CHALMERS
UNIVERSITY OF TECHNOLOGY

Conditional Mutual Information-Based Generalization Bound for Meta Learning

Citation for the original published paper (version of record):
Rezazadeh, A., Jose, S., Durisi, G. et al (2021). Conditional Mutual Information-Based Generalization Bound for Meta Learning. IEEE International Symposium on Information Theory - Proceedings, 2021-July: 1176-1181. http://dx.doi.org/10.1109/ISIT45174.2021.9518020

Conditional Mutual Information-Based Generalization Bound for Meta Learning

Arezou Rezazadeh[1], Sharu Theresa Jose[2], Giuseppe Durisi[1], Osvaldo Simeone[2]

[1] Chalmers University of Technology, [2] Kings College London
arezour@chalmers.se,sharu.jose@kcl.ac.uk, durisi@chalmers.se,osvaldo.simeone@kcl.ac.uk

***Abstract*—Meta-learning optimizes an inductive bias—typically in the form of the hyperparameters of a base-learning algorithm—by observing data from a finite number of related tasks. This paper presents an information-theoretic bound on the generalization performance of any given meta-learner, which builds on the conditional mutual information (CMI) framework of Steinke and Zakynthinou (2020). In the proposed extension to meta-learning, the CMI bound involves a training *meta-supersample* obtained by first sampling $2\mathrm{N}$ independent tasks from the task environment, and then drawing $2\mathrm{M}$ independent training samples for each sampled task. The meta-training data fed to the meta-learner is modelled as being obtained by randomly selecting N tasks from the available $2\mathrm{N}$ tasks and M training samples per task from the available $2\mathrm{M}$ training samples per task. The resulting bound is explicit in two CMI terms, which measure the information that the meta-learner output and the base-learner output provide about which training data are selected, given the entire meta-supersample. Finally, we present a numerical example that illustrates the merits of the proposed bound in comparison to prior information-theoretic bounds for meta-learning.**

I. Introduction

Meta-learning refers to the process of automatically optimizing the hyperparameters of a training algorithm by observing data from a number of related tasks, so as to "speed up" the learning of a new, previously unseen task [1], [2]. Hyperparameters include the initialization and the learning rate of a training algorithm [3], [4]. As in prior information-theoretic analyses of learning systems [5]–[7], we fix a training algorithm, also referred to as the *base-learner*, as a stochastic mapping $\mathrm{P}_{W|\boldsymbol{Z}^{\mathrm{M}},U=u}$ from the input training data $\boldsymbol{Z}^{\mathrm{M}}$ to the space of model parameters $\mathcal{W}$ for a given hyperparameter vector u. The meta-learner observes a meta-training data set $\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}$, comprising of M data samples, each from one of N related tasks, to optimize the hyperparameter u. The tasks are assumed to belong to a *task environment*, which is defined by a task distribution P_T over a set $\mathcal{T}$ of tasks and by per-task data distributions $\{\mathrm{P}_{Z|T=\tau}\}_{\tau\in\mathcal{T}}$.

The goal of the meta-learner is to minimize the *meta-generalization loss* $\mathrm{L}(u)$, i.e., the average loss incurred by the hyperparameter u when used by the base-learner on a new task T drawn from the same task environment. However, this quantity is not computable since the task-environment distribution is unknown. Instead, the meta-learner can evaluate the empirical *meta-training loss* $\mathrm{L}_{\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}}(u)$ for hyperparameter u based on the meta-training set $\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}$. The difference between the meta-generalization loss and the meta-training loss, $\Delta\mathrm{L}(u|\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}) = \mathrm{L}(u) - \mathrm{L}_{\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}}(u)$, is the *meta-generalization gap*. If the meta-generalization gap is small, the performance of the meta-learner on the meta-training set can be taken as a reliable measure of the meta-generalization loss. As in [5]–[7], this paper studies information-theoretic bounds on the average meta-generalization gap, $\Delta\mathrm{L}^{\text{avg}} = \mathbb{E}_{\mathrm{P}_{U\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}}}\left[\Delta\mathrm{L}(U|\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}})\right]$, where the average is taken over the meta-training set and the hyperparameters.

Related Work: Following the initial work of Russo and Zhou [6], information-theoretic bounds on the average generalization gap for conventional learning have been widely investigated in recent years [5], [7], [8]. While the bound in [5] depends on the mutual information (MI) between the output of the learning algorithm and the training set, Bu *et al.* [7] tightened the bound via the individual sample mutual information (ISMI) between the algorithm output and each individual sample of the training set. The MI and ISMI-based approaches have been extended to meta-learning in [9].

Directly relevant to this paper is the approach recently introduced in [10] for conventional learning, which will be referred to as *conditional mutual information* (CMI) framework. Differently from MI-based bounds, the CMI-based bound given in [10, Thm. 2] is always bounded. This is because the bound depends on the MI between the trained model and discrete data selection indices, rather than between trained model and training data as in [10]. To this end, the CMI framework introduces a *supersample* of $2\mathrm{M}$ samples generated independently according to the underlying data distribution, and the training samples are chosen by selecting M samples from the supersample at random. The resulting bound [10, Thm. 2] depends on the CMI between the algorithm output and the random vector that determines which training data are selected from the supersample, given the supersample.

Contributions: Inspired by [10], we present a novel information-theoretic bound on the average meta-

This work has been funded by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 893082 and by the Wallenberg AI, autonomous systems, and software program. It was also supported by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 725731).

generalization gap that extends the CMI-based bounds for conventional learning to meta-learning. In line with [9], we show that there are two sources of generalization errors that contribute to the meta-generalization gap: (*i*) the *environment-level* meta-generalization gap, resulting from the observation of a finite number N of tasks; and (*ii*) the *task-level* generalization gap, resulting from the observation of a finite number M of data samples per task. Unlike prior work [9], these two contributions are quantified in our analysis via CMI terms that depend on a *meta-supersample* of per-task training data sets and on data selection indices for meta-learner and base-learners. The derived bound inherits the advantage of the CMI bound for conventional learning, including its boundedness. We finally demonstrate the usefulness of the proposed bound on an example.

