat{n}-1}{2}\left(\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}\lambda'^{i,-\hat{S}_i}_{j,-S^i_j}-\lambda'^{i,-\hat{S}_i}_{j,S^i_j}\right)^2\right)\right]$$
$$\leq D(Q^{-\hat{S}}\,||\,P^{-\hat{S}})+\log\left(\frac{\sqrt{n\hat{n}}}{\delta}\right). \quad (111)$$

Since $\lambda^{i,-\hat{S}_i}$ and $\lambda^{i',-\hat{S}_{i'}}$ are losses on separate, unobserved tasks, they are dependent only through U. Therefore, they are conditionally independent given $(Z,\hat{S},S^{\hat{S}})$. By the chain rule for the KL divergence, it follows that

$$D(Q^{-\hat{S}}\,||\,P^{-\hat{S}})=\sum_{i=1}^{\hat{n}}D(Q^{i,-\hat{S}_i}\,||\,P^{i,-\hat{S}_i}). \quad (112)$$

By combining (107)-(112), we get, after some arithmetic, that with probability at least $1-\delta$ under $(Z,\hat{S},S)$,

$$\left|L_{\mathcal{D}}-\widetilde{L}\right|\leq\sqrt{\frac{\sum_{i=1}^{\hat{n}}2D(Q^{i,-\hat{S}_i}\,||\,P^{i,-\hat{S}_i})+2\log\left(\frac{\sqrt{n\hat{n}}}{\delta}\right)}{n\hat{n}-1}}. \quad (113)$$

By the triangle inequality, $\left|L_{\mathcal{D}}-\widehat{L}\right|\leq\left|L_{\mathcal{D}}-\widetilde{L}\right|+\left|\widetilde{L}-\widehat{L}\right|$. By the union bound, (106) and (113) hold simultaneously with probability at least $1-2\delta$ under $(Z,\hat{S},S)$. Therefore, with $\delta\to\delta/2$, they

hold simultaneously with probability at least $1-\delta$. Thus, with probability at least $1-\delta$ under the draw of $(Z, \hat{S}, S)$,

$$\left|L_{\mathcal{D}} - \widehat{L}\right| \leq \sqrt{\frac{\frac{1}{n}\sum_{j=1}^{n} 2D(Q_{j,S_j} \,||\, P_{j,S_j}) + 2\log\left(\frac{2n\sqrt{\hat{n}}}{\delta}\right)}{\hat{n}-1}} + \sqrt{\frac{\sum_{i=1}^{\hat{n}} 2D(Q^{i,-\hat{S}_i} \,||\, P^{i,-\hat{S}_i}) + 2\log\left(\frac{2\sqrt{n\hat{n}}}{\delta}\right)}{n\hat{n}-1}}. \quad (114)$$

□

Having established Theorem 6, we now show how to use it to derive Theorem 4 under the assumption that the meta learner and base learner are invariant to the order of the samples.

Proof of Theorem 4. By the assumptions that the meta learner and base learner are invariant to the sample order and task index, we can reorder the data set so that $\max_j D(Q_{j,S_j} \,||\, P_{j,S_j}) = D(Q_{1,S_1} \,||\, P_{1,S_1})$. With this, $\frac{1}{n}\sum_{j=1}^{n} D(Q_{j,S_j} \,||\, P_{j,S_j}) \leq D(Q_{1,S_1} \,||\, P_{1,S_1})$. Similarly, we can reorder the data set so that $\max_i D(Q^{i,-\hat{S}_i} \,||\, P^{i,-\hat{S}_i}) = D(Q_\lambda^{1,-\hat{S}_1} \,||\, P_\lambda^{1,-\hat{S}_1})$, implying that $\sum_{i=1}^{\hat{n}} D(Q^{i,-\hat{S}_i} \,||\, P^{i,-\hat{S}_i}) \leq \hat{n} D(Q_\lambda^{1,-\hat{S}_1} \,||\, P_\lambda^{1,-\hat{S}_1})$. To obtain the final result, we note that for $n, \hat{n} \geq 2$, we have $1/(\hat{n}–1) \leq 2/\hat{n}$, $\log 2 \leq \log(n\sqrt{\hat{n}}/\delta)$, $1/(n–1) \leq 2/n$, $\log 2 \leq \log(\sqrt{n\hat{n}}/\delta)$, and $\log(\sqrt{n\hat{n}})/n\hat{n} \leq \log(\sqrt{n})/n$. Thus, we get the final result

$$\left|L_{\mathcal{D}} - \widehat{L}\right| \leq 2\sqrt{2}\sqrt{\frac{D(Q_{1,S_1}||P_{1,S_1}) + \log\left(\frac{n\sqrt{\hat{n}}}{\delta}\right)}{\hat{n}}} + 2\sqrt{2}\sqrt{\frac{D(Q^{1,-\hat{S}_1}||P^{1,-\hat{S}_1}) + \log\left(\frac{\sqrt{n}}{\delta}\right)}{n}}. \quad (115)$$

Thus, the bound holds with $C_1 = C_2 = 2\sqrt{2}$.

□

While the simplifying assumption of invariance to the order of samples leads to a simpler result, it does not hold for all learning algorithms. Therefore, we will use the more general form given in (114) as the basis of Corollary 6.

Note that the first term of (115) diverges as $n \to \infty$. This counter-intuitive behavior, which requires both n and $\hat{n}$ to be large for the bound to be nonvacuous, is common in PAC-Bayesian bounds for meta learning [29], and is seemingly an effect of the two-step approach. It is possible to obtain a different bound where this dependence is not explicit, similar to [28, 31], by simply not applying Jensen's inequality to the average over j in (100) nor the average over i in (107). However, the resulting environment-level KL divergence is different. In particular, if we were to use this alternative bound to derive minimax bounds in Section 4, the logarithmic dependence on n would be embedded in this KL divergence. In the proof of Corollary 6, we point out where this difference would come into play.

We present the alternative bound in the following remark.

Remark 1. *Let Q_S denote the conditional distribution of λ_S given $(Z, \hat{S}, S)$ and let $P_S = \mathbb{E}_{\hat{S}}[Q_S]$. Then,*

$$\left|L_{\mathcal{D}} - \widehat{L}\right| \leq 2\sqrt{2}\sqrt{\frac{D(Q_S \,||\, P_S) + \log\left(\frac{\sqrt{\hat{n}}}{\delta}\right)}{\hat{n}}} + 2\sqrt{2}\sqrt{\frac{D(Q^{1,-\hat{S}_1} \,||\, P^{1,-\hat{S}_1}) + \log\left(\frac{\sqrt{n}}{\delta}\right)}{n}}. \quad (116)$$

Proof. The proof follows that of Theorem 4, so we only detail the differences: In (100), the average over j is not moved outside the square when using Jensen's inequality; Donsker-Varadhan is now

used to perform a change of measure from Q_S to P_S in (101); and the union bound in (103) is no longer needed. $\square$

We now turn to Theorem 5.

Proof of Theorem 5. Recall that Q denotes the conditional distribution of λ given $(Z, \hat{S}, S)$, and that P denotes $\mathbb{E}_{\hat{S},S}[Q]$. Let λ be distributed according to Q. By Jensen's inequality,

$$\frac{n\hat{n}-1}{2}\Big(L_{\mathcal{D}}(Z,\hat{S},S)-\widehat{L}(Z,\hat{S},S)\Big)^2 \leq \mathbb{E}_{\lambda}\left[\frac{n\hat{n}-1}{2}\left(\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}\lambda_{j,-S_j^i}^{i,-\hat{S}_i}-\lambda_{j,S_j^i}^{i,\hat{S}_i}\right)^2\right]. \tag{117}$$

Next, let λ' be distributed according to P. By Donsker-Varadhan's variational representation of the KL divergence,

$$\mathbb{E}_{\lambda}\left[\frac{n\hat{n}-1}{2}\left(\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}\lambda_{j,-S_j^i}^{i,-\hat{S}_i}-\lambda_{j,S_j^i}^{i,\hat{S}_i}\right)^2\right] \leq D(Q\,||\,P)+\log\mathbb{E}_{\lambda'}\left[\frac{n\hat{n}-1}{2}\left(\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}\lambda'^{i,-\hat{S}_i}_{j,-S_j^i}-\lambda'^{i,\hat{S}_i}_{j,S_j^i}\right)^2\right]. \tag{118}$$