We conclude this section by further clarifying the relation to previous works [3], [11]–[21]. The information-theoretic bounds on the average meta-generalization gap derived here and in [9] hold for arbitrary, fixed base-learners and meta-learners. In contrast, the probably-approximately-correct (PAC) bound of [11] holds uniformly over a class of base-learners and meta-learners [22]. Our bound is related to, although not directly comparable with, the high-probability PAC-Bayes bounds reported in [12]–[14], which also apply to arbitrary, fixed, base-learners and meta-learners.

References [3], [15]–[21] derive bounds on the optimality gap, $\mathbb{E}_{\mathrm{P}_{U|\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}}[\mathrm{L}(U)] - \min_{u\in\mathcal{U}} \mathrm{L}(u)$, on average or with high probability, for specific meta-learning instances and algorithms, such as ridge regression with meta-learned bias in [15]–[18], gradient based meta-learning methods in [3], [19], and online meta learning in [20], [21]. In contrast, our derived bounds provide general information-theoretic insight into the number of tasks and number of samples per task required to ensure that the average meta-generalization gap for arbitrary base-learners and meta-learners is sufficiently small.

II. Problem Definition

In this section, we provide the key definitions for our setup.

Base-learner: Each task τ within some set of tasks $\mathcal{T}$ is associated with an underlying unknown data distribution $\mathrm{P}_{Z|T=\tau}$ on a set $\mathcal{Z}$. Throughout, sets can be discrete or continuous. For a given task τ_i, the *base-learner* observes a data set $\boldsymbol{Z}_i^{\mathrm{M}} = (Z_i^1, \ldots, Z_i^{\mathrm{M}})$ of M independently and identically distributed (i.i.d.) samples from $\mathrm{P}_{Z|T=\tau_i}$. The base-learner uses the training set $\boldsymbol{Z}_i^{\mathrm{M}}$ to infer a model parameter $w \in \mathcal{W}$. We assume that the base-learner depends on an inductive bias that is defined by a hyperparameter vector $u \in \mathcal{U}$. The performance of the model parameter w on a data sample z is measured by the loss function $\ell : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}^+$. The goal of the base-learner is to infer the model parameter $w \in \mathcal{W}$ that minimizes the per-task generalization loss

$$\mathrm{L}_{\mathrm{P}_{Z|T=\tau_i}}(w) = \mathbb{E}_{\mathrm{P}_{Z|T=\tau_i}}[\ell(w, Z)], \tag{1}$$

where the average is taken over a test sample $Z \sim \mathrm{P}_{Z|T=\tau_i}$ drawn independently from $\boldsymbol{Z}_i^{\mathrm{M}}$. Since $\mathrm{P}_{Z|T}$ is unknown, the generalization loss $\mathrm{L}_{\mathrm{P}_{Z|T=\tau_i}}(w)$ cannot be computed. Instead, the base-learner evaluates the training loss

$$\mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(w) = \frac{1}{\mathrm{M}} \sum_{j=1}^{\mathrm{M}} \ell(w, Z_i^j). \tag{2}$$

The difference between the generalization loss and the training loss is referred to as the *generalization gap*:

$$\Delta\mathrm{L}(w|\boldsymbol{Z}_i^{\mathrm{M}}, u, \tau_i) = \mathrm{L}_{\mathrm{P}_{Z|T=\tau_i}}(w) - \mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(w). \tag{3}$$

In this paper, we model the base-learner as a stochastic map $\mathrm{P}_{W|\boldsymbol{Z}_i^{\mathrm{M}},U=u}$ from the input training data set $\boldsymbol{Z}_i^{\mathrm{M}}$ to the model class $\mathcal{W}$. As mentioned, this map depends on u.

Meta-Learner: The goal of meta-learning is to automatically infer the hyperparameter u of the base-learner $\mathrm{P}_{W|\boldsymbol{Z}^{\mathrm{M}},U=u}$ from training data pertaining a number of related tasks. The tasks are assumed to belong to a *task environment*, which is defined by a task distribution P_T on the space of tasks $\mathcal{T}$, and by the per-task data distributions $\{\mathrm{P}_{Z|T=\tau}\}_{\tau\in\mathcal{T}}$. The meta-learner observes a meta-training set $\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}} = (\boldsymbol{Z}_1^{\mathrm{M}}, \ldots, \boldsymbol{Z}_{\mathrm{N}}^{\mathrm{M}})$ of N data sets. Each $\boldsymbol{Z}_i^{\mathrm{M}}$ is generated independently by first drawing a task $T_i \sim \mathrm{P}_T$ and then a task-specific dataset $\boldsymbol{Z}_i^{\mathrm{M}} \sim \mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}|T_i}$.

The meta-learner uses the meta-training set $\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}$ to infer the hyperparameter u. This is done with the goal of ensuring that, using the inferred hyperparameter u, the base-learner $\mathrm{P}_{W|\boldsymbol{Z}^{\mathrm{M}},U=u}$ can efficiently learn on a new meta-test task $T \sim \mathrm{P}_T$ given the corresponding training dataset $\boldsymbol{Z}^{\mathrm{M}}$.