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

$$D(Q\,||\,P)+\log\mathbb{E}_{\lambda'}\left[\exp\left(\frac{n\hat{n}-1}{2}\left(\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}\lambda'^{i,-\hat{S}_i}_{j,-S_j^i}-\lambda'^{i,\hat{S}_i}_{j,S_j^i}\right)^2\right)\right] \leq D(Q\,||\,P)+\log\mathbb{E}_{\lambda',Z,\hat{S},S}\left[\frac{1}{\delta}\exp\left(\frac{n\hat{n}-1}{2}\left(\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}\lambda'^{i,-\hat{S}_i}_{j,-S_j^i}-\lambda'^{i,\hat{S}_i}_{j,S_j^i}\right)^2\right)\right]. \tag{119}$$

Now, note that $(\hat{S}, S)$ are independent from λ', Z. Furthermore, $\lambda'^{i,-\hat{S}_i}_{j,-S_j^i}-\lambda'^{i,\hat{S}_i}_{j,S_j^i}$ is bounded to $[-1,1]$, and $\mathbb{E}_{\hat{S}_i,S_j^i}\left[\lambda'^{i,-\hat{S}_i}_{j,-S_j^i}-\lambda'^{i,\hat{S}_i}_{j,S_j^i}\right]=0$. Thus, $\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}\lambda'^{i,-\hat{S}_i}_{j,-S_j^i}-\lambda'^{i,\hat{S}_i}_{j,S_j^i}$ is a $1/\sqrt{n\hat{n}}$-sub-Gaussian random variable, from which it follows that [38, Thm. 2.6.(IV)]

$$\log\mathbb{E}_{\lambda',Z,\hat{S},S}\left[\exp\left(\frac{n\hat{n}-1}{2}\left(\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}\lambda'^{i,-\hat{S}_i}_{j,-S_j^i}-\lambda'^{i,\hat{S}_i}_{j,S_j^i}\right)^2\right)\right] \leq \log\sqrt{n\hat{n}}. \tag{120}$$

By combining (117)-(120), we get

$$\frac{n\hat{n}-1}{2}\Big(L_{\mathcal{D}}(Z,\hat{S},S)-\widehat{L}(Z,\hat{S},S)\Big)^2 \leq D(Q\,||\,P)+\log\frac{\sqrt{n\hat{n}}}{\delta}. \tag{121}$$

The desired result now follows after some arithmetic.

$\square$

A.4 Proofs for Section 4

Proof of Corollary 4. We begin with (20). To establish this inequality, we bound the two sums on the right-hand side of (4) separately. First, by Jensen's inequality, we find that

$$\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}\mathbb{E}_{Z,S_j^i}\left[\sqrt{I^{Z,S_j^i}\Big(\lambda_{j,S_j^i}^i;\hat{S}_i\Big)}\right]\leq\sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}I\Big(\lambda_{j,S_j^i}^i;\hat{S}_i|Z,S\Big)}. \tag{122}$$

Let $X_{j,l}^{i,k}$ denote the projection of $Z_{j,l}^{i,k}$ onto $\mathcal{X}$, i.e., $X_{j,l}^{i,k}$ contains the unlabelled instances from Z. The notation for $X_{j,l}^{i,k}$ is inherited from the notation for $Z_{j,l}^{i,k}$ introduced in Section 2. Let $f(\mathcal{A}(Z_{S^i}^{(i,k)},R_i,\hat{\mathcal{A}}(Z_S^{\hat{S}},\hat{R})),\cdot)$ denote the function from $\mathcal{F}$ that is selected by $\hat{\mathcal{A}}$ and $\mathcal{A}$ for task (i,k) on the basis of $(Z,\hat{S},S,R,\hat{R})$. We let $F_{j,S_j^i}^{i,k}$ denote the predicted label that the meta learner and the base learner produce for $X_{j,S_j^i}^{i,k}$. Furthermore, we let $F_{j,S_j^i}^i=(F_{j,S_j^i}^{i,0},F_{j,S_j^i}^{i,1})$. Again, $F_{j,l}^{i,k}$ inherits the notational conventions that we use for $Z_{j,l}^{i,k}$. Note that, given Z, the losses $\lambda_{j,S_j^i}^i$ are a function of $F_{j,S_j^i}^i$. Thus, by the data-processing inequality,

$$\sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}I\Big(\lambda_{j,S_j^i}^i;\hat{S}_i|Z,S\Big)}\leq\sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}I\Big(F_{j,S_j^i}^i;\hat{S}_i|Z,S\Big)}. \tag{123}$$

Next, Let $h(\hat{\mathcal{A}}(Z_S^{\hat{S}},\hat{R}),\cdot)$ denote the function from $\mathcal{H}$ that is selected by $\hat{\mathcal{A}}$ on the basis of $(Z,\hat{S},S,R,\hat{R})$. We denote the representation that the meta learner induces on $X_{j,S_j^i}^i$ as $H_{j,S_j^i}^i$, the elements of which is given by, for $k\in\{0,1\}$,

$$H_{j,S_j^i}^{i,k}=h(\hat{\mathcal{A}}(Z_S^{\hat{S}},\hat{R}),X_{j,S_j^i}^{i,k}). \tag{124}$$

Note that, given Z, S^i and R_i, the predictions in $F_{j,S_j^i}^i$ are a deterministic function of the intermediate representations $H_{j,S_j^i}^i$. Therefore, using the independence of R_i and $\hat{S}_i$,

$$\sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}I\Big(F_{j,S_j^i}^i;\hat{S}_i|Z,S\Big)}\leq\sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}I\Big(F_{j,S_j^i}^i;\hat{S}_i|Z,S,R_i\Big)} \tag{125}$$

$$\leq\sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}I\Big(H_{j,S_j^i}^i;\hat{S}_i|Z,S\Big)}, \tag{126}$$

where R_i disappears from the conditioning due to the independence of $H_{j,S_j^i}^i$ and R_i. Next, by adding random variables and using Lemma 3, we get

$$\sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}I\Big(H_{j,S_j^i}^i;\hat{S}_i|Z,S\Big)}\leq\sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}I\Big(H_{j,S_j^i};\hat{S}_i|Z,S\Big)} \tag{127}$$

$$\leq\sqrt{\frac{\frac{1}{n}\sum_{j=1}^{n}I\Big(H_{j,S_j};\hat{S}|Z,S\Big)}{\hat{n}}}. \tag{128}$$

For a given j, Z, and S, the $2\hat{n}$ inputs that give rise to H_{j,S_j^i} are fixed. Thus, the number of possible different values that H_{j,S_j^i} can take is at most $g_{\mathcal{H}}(2\hat{n})$, where $g_{\mathcal{H}}(\cdot)$ is the growth function of $\mathcal{H}$.

From this, it follows that

$$I\Big(H_{j,S_j};\hat{S}|Z,S\Big) \leq \mathcal{H}\Big(H_{j,S_j}|Z,S\Big) \tag{129}$$

$$\leq \log g_{\mathcal{H}}(2\hat{n}) \tag{130}$$

$$\leq d_N \log\left(\binom{N}{2}\frac{2e\hat{n}}{d_N}\right). \tag{131}$$

Here, $\mathcal{H}(H_{j,S_j}|Z,S)$ denotes the conditional entropy of H_{j,S_j} given (Z,S), and the last inequality follows from Lemma 4. Since (129) does not depend on j, we find that

$$\sqrt{\frac{\frac{1}{n}\sum_{j=1}^{n} I\Big(H_{j,S_j};\hat{S}|Z,S\Big)}{\hat{n}}} \leq \sqrt{\frac{d_N \log\left(\binom{N}{2}\frac{2e\hat{n}}{d_N}\right)}{\hat{n}}}. \tag{132}$$

By combining (122)-(132), we get

$$\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} \mathbb{E}_{Z,S_j^i}\left[\sqrt{I^{Z,S_j^i}\Big(\lambda_{j,S_j^i}^{i};\hat{S}_i\Big)}\right] \leq \sqrt{\frac{d_N \log\left(\binom{N}{2}\frac{2e\hat{n}}{d_N}\right)}{\hat{n}}}. \tag{133}$$

Next, we turn to the second sum on the right-hand side of (4). First, by Jensen's inequality,

$$\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}\mathbb{E}_{Z,\hat{S}_i}\left[\sqrt{I^{Z,\hat{S}_i}\Big(\lambda_j^{i,-\hat{S}_i};S_j^{i,-\hat{S}_i}\Big)}\right] \leq \sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} I\Big(\lambda_j^{i,-\hat{S}_i};S_j^{i,-\hat{S}_i}|Z,\hat{S}_i\Big)}. \tag{134}$$

Note that, given Z, the losses $\lambda_j^{i,-\hat{S}_i}$ are a function of $F_j^{i,-\hat{S}_i}$. Therefore,

$$\sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} I\Big(\lambda_j^{i,-\hat{S}_i};S_j^{i,-\hat{S}_i}|Z,\hat{S}_i\Big)} \leq \sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} I\Big(F_j^{i,-\hat{S}_i};S_j^{i,-\hat{S}_i}|Z,\hat{S}_i\Big)} \tag{135}$$

$$\leq \sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} I\Big(F^{i,-\hat{S}_i};S_j^{i,-\hat{S}_i}|Z,\hat{S}_i\Big)}, \tag{136}$$

where we used the fact that adding random variables cannot decrease mutual information. By the independence of $S_j^{i,-\hat{S}_i}$ for different j and Lemma 3,

$$\sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} I\Big(F^{i,-\hat{S}_i};S_j^{i,-\hat{S}_i}|Z,\hat{S}_i\Big)} \leq \sqrt{\frac{\frac{1}{\hat{n}}\sum_{i=1}^{\hat{n}} I\Big(F^{i,-\hat{S}_i};S^{i,-\hat{S}_i}|Z,\hat{S}_i\Big)}{n}}. \tag{137}$$

Now, note that given Z and $\hat{S}_i$, the $2n$ inputs that give rise to $F^{i,-\hat{S}_i}$ are fixed. Recall that $g_{\mathcal{F}}(\cdot)$ denotes the growth function of $\mathcal{F}$. Then,

$$\sqrt{\frac{\frac{1}{\hat{n}}\sum_{i=1}^{\hat{n}} I\Big(F^{i,-\hat{S}_i};S^{i,-\hat{S}_i}|Z,\hat{S}_i\Big)}{n}} \leq \sqrt{\frac{\frac{1}{\hat{n}}\sum_{i=1}^{\hat{n}} \mathcal{H}\Big(F^{i,-\hat{S}_i}|Z,\hat{S}_i\Big)}{n}} \tag{138}$$

$$\leq \sqrt{\frac{\frac{1}{\hat{n}}\sum_{i=1}^{\hat{n}} g_{\mathcal{F}}(2n)}{n}} \tag{139}$$

$$\leq \sqrt{\frac{\frac{1}{\hat{n}}\sum_{i=1}^{\hat{n}} d_{\text{VC}} \log\Big(\frac{2en}{d_{\text{VC}}}\Big)}{n}}, \tag{140}$$

where we used Lemma 4. By combining (134)-(140), we get

$$\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}\mathbb{E}_{Z,\hat{S}_i}\left[\sqrt{I^{Z,\hat{S}_i}\left(\lambda_j^{i,-\hat{S}_i};S_j^{i,-\hat{S}_i}\right)}\right]\leq\sqrt{\frac{\frac{1}{\hat{n}}\sum_{i=1}^{\hat{n}}d_{\text{VC}}\log\left(\frac{2en}{d_{\text{VC}}}\right)}{n}}. \tag{141}$$

The result in (20) now follows by combining (4), (133) and (141).

We now turn to (21). First, by Jensen's inequality,

$$\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}\mathbb{E}_Z\left[\sqrt{2I^Z(\lambda_j^i;\hat{S}_i,S_j^i)}\right]\leq\sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}2I(\lambda_j^i;\hat{S}_i,S_j^i|Z)} \tag{142}$$

$$\leq\sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}2I(\lambda;\hat{S}_i,S_j^i|Z)}, \tag{143}$$

where in the second step, we used that adding random variables does not decrease mutual information. Next, by the independence of the S_j^i over j and of the $(\hat{S}_i,S^i)$ over i,

$$\sqrt{\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n}2I(\lambda;\hat{S}_i,S_j^i|Z)}\leq\sqrt{\frac{2I(\lambda;\hat{S},S|Z)}{n\hat{n}}}. \tag{144}$$

Now, note that given Z, the losses λ are a function of the predictions $F=\{F_j^i\}_{j=1:n}^{i=1:\hat{n}}$. Hence,

$$\sqrt{\frac{2I(\lambda;\hat{S},S|Z)}{n\hat{n}}}\leq\sqrt{\frac{2I(F;\hat{S},S|Z)}{n\hat{n}}} \tag{145}$$

$$\leq\sqrt{\frac{2I(F,H;\hat{S},S|Z)}{n\hat{n}}} \tag{146}$$

where $H=\{H_j^i\}_{j=1:n}^{i=1:\hat{n}}$ and the second step follows by adding random variables. By the chain rule,

$$\sqrt{\frac{2I(F,H;\hat{S},S|Z)}{n\hat{n}}}\leq\sqrt{\frac{2I(H;\hat{S},S|Z)+2I(F;\hat{S},S|Z,H)}{n\hat{n}}} \tag{147}$$

$$\leq\sqrt{\frac{2\mathcal{H}(H|Z)+2\mathcal{H}(F|Z,H)}{n\hat{n}}}. \tag{148}$$

Since H is given by the elementwise application of some $h\in\mathcal{H}$ to $X=\{X_j^i\}_{j=1:n}^{1:\hat{n}}$, it can take at most $g_{\mathcal{H}}(4n\hat{n})$ different values, similar to previous arguments. This implies that

$$\mathcal{H}(H|Z)\leq\log(g_{\mathcal{H}}(4n\hat{n}))\leq d_N\log\left(\binom{N}{2}\frac{4en\hat{n}}{d_N}\right), \tag{149}$$

where the last inequality is again due to Lemma 4. Given H and Z, the predictions F can take at most $(g_{\mathcal{F}}(2n))^{2\hat{n}}$ different values, since the $2n$ inputs to each of the $2\hat{n}$ task-specific functions are fixed. This implies that

$$\mathcal{H}(F|H,Z)\leq 2\hat{n}\log(g_{\mathcal{F}}(2n))\leq 2\hat{n}d_{\text{VC}}\log\left(\frac{2en}{d_{\text{VC}}}\right), \tag{150}$$

where we again used Lemma 4. The desired result follows by combining (6) with (142)-(150).

□

Proof of Corollary 5. By the same steps as in (142)-(150), we find that

$$\frac{1}{n\hat{n}}\sum_{i,j=1}^{\hat{n},n} I(\lambda_j^i;\hat{S}_i,S_j^i|Z) \leq d_N \log\left(\binom{N}{2}\frac{4en\hat{n}}{d_N}\right) + 2\hat{n}d_{\text{VC}}\log\left(\frac{2en}{d_{\text{VC}}}\right). \tag{151}$$

By combining this with (12), we find that

$$L_{\mathcal{D}} \leq \frac{4d_N \log\left(\binom{N}{2}\frac{4en\hat{n}}{d_N}\right) + 8\hat{n}d_{\text{VC}}\log\left(\frac{2en}{d_{\text{VC}}}\right)}{n\hat{n}}. \tag{152}$$

This establishes the desired result. □

Proof of Corollary 6. First, we establish (23). As mentioned in the proof of Theorem 4, we start the derivation from the more general bound given in (114) rather than the simplified bound given in (15). We begin by bounding $D(Q_{j,S_j} \,||\, P_{j,S_j})$. Recall that Q_{j,S_j} denotes the conditional distribution of λ_{j,S_j} given $(Z,\hat{S},S)$ and $P_{j,S_j} = \mathbb{E}_{\hat{S}}\big[Q_{j,S_j}\big]$. Let λ_{j,S_j} be distributed according to Q_{j,S_j}. By Jensen's inequality,

$$D(Q_{j,S_j} \,||\, P_{j,S_j}) = \mathbb{E}_{\lambda_{j,S_j}}\left[\log\frac{Q_{j,S_j}(\lambda_{j,S_j})}{P_{j,S_j}(\lambda_{j,S_j})}\right] \leq \log \mathbb{E}_{\lambda_{j,S_j}}\left[\frac{Q_{j,S_j}(\lambda_{j,S_j})}{P_{j,S_j}(\lambda_{j,S_j})}\right]. \tag{153}$$