Formally, the objective of the meta-learner is to infer the hyperparameter u that minimizes the *meta-generalization loss*

$$\mathrm{L}(u) = \mathbb{E}_{\mathrm{P}_T \mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}|T}}\left[\mathbb{E}_{\mathrm{P}_{W|\boldsymbol{Z}^{\mathrm{M}},U=u}}[\mathrm{L}_{\mathrm{P}_{Z|T}}(W)]\right], \tag{4}$$

where the expectation is taken over an independently generated meta-test task $T \sim \mathrm{P}_T$, over the associated data set $\boldsymbol{Z}^{\mathrm{M}} \sim \mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}|T}$, and over the output of the base-learner. Since P_T and $\{\mathrm{P}_{Z|T=\tau}\}_{\tau\in\mathcal{T}}$ are unknown, the meta-generalization loss (4) cannot be computed. Instead, the meta-learner can evaluate the *meta-training loss*, which for a given hyperparameter u, is defined as the average training loss on the meta-training set

$$\mathrm{L}_{\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}}(u) = \frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \mathbb{E}_{\mathrm{P}_{W|\boldsymbol{Z}_i^{\mathrm{M}},U=u}}[\mathrm{L}_{\boldsymbol{Z}_i^{\mathrm{M}}}(W)]. \tag{5}$$

Here, the average is taken over the output of the base-learner. The difference between the meta-generalization loss and the meta-training loss is the *meta-generalization gap*

$$\Delta\mathrm{L}(u|\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}) = \mathrm{L}(u) - \mathrm{L}_{\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}}(u). \tag{6}$$

Intuitively, if the meta-generalization gap is small, on average or with high probability, then the performance of the inferred hyperparameter u on the meta-training set can be taken as a reliable measure of the meta-generalization loss (4).

In this paper, we consider a stochastic meta-learner described by the conditional probability distribution $\mathrm{P}_{U|\boldsymbol{Z}_{1:\mathrm{N}}^{\mathrm{M}}}$,

which maps the meta-training set $\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}$ to the space of hyperparameters $\mathcal{U}$. We seek an information-theoretic upper bound on the average meta-generalization gap

$$\Delta\mathrm{L}^{\mathrm{avg}} = \mathbb{E}_{\mathrm{P}_{U\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}}\left[\Delta\mathrm{L}(U|\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}})\right], \tag{7}$$

where $\mathrm{P}_{U\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}} = \mathrm{P}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}\mathrm{P}_{U|\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}$ is the joint distribution of meta-training data and U induced by the meta-learner.

III. A CMI-Based Framework for Meta-Learning

In this section, we introduce the CMI framework for meta-learning. We start by noting that the average meta-generalization gap (7) can be decomposed as [9]

$$\begin{aligned}\Delta\mathrm{L}^{\mathrm{avg}} = &\mathbb{E}_{\mathrm{P}_{U\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}}\left[\mathrm{L}(U) - \mathbb{E}_{\mathrm{P}_{T\boldsymbol{Z}^{\mathrm{M}}}}[\widehat{\mathrm{L}}_{\boldsymbol{Z}^{\mathrm{M}}}(U)]\right]\\ &+\mathbb{E}_{\mathrm{P}_{U\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}}\left[\mathbb{E}_{\mathrm{P}_{T\boldsymbol{Z}^{\mathrm{M}}}}[\widehat{\mathrm{L}}_{\boldsymbol{Z}^{\mathrm{M}}}(U)] - \mathrm{L}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}(U)\right],\end{aligned} \tag{8}$$

where $\widehat{\mathrm{L}}_{\boldsymbol{Z}^{\mathrm{M}}}(u)$ denotes the average per-task training loss:

$$\widehat{\mathrm{L}}_{\boldsymbol{Z}^{\mathrm{M}}}(u) = \mathbb{E}_{\mathrm{P}_{W|\boldsymbol{Z}^{\mathrm{M}},U=u}}\left[\mathrm{L}_{\boldsymbol{Z}^{\mathrm{M}}}(W)\right]. \tag{9}$$

This quantity differs from the meta-training loss in that the hyperparameter U does not depend on the per-task training sets $\boldsymbol{Z}^{\mathrm{M}}$ used to evaluate the loss. On the contrary, the meta-training loss is evaluated on the meta-training data used to determine U. It also differs from the meta-generalization loss, since it averages training losses and not test losses. The first expectation in (8) captures the average *within-task generalization gap* associated to meta-test task $T \sim \mathrm{P}_T$, caused by the fact that only M training samples per the task are available. The second expectation captures the average *environment-level generalization gap*, caused by the fact that the meta-learner observes only N tasks.

To obtain an upper bound on the average meta-generalization gap, we bound these two expectations separately using the CMI approach introduced in the next section.

A. Per-Task Supersample and Meta-Supersamples

In this section, we define the per-task supersample following [10] and we introduce the concept of meta-supersample.

For a given task τ_i, we define the per-task supersample as the collection $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_i = (Z^1_i, \ldots, Z^{2\mathrm{M}}_i)$ of 2M samples draw independently from $\mathrm{P}_{Z|T=\tau_i}$. Let $\boldsymbol{S}_i = (S^1_i, \ldots, S^{\mathrm{M}}_i)$ be an M-dimensional random vector whose elements are drawn independently from a Bernoulli distribution with parameter 0.5, independent of $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_i$. We use the vector $\boldsymbol{S}_i$ to partition the per-task supersample $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_i$ into a set of M input training samples fed to the base-learner, and a set of M test samples. The vector of M input training samples, which we denote by $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_i(\boldsymbol{S}_i)$ is obtained as $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_i(\boldsymbol{S}_i) = (Z_i^{1+\mathrm{M}S^1_i}, Z_i^{2+\mathrm{M}S^2_i}, \ldots, Z_i^{\mathrm{M}+\mathrm{M}S^{\mathrm{M}}_i})$. Let $\bar{\boldsymbol{S}}_i$ denotes the vector whose entries are the modulo-2 complement of the entries of $\boldsymbol{S}_i$. Then $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_i(\bar{\boldsymbol{S}}_i)$ stands for the vector of test data for task τ_i. We shall use the training sets $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_i(\boldsymbol{S}_i)$ to train the base-learner, while the within-task generalization loss will be evaluated on the test data $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_i(\bar{\boldsymbol{S}}_i)$.