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

$$D(Q_{j,S_j} \,||\, P_{j,S_j}) = \mathbb{E}_{\lambda_{j,S_j}}\left[\log\frac{Q_{j,S_j}(\lambda_{j,S_j})}{P_{j,S_j}(\lambda_{j,S_j})}\right] \leq \log\left(\frac{1}{\delta}\,\mathbb{E}_{\lambda_{j,S_j},Z,\hat{S},S}\left[\frac{Q_{j,S_j}(\lambda_{j,S_j})}{P_{j,S_j}(\lambda_{j,S_j})}\right]\right). \tag{154}$$

Since λ_{j,S_j} is a discrete random variable, $Q_{j,S_j}(\lambda_{j,S_j}) \leq 1$. Hence,

$$\log\left(\frac{1}{\delta}\,\mathbb{E}_{\lambda_{j,S_j},Z,\hat{S},S}\left[\frac{Q_{j,S_j}(\lambda_{j,S_j})}{P_{j,S_j}(\lambda_{j,S_j})}\right]\right) \leq \log\left(\frac{1}{\delta}\,\mathbb{E}_{\lambda_{j,S_j},Z,\hat{S},S}\left[\frac{1}{P_{j,S_j}(\lambda_{j,S_j})}\right]\right). \tag{155}$$

Recall that $\mathbb{E}_{\hat{S}}\big[Q_{j,S_j}\big] = P_{j,S_j}$. Let λ'_{j,S_j} be distributed according to P_{j,S_j}. Since the argument of the expectation is now independent of $\hat{S}$,

$$\log\left(\frac{1}{\delta}\,\mathbb{E}_{\lambda_{j,S_j},Z,\hat{S},S}\left[\frac{1}{P_{j,S_j}(\lambda_{j,S_j})}\right]\right) = \log\left(\frac{1}{\delta}\,\mathbb{E}_{\lambda'_{j,S_j},Z,S}\left[\frac{1}{P_{j,S_j}(\lambda_{j,S_j})}\right]\right) \tag{156}$$

$$\leq \log\left(\frac{1}{\delta}\sup_{Z,S}\mathbb{E}_{\lambda'_{j,S_j}}\left[\frac{1}{P_{j,S_j}(\lambda_{j,S_j})}\right]\right). \tag{157}$$

Now, let $\Lambda_{j,S_j}(Z,S)$ denote the set of all possible values that λ'_{j,S_j} can take given (Z,S). Then,

$$\log\left(\frac{1}{\delta}\sup_{Z,S}\mathbb{E}_{\lambda'_{j,S_j}}\left[\frac{1}{P_{j,S_j}(\lambda_{j,S_j})}\right]\right) = \log\left(\frac{1}{\delta}\sup_{Z,S}\sum_{\lambda'_{j,S_j}\in\Lambda_{j,S_j}(Z,S)}\frac{P_{j,S_j}(\lambda_{j,S_j})}{P_{j,S_j}(\lambda_{j,S_j})}\right) \tag{158}$$

$$= \log\left(\frac{1}{\delta}\sup_{Z,S}\big|\Lambda_{j,S_j}(Z,S)\big|\right). \tag{159}$$

Now, note that since λ_{j,S_j} is averaged over R, it is a function of H_{j,S_j} given (Z,S). Furthermore, the inputs X_{j,S_j} are fixed. Therefore, as argued in the proof of Corollary 4, the number of different values that H_{j,S_j} can take given (Z,S) is at most $g_{\mathcal{H}}(2\hat{n})$.[3] Thus, by combining (153)-(159), we get

$$D(Q_{j,S_j} \,||\, P_{j,S_j}) \leq \log\left(\frac{g_{\mathcal{H}}(2\hat{n})}{\delta}\right) \leq d_N \log\left(\binom{N}{2}\frac{2e\hat{n}}{d_N}\right) + \log\frac{1}{\delta}. \tag{160}$$

[3] If we had instead used the result of Remark 1 and followed analogous steps, we would instead get $g_{\mathcal{H}}(2n\hat{n})$ as an upper bound of the KL divergence.

Next, we turn to $D(Q^{-\hat{S}} \,||\, P^{-\hat{S}})$. Recall that $Q^{-\hat{S}}$ denotes the conditional distribution of $\lambda^{-\hat{S}}$ given $(Z, \hat{S}, S)$, and $P^{-\hat{S}} = \mathbb{E}_{S^{-\hat{S}}}\left[Q^{-\hat{S}}\right]$. Let $\lambda^{-\hat{S}}$ be distributed according to $Q^{-\hat{S}}$. Again, by Jensen's inequality,

$$D(Q^{-\hat{S}} \,||\, P^{-\hat{S}}) = \mathbb{E}_{\lambda^{-\hat{S}}}\left[\log \frac{Q^{-\hat{S}}(\lambda^{-\hat{S}})}{P^{-\hat{S}}(\lambda^{-\hat{S}})}\right] \leq \log \mathbb{E}_{\lambda^{-\hat{S}}}\left[\frac{Q^{-\hat{S}}(\lambda^{-\hat{S}})}{P^{-\hat{S}}(\lambda^{-\hat{S}})}\right]. \tag{161}$$

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

$$D(Q^{-\hat{S}} \,||\, P^{-\hat{S}}) = \mathbb{E}_{\lambda^{-\hat{S}}}\left[\log \frac{Q^{-\hat{S}}(\lambda^{-\hat{S}})}{P^{-\hat{S}}(\lambda^{-\hat{S}})}\right] \leq \log\left(\frac{1}{\delta}\,\mathbb{E}_{\lambda^{-\hat{S}},Z,\hat{S},S}\left[\frac{Q^{-\hat{S}}(\lambda^{-\hat{S}})}{P^{-\hat{S}}(\lambda^{-\hat{S}})}\right]\right). \tag{162}$$

Since $\lambda^{-\hat{S}}$ is a discrete random variable, $Q^{-\hat{S}}(\lambda^{-\hat{S}}) \leq 1$. Therefore,

$$\log\left(\frac{1}{\delta}\,\mathbb{E}_{\lambda^{-\hat{S}},Z,\hat{S},S}\left[\frac{Q^{-\hat{S}}(\lambda^{-\hat{S}})}{P^{-\hat{S}}(\lambda^{-\hat{S}})}\right]\right) \leq \log\left(\frac{1}{\delta}\,\mathbb{E}_{\lambda^{-\hat{S}},Z,\hat{S},S}\left[\frac{1}{P^{-\hat{S}}(\lambda^{-\hat{S}})}\right]\right). \tag{163}$$

Now, let $\lambda'^{-\hat{S}}$ be distributed according to $P^{-\hat{S}}$. Since the argument of the expectation is now independent of $S^{-\hat{S}}$,

$$\log\left(\frac{1}{\delta}\,\mathbb{E}_{\lambda^{-\hat{S}},Z,\hat{S},S}\left[\frac{1}{P^{-\hat{S}}(\lambda^{-\hat{S}})}\right]\right) = \log\left(\frac{1}{\delta}\,\mathbb{E}_{\lambda'^{-\hat{S}},Z,\hat{S},S^{\hat{S}}}\left[\frac{1}{P^{-\hat{S}}(\lambda^{-\hat{S}})}\right]\right) \tag{164}$$

$$\leq \log\left(\frac{1}{\delta}\sup_{Z,\hat{S},S^{\hat{S}}}\mathbb{E}_{\lambda'^{-\hat{S}}}\left[\frac{1}{P^{-\hat{S}}(\lambda^{-\hat{S}})}\right]\right). \tag{165}$$

Now, let $\Lambda^{-\hat{S}}(Z,S)$ denote the set of all possible values that $\lambda'^{-\hat{S}}$ can take given $(Z, \hat{S}, S^{\hat{S}})$. Then,

$$\log\left(\frac{1}{\delta}\sup_{Z,\hat{S},S^{\hat{S}}}\mathbb{E}_{\lambda'^{-\hat{S}}}\left[\frac{1}{P^{-\hat{S}}(\lambda^{-\hat{S}})}\right]\right) = \log\left(\frac{1}{\delta}\sup_{Z,\hat{S},S^{\hat{S}}}\sum_{\lambda'^{-\hat{S}}\in\Lambda^{-\hat{S}}(Z,S)}\frac{P^{-\hat{S}}(\lambda^{-\hat{S}})}{P^{-\hat{S}}(\lambda^{-\hat{S}})}\right) \tag{166}$$

$$= \log\left(\frac{1}{\delta}\sup_{Z,\hat{S},S^{\hat{S}}}\left|\Lambda^{-\hat{S}}(Z,S)\right|\right). \tag{167}$$