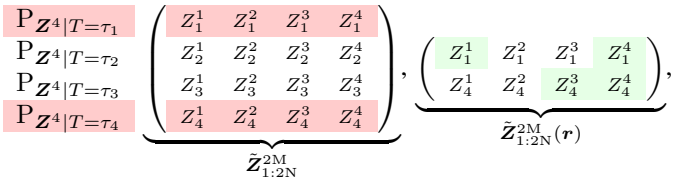

Figure 1. An example of meta-supersample $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}$ and $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}(\boldsymbol{R})$ for the case $\mathrm{M} = 2$ and $\mathrm{N} = 2$, $\boldsymbol{r} = (0, 1)$, $\boldsymbol{s}_1 = (0, 1)$ and $\boldsymbol{s}_2 = (1, 1)$. The rows of the meta-supersample matrix $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}$ contain data samples from four tasks. The shaded rows—the first and the fourth—are selected by $\boldsymbol{r} = (0, 1)$ to form the meta-training tasks. Then, the vectors $\boldsymbol{s}_1 = (0, 1)$ and $\boldsymbol{s}_2 = (1, 1)$ select the highlighted elements in $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}(\boldsymbol{r})$ to obtain the meta-training set $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}(\boldsymbol{r}, \boldsymbol{s}_{1:\mathrm{N}}) = \{\{Z^1_1, Z^4_1\}, \{Z^3_4, Z^4_4\}\}$.

We now describe the construction of the *meta-supersample*. We start by sampling 2N tasks, $T_1, \ldots, T_{2\mathrm{N}}$, independently from P_T. For each T_i, we generate a per-task supersample $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_i$ as detailed above. The meta-supersample is then defined as $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}} = (\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_1, \ldots, \tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{2\mathrm{N}})$. Let now $\boldsymbol{R} = (R_1, \ldots, R_{\mathrm{N}})$ be a N-dimensional random vector whose elements are drawn independently according to a Bernoulli distribution with parameter 0.5, independent of $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}$. We use the random vector $\boldsymbol{R}$ to partition the meta-supersample into N meta-training task datasets, and N meta-test datasets. Specifically, the meta-training task datasets $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}(\boldsymbol{R}) = (\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_1(R_1), \ldots, \tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{\mathrm{N}}(R_{\mathrm{N}}))$ are obtained by setting $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_i(R_i) = \tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{i+R_i\mathrm{N}}$, $i = 1, \ldots, \mathrm{N}$. Finally, the meta-training data fed to the meta-learner is $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}(\boldsymbol{R}, \boldsymbol{S}_{1:\mathrm{N}}) = (\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_1(R_1, \boldsymbol{S}_1), \ldots, \tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{\mathrm{N}}(R_{\mathrm{N}}, \boldsymbol{S}_{\mathrm{N}}))$, where $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_i(R_i, \boldsymbol{S}_i) = \tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{i+R_i\mathrm{N}}(\boldsymbol{S}_i)$. As before, we let $\bar{\boldsymbol{R}}$ be the vector whose entries are the modulo-2 complement of the entries of $\boldsymbol{R}$, and use $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}(\bar{\boldsymbol{R}})$ to denote all elements of $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}$ that are not in $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}(\boldsymbol{R})$.

To sum up, the meta-training set $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}(\boldsymbol{R}, \boldsymbol{S}_{1:\mathrm{N}})$ is generated by first choosing N tasks from the meta-supersample $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}$ according to $\boldsymbol{R}$, and then choosing M samples per task from $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_i$ according to $\boldsymbol{S}_i$, for $i = 1, \ldots, \mathrm{N}$. Fig. 1 shows an example of meta supersample.

IV. CMI-Based Bounds on $\Delta\mathrm{L}^{\mathrm{avg}}$

In this section, we introduce a CMI-based bound on the average meta-generalization gap $\Delta\mathrm{L}^{\mathrm{avg}}$ in (7). Throughout, we assume that the loss function $\ell(\cdot, \cdot)$ is bounded on the interval $[a, b]$. As discussed in Section III, to obtain an upper bound on the average meta-generalization gap, we upper-bound separately the within-task and the environment-level generalization gaps. Towards this goal, we leverage the exponential-inequality-based approach introduced in [23].

A. Main Result

Theorem 1: Under the assumption that the loss function is bounded as $0 \leq a \leq \ell(\cdot, \cdot) \leq b < \infty$, the following bound on the average meta-generalization gap (7) holds:

$$\Delta\mathrm{L}^{\mathrm{avg}} \leq \sqrt{2(b-a)^2\left(\frac{\mathrm{I}(U; \boldsymbol{R}, \boldsymbol{S}_{1:\mathrm{N}}|\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}})}{\mathrm{N}}\right)}$$

$$+\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\sqrt{2(b-a)^2\left(\frac{\mathrm{I}(W;\boldsymbol{S}_i|\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}},R_i)}{\mathrm{M}}\right)}. \quad (10)$$

The theorem is proved using the exponential inequalities that we report in the next subsection and Appendix B. The first term in (10) accounts for the environment-level generalization gap via the CMI $\mathrm{I}(U;\boldsymbol{R},\boldsymbol{S}_{1:\mathrm{N}}|\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}})$ divided by the number N of meta-training input tasks. The CMI $\mathrm{I}(U;\boldsymbol{R},\boldsymbol{S}_{1:\mathrm{N}}|\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}})$ measures the information the meta-learner output reveals about the environment-level partition $\boldsymbol{R}$ and per-task partition $\boldsymbol{S}_{1:\mathrm{N}}$ of the meta-supersample $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}$, when the supersample $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}$ is given. The second term in (10) accounts for the within-task generalization gap through the CMI $\mathrm{I}(W;\boldsymbol{S}_i|\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}},R_i)$, or equivalently $\mathrm{I}(W;\boldsymbol{S}_i|\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{i+R_i\mathrm{N}})$. Consistent with the per-task generalization gap bounds of [10], this second term measures the information the base-learner output reveals about the partitioning $\boldsymbol{S}_i$ of the per-task supersample $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{i+R_i\mathrm{N}}$.

The CMI-based bound (10) recovers the CMI bound for conventional learning in [10, Thm. 2] as special case. For conventional learning, the hyperparameter u is fixed *a priori*. Moreover, the task environment distribution can be assumed to be a delta function centered at some task $\tau\in\mathcal{T}$, and the meta-training set with $\mathrm{N}=1$ reduces to the training set of task τ. Hence, the first MI in (10) is zero, and (10) reduces to

$$\mathbb{E}_{\mathrm{P}_{W\boldsymbol{Z}^{\mathrm{M}}}}\left[\Delta\mathrm{L}(W|\boldsymbol{Z}^{\mathrm{M}})\right]\leq\sqrt{\frac{2(b-a)^2\mathrm{I}(W;\boldsymbol{S}|\tilde{\boldsymbol{Z}}^{2\mathrm{M}})}{\mathrm{M}}}, \quad (11)$$

which recovers the bound in [10, Thm. 2]. It can be verified that the CMI-based bound (10) is bounded by $\sqrt{2(b-a)^2\log 2}(\sqrt{1+\mathrm{M}}+1)$. A comparison with other information-theoretic bounds is presented in Section V.

B. Exponential Inequalities

The proof of Theorem 1 relies on exponential inequalities that are used to bound the two terms in the decomposition (8).

We start by expressing the within-task generalization gap, *i.e.*, the first expectation in the decomposition (8), in the following equivalent form, which makes explicit its dependence on the per-task supersamples and meta-supersample,

$$\begin{aligned}
&\mathbb{E}_{\mathrm{P}_{U\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}}\left[\mathrm{L}(U)-\mathbb{E}_{\mathrm{P}_{T\boldsymbol{Z}^{\mathrm{M}}}}[\widehat{\mathrm{L}}_{\boldsymbol{Z}^{\mathrm{M}}}(U)]\right]\\
&=\mathbb{E}_{\mathrm{P}_{U\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}}\mathbb{E}_{\mathrm{P}_{T\boldsymbol{Z}^{\mathrm{M}}}}\mathbb{E}_{\mathrm{P}_{W|\boldsymbol{Z}^{\mathrm{M}}U}}\left[\mathrm{L}_{\mathrm{P}_{Z|T}}(W)-\mathrm{L}_{\boldsymbol{Z}^{\mathrm{M}}}(W)\right] \quad (12)\\
&=\mathbb{E}_{\mathrm{P}_{U\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}\boldsymbol{R}\boldsymbol{S}_{1:\mathrm{N}}}}\left[\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\mathbb{E}_{\mathrm{P}_{W|\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_i(\bar{R}_i,\boldsymbol{S}_i),U}}\left[\mathrm{L}_{\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_i(\bar{R}_i,\bar{\boldsymbol{S}}_i)}(W)\right.\right.\\
&\qquad\left.\left.-\mathrm{L}_{\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_i(\bar{R}_i,\boldsymbol{S}_i)}(W)\right]\right] \quad (13)\\
&=\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\mathbb{E}_{\mathrm{P}_{\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}R_i\boldsymbol{S}_i}\mathrm{P}_{W|\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_i(\bar{R}_i,\boldsymbol{S}_i)}}\left[\mathrm{L}_{\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_i(\bar{R}_i,\bar{\boldsymbol{S}}_i)}(W)\right.\\
&\qquad\left.-\mathrm{L}_{\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_i(\bar{R}_i,\boldsymbol{S}_i)}(W)\right]. \quad (14)
\end{aligned}$$

The first equality (13) holds since the meta-learner is trained on the meta-supersample $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}(\boldsymbol{R},\boldsymbol{S}_{1:\mathrm{N}})$, and since the average over $\mathrm{P}_{T\boldsymbol{Z}^{\mathrm{M}}}$ in (12) can be evaluated on the independent test data set $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}(\bar{\boldsymbol{R}},\boldsymbol{S}_{1:\mathrm{N}})$. Note that the sets $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}(\boldsymbol{R},\boldsymbol{S}_{1:\mathrm{N}})$ and $\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}(\bar{\boldsymbol{R}},\boldsymbol{S}_{1:\mathrm{N}})$ represent independent datasets from two different environment-level tasks, even if they share the same $\boldsymbol{S}_{1:\mathrm{N}}$. Finally, we obtain (14) by taking the expectation inside the sum and by marginalizing over U.

To keep notation compact, let the term within the average in (14) be defined as

$$\widehat{\Delta\mathrm{L}}(W,\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}},R_i,\boldsymbol{S}_i)=\mathrm{L}_{\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_i(R_i,\bar{\boldsymbol{S}}_i)}(W)-\mathrm{L}_{\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_i(R_i,\boldsymbol{S}_i)}(W). \quad (15)$$

We now present a task-level exponential inequality that will be useful to bound the expectation of this term. For generality, the bound is expressed in terms of the Radon–Nikodym derivatives of the relevant distributions (see, e.g., [24, Sec. 17.1]).