Note that, given Z, the losses $\lambda^{-i,\hat{S}_i}$ are a function of the predictions $F^{i,-\hat{S}_i}$. Furthermore, given $(Z, \hat{S}, S^{\hat{S}})$, the inputs $H^{i,-\hat{S}_i}$ are fixed. This is the case since U is independent from $S^{-\hat{S}}$. Thus, similar to previous arguments, given $(Z, \hat{S}, S^{\hat{S}})$, $F^{i,-\hat{S}_i}$ can take at most $g_{\mathcal{F}}(2n)$ different values for each i. Therefore, $F^{-\hat{S}}$ can take at most $g_{\mathcal{F}}(2n)^{\hat{n}}$ values. Thus, by combining (161)-(167), we get

$$D(Q^{-\hat{S}} \,||\, P^{-\hat{S}}) \leq \log\left(\frac{g_{\mathcal{F}}(2n)^{\hat{n}}}{\delta}\right) \leq \hat{n} d_{\text{VC}} \log\left(\frac{2en}{d_{\text{VC}}}\right) + \log\frac{1}{\delta}, \tag{168}$$

where we used Lemma 4. Thus, by using a union bound, we can combine (114), (160), and (168), with $\delta \to \delta/3$, to conclude that with probability at least $1-\delta$ under the draw of $(Z, \hat{S}, S)$,

$$\left|L_{\mathcal{D}} - \widehat{L}\right| \leq \sqrt{\frac{2d_N \log\left(\binom{N}{2}\frac{2e\hat{n}}{d_N}\right) + \log\frac{3}{\delta} + 2\log\left(\frac{6n\sqrt{\hat{n}}}{\delta}\right)}{\hat{n}-1}} + \sqrt{\frac{2\hat{n} d_{\text{VC}} \log\left(\frac{2en}{d_{\text{VC}}}\right) + \log\frac{3}{\delta} + 2\log\left(\frac{6\sqrt{n\hat{n}}}{\delta}\right)}{n\hat{n}-1}}. \tag{169}$$

Under the assumption that $n, \hat{n} \geq 2$, by similar arguments as in the proof of Theorem 4, we find that, for some constants C_1 and C_2,

$$\left|L_{\mathcal{D}} - \widehat{L}\right| \leq C_1 \sqrt{\frac{d_N \log\left(\binom{N}{2}\frac{\hat{n}}{d_N}\right) + \log\left(\frac{n\sqrt{\hat{n}}}{\delta}\right)}{\hat{n}}} + C_2 \sqrt{\frac{d_{\text{VC}} \log\left(\frac{n}{d_{\text{VC}}}\right) + \log\left(\frac{\sqrt{n}}{\delta}\right)}{n}}. \tag{170}$$

This establishes (23).

We now turn to (24). Let λ be distributed according to Q. First, by Jensen's inequality,

$$D(Q \,||\, P) = \mathbb{E}_\lambda\left[\log \frac{Q(\lambda)}{P(\lambda)}\right] \leq \log \mathbb{E}_\lambda\left[\frac{Q(\lambda)}{P(\lambda)}\right]. \tag{171}$$

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

$$D(Q \,||\, P) = \mathbb{E}_\lambda\left[\log \frac{Q(\lambda)}{P(\lambda)}\right] \leq \log\left(\frac{1}{\delta}\,\mathbb{E}_{\lambda, Z, \hat{S}, S}\left[\frac{Q(\lambda)}{P(\lambda)}\right]\right). \tag{172}$$

Since λ is a discrete random variable, $Q(\lambda) \leq 1$. Hence,

$$D(Q \,||\, P) = \mathbb{E}_\lambda\left[\log \frac{Q(\lambda)}{P(\lambda)}\right] \leq \log\left(\frac{1}{\delta}\,\mathbb{E}_{\lambda, Z, \hat{S}, S}\left[\frac{1}{P(\lambda)}\right]\right). \tag{173}$$

Recall that $\mathbb{E}_{\hat{S}, S}[Q] = P$. Let λ' be distributed according to P. Since the argument of the expectation is now independent of $(\hat{S}, S)$,

$$\log\left(\frac{1}{\delta}\,\mathbb{E}_{\lambda, Z, \hat{S}, S}\left[\frac{1}{P(\lambda)}\right]\right) \leq \log\left(\frac{1}{\delta}\,\mathbb{E}_{\lambda', Z}\left[\frac{1}{P(\lambda')}\right]\right) \tag{174}$$

$$\leq \log\left(\frac{1}{\delta}\sup_Z \mathbb{E}_{\lambda'}\left[\frac{1}{P(\lambda')}\right]\right). \tag{175}$$

Let $\Lambda(Z)$ denote the set of all possible values that λ' can take given Z. Then,

$$\log\left(\frac{1}{\delta}\sup_Z \mathbb{E}_{\lambda'}\left[\frac{1}{P(\lambda')}\right]\right) = \log\left(\frac{1}{\delta}\sup_Z \sum_{\lambda' \in \Lambda(Z)} \frac{P(\lambda')}{P(\lambda')}\right) \tag{176}$$

$$= \log\left(\frac{1}{\delta}\sup_Z |\Lambda(Z)|\right). \tag{177}$$

Since the map from predictions to losses is surjective, $|\Lambda(Z)|$ is bounded by the number of possible predictions F given Z. We can bound this as follows. First, the number of possible different values for H given Z is at most $g_{\mathcal{H}}(4n\hat{n})$. Given a fixed H, the number of possible values that F can take is at most $(g_{\mathcal{F}}(2n))^{2\hat{n}}$, since the $2n$ inputs to each of the $2\hat{n}$ task-specific functions are fixed. Therefore, the total number of possible values for F given Z is at most $g_{\mathcal{H}}(4n\hat{n})(g_{\mathcal{F}}(2n))^{2\hat{n}}$. Hence,

$$\log\left(\frac{1}{\delta}\sup_Z |\Lambda(Z)|\right) \leq \log(g_{\mathcal{H}}(4n\hat{n})) + 2\hat{n}\log(g_{\mathcal{F}}(2n)) + \log\frac{1}{\delta} \tag{178}$$

$$\leq d_N \log\left(\binom{N}{2}\frac{4en\hat{n}}{d_N}\right) + 2\hat{n}d_{\text{VC}}\log\left(\frac{2en}{d_{\text{VC}}}\right) + \log\frac{1}{\delta} \tag{179}$$

where we used Lemma 4. Substituting this into (18), using a union bound and letting $\delta \to \delta/2$, we find that with probability at least $1-\delta$ under the draw of $(Z, \hat{S}, S)$,

$$\left|L_{\mathcal{D}}(Z,\hat{S},S) - \widehat{L}(Z,\hat{S},S)\right| \leq \sqrt{\frac{2d_N \log\left(\binom{N}{2}\frac{4en\hat{n}}{d_N}\right) + 4\hat{n}d_{\mathrm{VC}}\log\left(\frac{2en}{d_{\mathrm{VC}}}\right) + 2\log\frac{2}{\delta} + 2\log\left(\frac{2\sqrt{n\hat{n}}}{\delta}\right)}{n\hat{n}-1}}. \quad (180)$$

Assuming that $n, \hat{n} \geq 2$, the desired result in (21) follows by upper-bounding constants by using similar arguments as in the proof of Theorem 4.

□

B Bound for the Excess Risk

We now turn to excess risk bounds. In Corollary 7, we present the formal statement of the excess risk bound in Section 4.2. In Corollary 8, we state an excess risk bound for a randomly drawn new task, which obviates the need of a task diversity assumption.

In order to derive excess risk bounds, we need to introduce some technical tools. First, we need to consider *oracle* algorithms, that is, algorithms that output minimizers of the population loss. Specifically, the oracle meta learner knows the task distribution $\mathcal{D}$, while the oracle base learner knows the indexed set of in-task distributions $\{D_\tau : \tau \in \mathcal{T}\}$. While these algorithms have access to the data distributions, and are thus not of practical interest, they are useful as a proof technique, and can be analyzed in the same way as realistic algorithms. Second, in order to allow the oracle base learner to minimize the population loss for a given task, we need to extend the input to the base learner to include the identity of the task $\tau \in \mathcal{T}$. Thus, the base learner is a mapping $\mathcal{A} : \mathcal{Z}^n \times \mathcal{T} \times \mathcal{R} \times \mathcal{U} \to \mathcal{W}$. For the case of a base learner that minimizes the empirical risk, the task identity is irrelevant, so the input from $\mathcal{T}$ does not affect the output. Conversely, for an oracle base learner, the training samples are irrelevant, so only the input from $\mathcal{T}$ affects the output. Finally, our information-theoretic bounds pertain to a test loss, rather than the population loss. While these are equal for average bounds, there is a small discrepancy for the high-probability bounds. In order to handle excess risk bounds and oracle algorithms that depend on the population loss, we need to convert between the two by using a Hoeffding bound, as discussed in [23, Thm. 3]. The extra terms that this additional step leads to are typically negligible compared to the dominant complexity terms.

For concreteness, we focus only on high-probability excess risk bounds derived on the basis of the one-step square-root bound in Corollary 6. However, note that excess risk bounds based on the other high-probability bounds can be obtained by suitably substituting these alternative bounds in the proofs. Average excess risk bounds can also be derived by an analogous procedure. First, using the task diversity assumption of [5], we derive an excess risk bound for a fixed target task.

Corollary 7. *Consider the setting of Corollary 6 and a fixed task τ_0. Let $Z^0 \in \mathcal{Z}^{2\times m}$ be a matrix of $2m$ samples generated independently according to $\mathcal{D}_{\tau_0}$, the data distribution for task τ_0. Let S^0 be an m-dimensional random vector with elements generated independently from a $\mathrm{Bern}(1/2)$ distribution, and let the training set $Z^0_{S^0}$ and test set $Z^0_{-S^0}$ be constructed in the same way as the training and test sets for tasks $1, \ldots, \hat{n}$. To simplify notation, let $Z^{0,0} = Z^{0,1} = Z^0$. Denote the population loss for the ith observed task when using the base learner $\mathcal{A}'$ with the representation h' as*

$$L_{\mathcal{D}}(i, \mathcal{A}', h') = \mathbb{E}_{R_i, \tilde{Z}\sim\mathcal{D}_{\tau_{i,\hat{S}_i}}}\left[\ell\left(\mathcal{A}'\left(Z^{i,\hat{S}_i}_{S^i}, \tau_{i,\hat{S}_i}, R_i, h'\right), \tilde{Z}\right)\right]. \quad (181)$$

Similarly, denote the population loss for the ith unobserved task when using the base learner $\mathcal{A}'$ with the representation h' as

$$L_{\mathcal{D}}(-i, \mathcal{A}', h') = \mathbb{E}_{R_i, \tilde{Z}\sim\mathcal{D}_{\tau_{i,-\hat{S}_i}}}\left[\ell\left(\mathcal{A}'\left(Z^{i,-\hat{S}_i}_{S^i}, \tau_{i,-\hat{S}_i}, R_i, h'\right), \tilde{Z}\right)\right]. \quad (182)$$

Let $\mathcal{A}^*$ denote an oracle learner that satisfies $L_{\mathcal{D}}(i, \mathcal{A}^*, h') = \min_{\mathcal{A}'} L_{\mathcal{D}}(i, \mathcal{A}', h')$ for all h'. Assume that $h^* = \arg\min_{h'} \min_{\mathcal{A}'} L_{\mathcal{D}}(i, \mathcal{A}', h')$ for all $\tau_{i,\hat{S}_i}$ and that $h^* \in \mathcal{H}$. Thus, the same representation h^* is optimal for all tasks. Let $\hat{\mathcal{A}}$ and $\mathcal{A}$ be empirical risk minimizers, and let $\hat{h} = \hat{\mathcal{A}}(\hat{R}, Z_S^{\hat{S}})$. Finally, assume that the supersample satisfies a task-diversity assumption, so that for some ν and ϵ,

$$\sup_{\tau_0} L_{\mathcal{D}}(0, \mathcal{A}^*, \hat{h}) - L_{\mathcal{D}}(0, \mathcal{A}^*, h^*) \leq \nu^{-1} \left(L_{\mathcal{D}}(-1{:}\,\hat{n}, \mathcal{A}^*, \hat{h}) - L_{\mathcal{D}}(-1{:}\,\hat{n}, \mathcal{A}^*, h^*) \right) + \epsilon. \quad (183)$$

Then, there exist constants C_1 and C_2 such that, with probability at least $1 - \delta$ under the draw of $(Z, \hat{S}, S, Z^0, S^0)$, we have

$$L_{\mathcal{D}}(0, \mathcal{A}, \hat{h}) - L_{\mathcal{D}}(0, \mathcal{A}^*, h^*) \leq C_1 \sqrt{\frac{d_{\mathrm{VC}} \log\left(\frac{\sqrt{m}}{d_{\mathrm{VC}}}\right) + \log\left(\frac{\sqrt{m}}{\delta}\right)}{m}}$$
$$+ C_2 \nu^{-1} \sqrt{\frac{d_N \log\left(\binom{N}{2} \frac{n\hat{n}}{d_N}\right) + \hat{n} d_{\mathrm{VC}} \log\left(\frac{n}{d_{\mathrm{VC}}}\right) + \log\left(\frac{\sqrt{n\hat{n}}}{\delta}\right)}{n\hat{n}}} + \epsilon. \quad (184)$$

Proof. We will use the following shorthands. When using the algorithm $\mathcal{A}'$ for task i based on the representation h', we let $L_{\mathcal{D}}(i, \mathcal{A}', h')$ denote the population loss, $\hat{L}(i, \mathcal{A}', h')$ denote the training loss, and $\widetilde{L}(i, \mathcal{A}', h')$ denote the test loss on a test set of the same size as the training set. Formally,

$$L_{\mathcal{D}}(i, \mathcal{A}', h') = \mathbb{E}_{R_i, \tilde{Z} \sim \mathcal{D}_{\tau_{i,\hat{S}_i}}} \left[\ell\left(\mathcal{A}'\left(Z_{S^i}^{i,\hat{S}_i}, \tau_{i,\hat{S}_i}, R_i, h' \right), \tilde{Z} \right) \right], \quad (185)$$

$$\hat{L}(i, \mathcal{A}', h') = \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}_{R_i} \left[\ell\left(\mathcal{A}'\left(Z_{S^i}^{i,\hat{S}_i}, \tau_{i,\hat{S}_i}, R_i, h' \right), Z_{j,S_j^i}^{i,\hat{S}_i} \right) \right], \quad (186)$$

$$\widetilde{L}(i, \mathcal{A}', h') = \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}_{R_i} \left[\ell\left(\mathcal{A}'\left(Z_{S^i}^{i,\hat{S}_i}, \tau_{i,\hat{S}_i}, R_i, h' \right), Z_{j,-S_j^i}^{i,\hat{S}_i} \right) \right]. \quad (187)$$

As a shorthand, we let $L_{\mathcal{D}}(1{:}\,\hat{n}, \mathcal{A}', h') = \frac{1}{\hat{n}} \sum_{i=1}^{\hat{n}} L_{\mathcal{D}}(i, \mathcal{A}', h')$, and we use the same convention for $\hat{L}(1{:}\,\hat{n}, \mathcal{A}', h')$ and $\widetilde{L}(1{:}\,\hat{n}, \mathcal{A}', h')$. Furthermore, to indicate losses on unobserved tasks we negate the task index. Thus,

$$\hat{L}(-i, \mathcal{A}', h') = \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}_{R_i} \left[\ell\left(\mathcal{A}'\left(Z_{S^i}^{i,-\hat{S}_i}, \tau_{i,-\hat{S}_i}, R_i, h' \right), Z_{j,S_j^i}^{i,-\hat{S}_i} \right) \right], \quad (188)$$

with analogous notation for the test and population losses.

The base learner $\mathcal{A}$ that we consider is an empirical risk minimizer, which satisfies for all h'

$$\hat{L}(i, \mathcal{A}, h') = \min_{\mathcal{A}'} \hat{L}(i, \mathcal{A}', h'). \quad (189)$$

For our analysis, we use an oracle learner $\mathcal{A}^*$, which outputs the minimizer of the population loss for the given task. While this is not a realistic learning algorithm in practice, as it depends on the data distribution, it is useful as an analysis tool. Formally, for all h',

$$L_{\mathcal{D}}(i, \mathcal{A}^*, h') = \min_{\mathcal{A}'} L_{\mathcal{D}}(i, \mathcal{A}', h'). \quad (190)$$

Finally, we let $\hat{h}$ be a representation that minimizes the empirical risk over the $\hat{n}$ training tasks and h^* be an optimal representation, i.e.

$$\hat{h} \in \arg\min_{h'} \hat{L}(1{:}\,\hat{n}, \mathcal{A}, h'). \quad (191)$$

$$h^* \in \arg\min_{h'} L_{\mathcal{D}}(i, \mathcal{A}^*, h'). \quad (192)$$

By assumption, h^* is the same for any task $\tau_{i,k}$.