Proposition 1: For all $\lambda\in\mathbb{R}$, we have

$$\mathbb{E}_{\mathrm{P}_{W\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}\boldsymbol{S}_iR_i}}\left[\exp\Big(\lambda\widehat{\Delta\mathrm{L}}(W,\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}},R_i,\boldsymbol{S}_i)-\frac{\lambda^2(b-a)^2}{2\mathrm{M}}-\imath(W;\boldsymbol{S}_i|\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}},R_i)\Big)\right]\leq 1,$$

where $\imath(W;\boldsymbol{S}|\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}},R)$ is the *conditional information density*

$$\imath(W;\boldsymbol{S}|\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}},R)=\log\frac{\mathrm{P}_{W\boldsymbol{S}|\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}R}}{\mathrm{P}_{W|\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}R}\mathrm{P}_{\boldsymbol{S}}}. \quad (16)$$

Proof: See Appendix A. ∎

We now derive a similar exponential inequality to bound the average environment-level generalization gap, i.e., the second expectation on the right-hand side of (8). Let

$$\begin{aligned}
&\widetilde{\Delta\mathrm{L}}(U,\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}},\boldsymbol{R},\boldsymbol{S}_{1:\mathrm{N}})\\
&\qquad=\mathrm{L}_{\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}(\bar{\boldsymbol{R}},\bar{\boldsymbol{S}}_{1:\mathrm{N}})}(U)-\mathrm{L}_{\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}(\boldsymbol{R},\boldsymbol{S}_{1:\mathrm{N}})}(U), \quad (17)
\end{aligned}$$

where $\mathrm{L}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}(u)$ was defined in (5). Using (17), and following steps similar to the ones leading to (14), we can express the environment-level generalization gap as

$$\begin{aligned}
&\mathbb{E}_{\mathrm{P}_{U\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}}\left[\mathbb{E}_{\mathrm{P}_{T,\boldsymbol{Z}^{\mathrm{M}}}}\left[\widehat{\mathrm{L}}_{\boldsymbol{Z}^{\mathrm{M}}}(U)\right]-\frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\widehat{\mathrm{L}}_{\boldsymbol{Z}^{\mathrm{M}}_i}(U)\right]=\\
&\mathbb{E}_{\mathrm{P}_{\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}\boldsymbol{R}\boldsymbol{S}_{1:\mathrm{N}}}\mathrm{P}_{U|\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}\boldsymbol{R}\boldsymbol{S}_{1:\mathrm{N}}}}\left[\widetilde{\Delta\mathrm{L}}(U,\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}},\boldsymbol{R},\boldsymbol{S}_{1:\mathrm{N}})\right]. \quad (18)
\end{aligned}$$

The next proposition states the relevant inequality. The proof is similar to that of Proposition 1 and, hence, omitted.

Proposition 2: For all $\lambda\in\mathbb{R}$, the following exponential inequality holds

$$\mathbb{E}_{\mathrm{P}_{U\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}\boldsymbol{R}\boldsymbol{S}_{1:\mathrm{N}}}}\left[\exp\Big(\lambda\widetilde{\Delta\mathrm{L}}(U,\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}},\boldsymbol{R},\boldsymbol{S}_{1:\mathrm{N}})-\frac{\lambda^2(b-a)^2}{2\mathrm{N}}-\imath\big(U;\boldsymbol{R},\boldsymbol{S}_{1:\mathrm{N}}|\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}\big)\Big)\right]\leq 1, \quad (19)$$

where $\mathrm{P}_{U\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}\boldsymbol{R}\boldsymbol{S}_{1:\mathrm{N}}}=\mathrm{P}_{\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}\boldsymbol{R}\boldsymbol{S}_{1:\mathrm{N}}}\mathrm{P}_{U|\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}\boldsymbol{R}\boldsymbol{S}_{1:\mathrm{N}}}$ and $\imath(U;\boldsymbol{R},\boldsymbol{S}_{1:\mathrm{N}}|\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}})$ is the conditional information density,

$$\imath\big(U;\boldsymbol{R},\boldsymbol{S}_{1:\mathrm{N}}|\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}\big)=\log\frac{\mathrm{P}_{U\boldsymbol{R}\boldsymbol{S}_{1:\mathrm{N}}|\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}}}{\mathrm{P}_{U|\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}}\mathrm{P}_{\boldsymbol{R}\boldsymbol{S}_{1:\mathrm{N}}}}. \quad (20)$$

Using these inequalities, the proof of Theorem 1 can be completed as detailed in Appendix B.

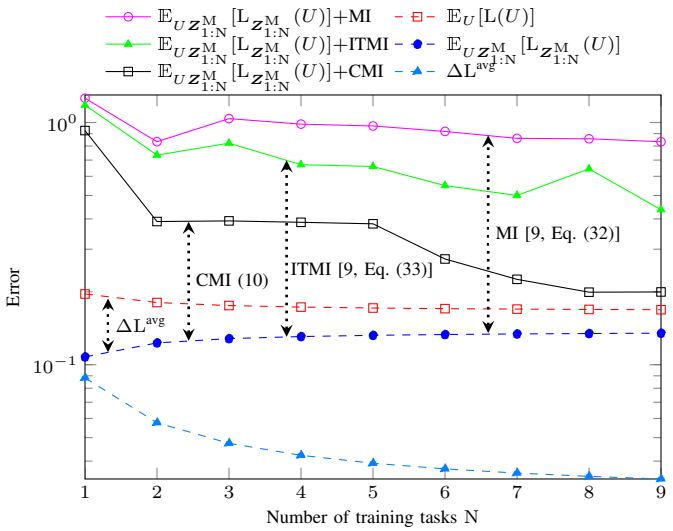

Figure 2. Comparison between the CMI-based bound (10) and both the MI-based and the ITMI-based bounds reported in [9, Eq. (32)] and [9, Eq. (33)], respectively. In the figure, $|\mathcal{T}| = 10$ with $\mathrm{P}_T = [0.05, 0.1, 0.02, 0.2, 0.01, 0.05, 0.02, 0.15, 0.1, 0.3]$, $\mathrm{M} = 5$, and $\alpha = 0.5$.