In the proof, we need to convert between test losses and population losses. By definition, test data is independent from the hypothesis, so standard concentration inequalities can be applied to bound the difference between the test and population loss. The following lemma follows immediately from Hoeffding's inequality [38, Prop. 2.5], as argued in [23, Thm. 3].

Lemma 5. *Let $\widetilde{L}(i, \mathcal{A}', h')$ be a test loss based on m samples. Then, with probability at least $1-\delta$,*

$$\left|\widetilde{L}(i, \mathcal{A}', h') - L_{\mathcal{D}}(i, \mathcal{A}', h')\right| \leq \sqrt{\frac{\log \frac{2}{\delta}}{2m}}. \tag{193}$$

Proof. The test loss $\widetilde{L}(i, \mathcal{A}', h')$ is the average of m independent samples of a bounded random variable with mean $L_{\mathcal{D}}(i, \mathcal{A}', h')$. Therefore, the result follows by Hoeffding's inequality [38, Prop. 2.5]. □

This result allows us to convert between test losses and population losses at the cost of a term that is typically negligible in comparison to the complexity terms.

With these tools and notations in place, we are ready to derive excess risk bounds. The aim is to upper-bound the excess risk by an expression consisting of differences between training and test losses, for which we can apply our generalization bounds. Starting from the excess risk on task τ_0, which is our fixed target task, we get

$$\begin{aligned} L_{\mathcal{D}}(0, \mathcal{A}, \hat{h}) - L_{\mathcal{D}}(0, \mathcal{A}^*, h^*) &= L_{\mathcal{D}}(0, \mathcal{A}, \hat{h}) - L_{\mathcal{D}}(0, \mathcal{A}^*, \hat{h}) + L_{\mathcal{D}}(0, \mathcal{A}^*, \hat{h}) - L_{\mathcal{D}}(0, \mathcal{A}^*, h^*) \\ &\leq L_{\mathcal{D}}(0, \mathcal{A}, \hat{h}) - L_{\mathcal{D}}(0, \mathcal{A}^*, \hat{h}) + D. \end{aligned} \tag{194}$$

Here, $D = \sup_{\tau_0} L_{\mathcal{D}}(0, \mathcal{A}^*, \hat{h}) - L_{\mathcal{D}}(0, \mathcal{A}^*, h^*)$ is the worst-case representation difference [5], which we will later bound using a task diversity assumption. Next, by Lemma 5, with probability at least $1-2\delta$,

$$L_{\mathcal{D}}(0, \mathcal{A}, \hat{h}) - L_{\mathcal{D}}(0, \mathcal{A}^*, \hat{h}) + D \leq \widetilde{L}(0, \mathcal{A}, \hat{h}) - \widetilde{L}(0, \mathcal{A}^*, \hat{h}) + D + 2\sqrt{\frac{\log \frac{2}{\delta}}{m}}. \tag{195}$$

Next, we use the risk decomposition

$$\widetilde{L}(0, \mathcal{A}, \hat{h}) - \widetilde{L}(0, \mathcal{A}^*, \hat{h}) \tag{196}$$

$$= \widetilde{L}(0, \mathcal{A}, \hat{h}) - \hat{L}(0, \mathcal{A}, \hat{h}) + \hat{L}(0, \mathcal{A}, \hat{h}) - \hat{L}(0, \mathcal{A}^*, \hat{h}) + \hat{L}(0, \mathcal{A}^*, \hat{h}) - \widetilde{L}(0, \mathcal{A}^*, \hat{h}) \tag{197}$$

$$\leq \widetilde{L}(0, \mathcal{A}, \hat{h}) - \hat{L}(0, \mathcal{A}, \hat{h}) + \hat{L}(0, \mathcal{A}^*, \hat{h}) - \widetilde{L}(0, \mathcal{A}^*, \hat{h}), \tag{198}$$

where the last step follows because $\hat{L}(0, \mathcal{A}, \hat{h}) \leq \hat{L}(0, \mathcal{A}^*, \hat{h})$, since $\mathcal{A}$ is an empirical risk minimizer. Notice that the resulting expression is the difference between test and training losses on task τ_0 for two different algorithms. These terms are simply the generalization gaps for a conventional learning setting. These terms can be bounded by applying Corollary 6, but for the case where $\hat{n} = 1$ and $\mathcal{H} = \{\hat{h}\}$, which implies that $d_N = 0$. We conclude that there exists a constant C_1 such that, with probability at least $1-\delta$,

$$\widetilde{L}(0, \mathcal{A}, \hat{h}) - \hat{L}(0, \mathcal{A}, \hat{h}) + \hat{L}(0, \mathcal{A}^*, \hat{h}) - \widetilde{L}(0, \mathcal{A}^*, \hat{h}) \leq C_1 \sqrt{\frac{d_{\text{VC}} \log\left(\frac{\sqrt{m}}{d_{\text{VC}}}\right) + \log\left(\frac{\sqrt{m}}{\delta}\right)}{m}}. \tag{199}$$

It remains to bound D. First, by the task diversity assumption,

$$D = \sup_{\tau_0} L_{\mathcal{D}}(0, \mathcal{A}^*, \hat{h}) - L_{\mathcal{D}}(0, \mathcal{A}^*, h^*) \tag{200}$$

$$\leq \nu^{-1}\left(L_{\mathcal{D}}(-1{:}\hat{n}, \mathcal{A}^*, \hat{h}) - L_{\mathcal{D}}(-1{:}\hat{n}, \mathcal{A}^*, h^*)\right) + \epsilon. \tag{201}$$

We note here that, while the way that [5] uses the assumption of task diversity requires that the difference between the minimum population losses for task τ_0 based on $\hat{h}$ and h^* is controlled by the corresponding risks for tasks $1{:}\,\hat{n}$, i.e., the tasks upon which $\hat{h}$ is chosen, we instead assume that it is controlled by the corresponding losses for tasks $-1{:}\,\hat{n}$, i.e., tasks that are independent from $\hat{h}$. In this sense, the diversity assumption that we use is arguably weaker.

By a risk decomposition, we get

$$
\begin{aligned}
& L_{\mathcal{D}}(-1{:}\,\hat{n}, \mathcal{A}^*, \hat{h}) - L_{\mathcal{D}}(-1{:}\,\hat{n}, \mathcal{A}^*, h^*) && (202)\\
=& L_{\mathcal{D}}(-1{:}\,\hat{n}, \mathcal{A}^*, \hat{h}) - L_{\mathcal{D}}(-1{:}\,\hat{n}, \mathcal{A}, \hat{h}) + L_{\mathcal{D}}(-1{:}\,\hat{n}, \mathcal{A}, \hat{h}) - L_{\mathcal{D}}(-1{:}\,\hat{n}, \mathcal{A}^*, h^*) && (203)\\
\leq& L_{\mathcal{D}}(-1{:}\,\hat{n}, \mathcal{A}, \hat{h}) - L_{\mathcal{D}}(-1{:}\,\hat{n}, \mathcal{A}^*, h^*), && (204)
\end{aligned}
$$

where the last step follows since $L_{\mathcal{D}}(-1:\,\hat{n}, \mathcal{A}^*, \hat{h}) \leq L_{\mathcal{D}}(-1:\,\hat{n}, \mathcal{A}, \hat{h})$. By Lemma 5, with probability $1 - 2\delta$,

$$
L_{\mathcal{D}}(-1{:}\,\hat{n}, \mathcal{A}, \hat{h}) - L_{\mathcal{D}}(-1{:}\,\hat{n}, \mathcal{A}^*, h^*) \leq \widetilde{L}(-1{:}\,\hat{n}, \mathcal{A}, \hat{h}) - \widetilde{L}(-1{:}\,\hat{n}, \mathcal{A}^*, h^*) + 2\sqrt{\frac{\log \frac{2}{\delta}}{2n}}. \quad (205)
$$

By a risk decomposition, we find that

$$
\begin{aligned}
& \widetilde{L}(-1{:}\,\hat{n}, \mathcal{A}, \hat{h}) - \widetilde{L}(-1{:}\,\hat{n}, \mathcal{A}^*, h^*)\\
\leq& \widetilde{L}(-1{:}\,\hat{n}, \mathcal{A}, \hat{h}) - \hat{L}(1{:}\,\hat{n}, \mathcal{A}, \hat{h}) + \hat{L}(1{:}\,\hat{n}, \mathcal{A}, \hat{h}) - \hat{L}(1{:}\,\hat{n}, \mathcal{A}^*, h^*) + \hat{L}(1{:}\,\hat{n}, \mathcal{A}^*, h^*) - \widetilde{L}(-1{:}\,\hat{n}, \mathcal{A}^*, h^*)\\
\leq& \widetilde{L}(-1{:}\,\hat{n}, \mathcal{A}, \hat{h}) - \hat{L}(1{:}\,\hat{n}, \mathcal{A}, \hat{h}) + \hat{L}(1{:}\,\hat{n}, \mathcal{A}^*, h^*) - \widetilde{L}(-1{:}\,\hat{n}, \mathcal{A}^*, h^*), && (206)
\end{aligned}
$$

where the last step follows since $\hat{L}(1{:}\,\hat{n}, \mathcal{A}, \hat{h}) \leq \hat{L}(1{:}\,\hat{n}, \mathcal{A}^*, h^*)$. Now, notice that the resulting expression consists of the differences between the unobserved test losses and observed training losses for two different learning algorithms. This means that we can apply Corollary 6 to find that there exists a constant C_2 such that, with probability at least $1 - \delta$,

$$
\begin{aligned}
& \widetilde{L}(-1{:}\,\hat{n}, \mathcal{A}, \hat{h}) - \hat{L}(1{:}\,\hat{n}, \mathcal{A}, \hat{h}) + \hat{L}(1{:}\,\hat{n}, \mathcal{A}^*, h^*) - \widetilde{L}(-1{:}\,\hat{n}, \mathcal{A}^*, h^*)\\
& \qquad \leq C_2 \sqrt{\frac{d_N \log\left(\binom{N}{2}\frac{n\hat{n}}{d_N}\right) + \hat{n} d_{\mathrm{VC}} \log\left(\frac{n}{d_{\mathrm{VC}}}\right) + \log\left(\frac{\sqrt{n\hat{n}}}{\delta}\right)}{n\hat{n}}}. \quad (207)
\end{aligned}
$$

Thus, by putting it all together, using a union bound to combine the probabilistic inequalities, we find that there exists constants C_1, C_2, C_3 such that, with probability at least $1 - \delta$,

$$
\begin{aligned}
L_{\mathcal{D}}(0, \mathcal{A}, \hat{h}) - L_{\mathcal{D}}(0, \mathcal{A}^*, h^*) \leq\ & C_1 \sqrt{\frac{d_{\mathrm{VC}} \log\left(\frac{\sqrt{m}}{d_{\mathrm{VC}}}\right) + \log\left(\frac{\sqrt{m}}{\delta}\right)}{m}}\\
& + C_2 \nu^{-1} \sqrt{\frac{d_N \log\left(\binom{N}{2}\frac{n\hat{n}}{d_N}\right) + \hat{n} d_{\mathrm{VC}} \log\left(\frac{n}{d_{\mathrm{VC}}}\right) + \log\left(\frac{\sqrt{n\hat{n}}}{\delta}\right)}{n\hat{n}}} + \epsilon, \quad (208)
\end{aligned}
$$

where we note that the penalty terms arising from the union bound and converting between test and population losses have been absorbed using the constants.

□

Thus, under the assumption of task diversity, we obtained an excess risk bound for a fixed target task, as was done in [5]. However, if we are interested in bounding the excess risk for a new, randomly drawn task, rather than a fixed target, task diversity is not necessary. In the following corollary, we demonstrate this by deriving an excess risk bound with respect to the population loss for a new, random task. While we only present a bound based on Corollary 6, similar excess risk bounds can be derived for the average case and from the other high-probability bounds.

Corollary 8. *Consider the setting of Corollary 6. Assume that $\mathcal{A}$ is an empirical risk minimizer, that $\mathcal{A}^*$ is an oracle algorithm, and let*

$$\hat{h} \in \arg\min_{h'} \hat{L}(1{:}\hat{n}, \mathcal{A}, h'), \tag{209}$$

$$h^* \in \arg\min_{h'} L_{\mathcal{D}}(-i, \mathcal{A}^*, h'). \tag{210}$$

Then, there exists a constant C such that, with probability at least $1-\delta$ under $(Z, \hat{S}, S)$,

$$L_{\mathcal{D}}(-i, \mathcal{A}, \hat{h}) - L_{\mathcal{D}}(-i, \mathcal{A}^*, h^*) \leq C\sqrt{\frac{d_N \log\left(\binom{N}{2}\frac{n\hat{n}}{d_N}\right) + \hat{n} d_{\text{VC}} \log\left(\frac{n}{d_{\text{VC}}}\right) + \log\left(\frac{\sqrt{n\hat{n}}}{\delta}\right)}{n\hat{n}}}. \tag{211}$$

Proof. We begin with the risk decomposition

$$\begin{aligned} L_{\mathcal{D}}(-i, \mathcal{A}, \hat{h}) - L_{\mathcal{D}}(-i, \mathcal{A}^*, h^*) &= L_{\mathcal{D}}(-i, \mathcal{A}, \hat{h}) - \hat{L}(1{:}\hat{n}, \mathcal{A}, \hat{h}) + \hat{L}(1{:}\hat{n}, \mathcal{A}, \hat{h}) \\ &\quad - \hat{L}(1{:}\hat{n}, \mathcal{A}^*, h^*) + \hat{L}(1{:}\hat{n}, \mathcal{A}^*, h^*) - L_{\mathcal{D}}(-i, \mathcal{A}^*, h^*) \\ &\leq L_{\mathcal{D}}(-i, \mathcal{A}, \hat{h}) - \hat{L}(1{:}\hat{n}, \mathcal{A}, \hat{h}) + \hat{L}(1{:}\hat{n}, \mathcal{A}^*, h^*) - L_{\mathcal{D}}(-i, \mathcal{A}^*, h^*), \end{aligned} \tag{212}$$

where we used that $\hat{L}(1{:}\hat{n}, \mathcal{A}, \hat{h}) \leq \hat{L}(1{:}\hat{n}, \mathcal{A}^*, h^*)$. Next, by Lemma 5,

$$\begin{aligned} &L_{\mathcal{D}}(-i, \mathcal{A}, \hat{h}) - \hat{L}(1{:}\hat{n}, \mathcal{A}, \hat{h}) + \hat{L}(1{:}\hat{n}, \mathcal{A}^*, h^*) - L_{\mathcal{D}}(-i, \mathcal{A}^*, h^*) \\ &\quad \leq \widetilde{L}(-1{:}\hat{n}, \mathcal{A}, \hat{h}) - \hat{L}(1{:}\hat{n}, \mathcal{A}, \hat{h}) + \hat{L}(1{:}\hat{n}, \mathcal{A}^*, h^*) - \widetilde{L}(-1{:}\hat{n}, \mathcal{A}^*, h^*) + 2\sqrt{\frac{\log(2\delta)}{2n\hat{n}}}. \end{aligned} \tag{213}$$

This expression consists of the differences between the unobserved test losses and observed training losses for two different learning algorithms. We can thus use Corollary 6 to conclude that there exists a constant C such that, with probability at least $1-\delta$,

$$\begin{aligned} &\widetilde{L}(-1{:}\hat{n}, \mathcal{A}, \hat{h}) - \hat{L}(1{:}\hat{n}, \mathcal{A}, \hat{h}) + \hat{L}(1{:}\hat{n}, \mathcal{A}^*, h^*) - \widetilde{L}(-1{:}\hat{n}, \mathcal{A}^*, h^*) + 2\sqrt{\frac{\log(2\delta)}{2n\hat{n}}} \\ &\qquad \leq C\sqrt{\frac{d_N \log\left(\binom{N}{2}\frac{n\hat{n}}{d_N}\right) + \hat{n} d_{\text{VC}} \log\left(\frac{n}{d_{\text{VC}}}\right) + \log\left(\frac{\sqrt{n\hat{n}}}{\delta}\right)}{n\hat{n}}}. \end{aligned} \tag{214}$$

Here, the penalty term from the conversion between population and test loss has been absorbed into the constant C. From this, the desired result follows. □