V. Example

We consider the example of mean estimation of a Bernoulli process studied in [9]. Specifically, we assume a finite set of tasks $\mathcal{T}$ and a task distribution P_T. For a given task $\tau_i \in \mathcal{T}$, we assume that the data follow a Bernoulli distribution with mean μ_{τ_i}. We also adopt the loss function $\ell(w, z) = (w - z)^2$. The base-learner's output W_i for task τ_i is chosen (deterministically) as a convex combination of the sample average $D_i = \frac{1}{\mathrm{M}}\sum_{j=1}^{\mathrm{M}} Z_i^j$, and of a hyperparameter vector u. Specifically, we set $W_i = \alpha D_i + (1-\alpha)u$ with some $\alpha \in [0, 1]$. The hyperparameter u is chosen (deterministically) so as to minimize the resulting meta-training empirical loss function $\mathrm{L}_{\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}(u) = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\frac{1}{\mathrm{M}}\sum_{j=1}^{\mathrm{M}}(W - Z_i^j)^2$, yielding $U = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} D_i$ [9].

In Fig. 2, we compare the CMI-based bound (10) with both the MI-based and the individual task mutual information (ITMI)-based bounds reported in [9, Eq. (32)] and in [9, Eq. (33)], respectively. We assume $|\mathcal{T}| = 10$, $\mathrm{P}_T = [0.05, 0.1, 0.02, 0.2, 0.01, 0.05, 0.02, 0.15, 0.1, 0.3]$, $\alpha = 0.5$, and $\mathrm{M} = 5$. The bounds are plotted as a function of the number N of available training tasks. As shown in the figure, the CMI-based bound provides a more accurate prediction of the average meta-generalization loss $\mathbb{E}[\mathrm{L}(U)]$, which can be computed in closed-form for this example [9], than the MI-based and the ITMI-based bounds.

VI. Conclusion

In this paper, we have derived a CMI-based bound on the average meta-generalization gap, which was demonstrated via an example to potentially result in a more accurate estimate of the meta-generalization performance as compared to previously derived ITMI- and MI-based bounds. The CMI-based bound scales as $O(1/\sqrt{\mathrm{N}}) + O(1/\sqrt{\mathrm{M}})$. A better scaling, for uniform-convergence bounds, has been recently reported in [25] under additional task-diversity assumptions. Adapting the CMI-based bound (10) to the PAC-Bayesian setting, along the lines of [26], may result in data-dependent high-probability bounds that are nonvacuous when applied to models such as deep neural networks. We leave this aspect to future work.

Appendix A
Proof of Proposition 1

Since $\ell(w, z)$ is bounded, $\widehat{\Delta\mathrm{L}}(w, \boldsymbol{y}^{2\mathrm{N}}, r_i, \boldsymbol{S}_i)$ is bounded on $[a - b, b - a]$, and hence, $\widehat{\Delta\mathrm{L}}(w, \boldsymbol{y}^{2\mathrm{N}}, r_i, \boldsymbol{S}_i)$ is subgaussian with parameter $(b-a)/\sqrt{\mathrm{M}}$ [27, Ex. 2.4]. Furthermore, since $\mathbb{E}_{\mathrm{P}_{\bar{\boldsymbol{S}}_i}}[\mathrm{L}_{\tilde{\boldsymbol{Z}}_i^{2\mathrm{M}}(r_i,\bar{\boldsymbol{S}}_i)}(W)] = \mathbb{E}_{\mathrm{P}_{\boldsymbol{S}_i}}[\mathrm{L}_{\tilde{\boldsymbol{Z}}_i^{2\mathrm{M}}(r_i,\boldsymbol{S}_i)}(W)]$, we have $\mathbb{E}_{\mathrm{P}_{\boldsymbol{S}_i}}[\widehat{\Delta\mathrm{L}}(w, \boldsymbol{y}^{2\mathrm{N}}, r_i, \boldsymbol{S}_i)] = 0$, Thus, for every w, $\boldsymbol{y}^{2\mathrm{N}}$, r_i,

$$\mathbb{E}_{\mathrm{P}_{\boldsymbol{S}_i}}\left[\exp\left(\lambda\widehat{\Delta\mathrm{L}}(w, \boldsymbol{y}^{2\mathrm{N}}, r_i, \boldsymbol{S}_i)\right)\right] \leq \exp\left(\frac{\lambda^2(b-a)^2}{2\mathrm{M}}\right). \tag{21}$$

Taking an additional expectation over $\mathrm{P}_{W\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}R_i}$, we find that

$$\mathbb{E}_{\mathrm{P}_{\boldsymbol{S}_i}\mathrm{P}_{W\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}R_i}}\left[\exp\left(\lambda\left(\widehat{\Delta\mathrm{L}}(W, \tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}, R_i, \boldsymbol{S}_i)\right) - \frac{\lambda^2(b-a)^2}{2\mathrm{M}}\right)\right] \leq 1. \tag{22}$$

The desired result then follows from a change of measure from $\mathrm{P}_{\boldsymbol{S}_i}\mathrm{P}_{W\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}R_i}$ to $\mathrm{P}_{W\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}\boldsymbol{S}_iR_i}$ [24, Prop. 17.1(4)].

Appendix B
Proof of Theorem 1

Substituting (15) into (14) and then (15) and (18) into (8),

$$\begin{aligned}
&\mathbb{E}_{\mathrm{P}_{U\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}}}}\left[\Delta\mathrm{L}(U|\boldsymbol{Z}^{\mathrm{M}}_{1:\mathrm{N}})\right] \\
&= \mathbb{E}_{\mathrm{P}_{\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}\boldsymbol{R}\boldsymbol{S}_{1:\mathrm{N}}}\mathrm{P}_{U|\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}\boldsymbol{R}\boldsymbol{S}_{1:\mathrm{N}}}}\left[\widetilde{\Delta\mathrm{L}}(U, \tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}, \boldsymbol{R}, \boldsymbol{S}_{1:\mathrm{N}})\right] \\
&+ \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}}\mathbb{E}_{\mathrm{P}_{W\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}\boldsymbol{S}_iR_i}}\left[\widehat{\Delta\mathrm{L}}(W, \tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}, R_i, \boldsymbol{S}_i)\right].
\end{aligned} \tag{23}$$

To bound the two terms on the right-hand side of (23), we use the exponential inequalities in Proposition 1 and 2. Specifically, applying Jensen's inequality to (16) and (19), we find the following inequalities

$$\exp\left(\lambda\mathbb{E}_{\mathrm{P}_{W\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}\boldsymbol{S}_iR_i}}\left[\widehat{\Delta\mathrm{L}}(W, \tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}, R_i, \boldsymbol{S}_i)\right] - \frac{\lambda^2(b-a)^2}{2\mathrm{M}} - \mathrm{I}(W; \boldsymbol{S}_i|\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}, R_i)\right) \leq 1, \tag{24}$$

and

$$\exp\left(\lambda\mathbb{E}_{\mathrm{P}_{U\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}\boldsymbol{R}\boldsymbol{S}_{1:\mathrm{N}}}}\left[\widetilde{\Delta\mathrm{L}}(U, \tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}}, \boldsymbol{R}, \boldsymbol{S}_{1:\mathrm{N}})\right] - \frac{\lambda^2(b-a)^2}{2\mathrm{N}} - \mathrm{I}(U; \boldsymbol{R}, \boldsymbol{S}_{1:\mathrm{M}}|\tilde{\boldsymbol{Z}}^{2\mathrm{M}}_{1:2\mathrm{N}})\right) \leq 1, \tag{25}$$

where (24) and (25) are valid for every $\lambda \in \mathbb{R}$. Taking the log on both sides of the above inequalities, and optimizing over λ, we obtain upper bounds on the two expectations in (23) (similar to the one reported in [23, Cor. 5]), which, when substituted into (23), give the desired bound.

REFERENCES

[1] J. Schmidhuber, "Evolutionary Principles in Self-Referential Learning, or On Learning How to Learn: The Meta-meta-... Hook," Ph.D. dissertation, Technische Universität München, 1987.

[2] S. Thrun and L. Pratt, "Learning to Learn: Introduction and Overview," in *Learning to Learn*. Springer, 1998, pp. 3–17.

[3] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Machine Learning (ICML)*, Sydney, Australia, Aug. 2017.

[4] Z. Li, F. Zhou, F. Chen, and H. Li, "Meta-SGD: Learning to learn quickly for few-shot learning," arXiv:1707.09835, 2017. [Online]. Available: http://arxiv.org/abs/1707.09835

[5] 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.

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

[7] Y. Bu, S. Zou, and V. V. Veeravalli, "Tightening mutual information based bounds on generalization error," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Paris, France, July 2019.

[8] 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.

[9] S. T. Jose and O. Simeone, "Information-theoretic generalization bounds for meta-learning and applications," arXiv:2005.04372, 2020. [Online]. Available: http://arxiv.org/abs/2005.04372

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

[11] J. Baxter, "A model of inductive bias learning," *Journal of Artif. Intell. Research*, vol. 12, pp. 149–198, Mar. 2000.

[12] A. Pentina and C. Lampert, "A PAC-Bayesian bound for lifelong learning," in *Proc. Int. Conf. on Machine Learning (ICML)*, Beijing, China, June 2014.

[13] J. Rothfuss, V. Fortuin, and A. Krause, "PACOH: Bayes-optimal meta-learning with PAC-guarantees," arXiv:2002.05551, 2020. [Online]. Available: http://arxiv.org/abs/2002.05551

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

[15] G. Denevi, C. Ciliberto, D. Stamos, and M. Pontil, "Incremental learning-to-learn with statistical guarantees," *UAI*, pp. 457–466, 2018.

[16] G. Denevi, C. Ciliberto, R. Grazzi, and M. Pontil, "Learning-to-learn stochastic gradient descent with biased regularization," in *Proc. Int. Conf. on Machine Learning (ICML)*, Long Beach, CA, USA, June 2019.

[17] G. Denevi, M. Pontil, and C. Ciliberto, "The advantage of conditional meta-learning for biased regularization and fine-tuning," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, USA, Dec. 2020.

[18] G. Denevi, C. Ciliberto, D. Stamos, and M. Pontil, "Learning to learn around a common mean," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Montreal, Canada, Dec. 2018.

[19] M. Konobeev, I. Kuzborskij, and C. Szepesvári, "On Optimality of Meta-Learning in Fixed-Design Regression with Weighted Biased Regularization," arXiv:2011.00344, 2020. [Online]. Available: https://arxiv.org/abs/2011.00344

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

[21] M. Khodak, M.-F. F. Balcan, and A. S. Talwalkar, "Adaptive gradient-based meta-learning methods," in *Proc. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, Canada, Dec. 2019.

[22] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *Proc. Int. Conf. Learning Representations, (ICLR) 2017*, Toulon, France, Apr. 2017.

[23] F. Hellström and G. Durisi, "Generalization bounds via information density and conditional information density," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 3, pp. 824–839, Nov. 2020.

[24] Y. Polyanskiy and Y. Wu, *Lecture notes on Information Theory*, MIT (6.441), UIUC (ECE 563), Yale (STAT 664), 2019.

[25] 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)*, Long Beach, CA, USA, 2020.

[26] F. Hellström and G. Durisi, "Nonvacuous loss bounds with fast rates for neural networks via conditional information measures," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Melbourne, Australia, Jul. 2021.

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