Information-Theoretic Generalization Bounds for Meta-Learning and Applications

Sharu Theresa Jose and Osvaldo Simeone

Abstract

Meta-learning, or “learning to learn”, refers to techniques that infer an inductive bias from data corresponding to multiple related tasks with the goal of improving the sample efficiency for new, previously unobserved, tasks. A key performance measure for meta-learning is the meta-generalization gap, that is, the difference between the average loss measured on the meta-training data and on a new, randomly selected task. This paper presents novel information-theoretic upper bounds on the meta-generalization gap. Two broad classes of meta-learning algorithms are considered that uses either separate within-task training and test sets, like MAML, or joint within-task training and test sets, like Reptile. Extending the existing work for conventional learning, an upper bound on the meta-generalization gap is derived for the former class that depends on the mutual information (MI) between the output of the meta-learning algorithm and its input meta-training data. For the latter, the derived bound includes an additional MI between the output of the per-task learning procedure and corresponding data set to capture within-task uncertainty. Tighter bounds are then developed, under given technical conditions, for the two classes via novel Individual Task MI (ITMI) bounds. Applications of the derived bounds are finally discussed, including a broad class of noisy iterative algorithms for meta-learning.

I. Introduction

A. Motivation

As formalized by the “no free lunch theorem”, any effective learning procedure must be based on prior assumptions on the task of interest [1]. These include the selection of a model class

The authors are with King’s Communications, Learning, and Information Processing (KCLIP) lab at the Department of Engineering of Kings College London, UK (emails: sharu.jose@kcl.ac.uk, osvaldo.simeone@kcl.ac.uk). The authors have received funding from the European Research Council (ERC) under the European Unions Horizon 2020 Research and Innovation Programme (Grant Agreement No. 725731).

and of the hyperparameters of a training algorithm, such as weight initialization and learning rate. In conventional single-task learning, these assumptions, collectively known as *inductive bias*, may rely on domain knowledge or validation [1]–[3]. Fixing a suitable inductive bias can significantly reduce the sample complexity of the learning process, and is thus crucial to any learning procedure. The goal of *meta-learning* is to automatically infer the inductive bias, thereby *learning to learn* from past experiences via the observation of a number of related tasks, so as to speed up learning a new and unseen task [4]–[8].

Following the standard setting of Baxter [9], meta-learning assumes the learning tasks to originate from a *task environment*, which defines a probability distribution on the (possible infinite) set of learning tasks. The past experience is modelled as the observation of data from a number of *meta-training tasks* which are sampled independently from the task environment. A meta-learner uses the *meta-training data set* to infer a hyperparameter U defining the inductive bias. The general goal is to ensure that this hyperparameter can be used to learn a new task, drawn from the same task environment, from fewer data samples.

The quality of the inferred hyperparameter U is measured by the *meta-generalization loss*, $\mathrm{L}_{P_{K,Z^m}}(U)$, which is the expected loss over task distribution, P_K, and conditional per-task data distribution, $P_{Z^m|K}$, incurred in learning a new task from the task environment. The notation will be formally introduced in Section II-B. While the goal of meta-learning is to infer a hyperparameter U that minimizes the meta-generalization loss $\mathrm{L}_{P_{K,Z^m}}(U)$, this is not computable, since the underlying task and data distributions are unknown. Instead, the meta-learner can evaluate an empirical estimate of the loss, $\mathrm{L}_{Z^m_{1:N}}(U)$, using the meta-training set $Z^m_{1:N}$, which is referred to as *meta-training loss*. The meta-generalization loss can be then decomposed as the sum of two terms

$$\mathrm{L}_{P_{K,Z^m}}(U) = \mathrm{L}_{Z^m_{1:N}}(U) + \Delta\mathrm{L}(U), \tag{1}$$

where the second term, $\Delta\mathrm{L}(U)$, is known as the *meta-generalization gap*. Minimizing simultaneously both of these terms is in general impossible due to their competing nature, particularly when the number of meta-training tasks N available is small: A small meta-training loss $\mathrm{L}_{Z^m_{1:N}}(U)$, requires the meta-learner to fit the meta-training set $Z^m_{1:N}$, while the meta-generalization gap $\Delta\mathrm{L}(U)$ measures how well the meta-learner generalize to new, previously unseen, tasks. A hyperparameter that is too sensitive to the specific meta-training set tasks and data set $Z^m_{1:N}$

may just memorize the tasks, and not generalize to new tasks [10]. The goal then is to strike a desirable balance between the two terms in (1).

In this paper, we study information-theoretic upper bounds on the meta-generalization gap $\Delta\mathrm{L}(U)$. Having analytical upper bounds on $\Delta\mathrm{L}(U)$ is of both theoretical and practical interest. At a theoretical level, meta-generalization gap bounds yield insights into the number of meta-training tasks and on the amount of per-task data required to ensure a sufficiently low meta-generalization loss in the decomposition (1) [9], [11]. At a practical level, bounds that do not depend on the data distribution can be used as regularizing terms in (1) in order to reduce meta-overfitting [12], [10]. This yields generalized (hierarchical) Bayesian inference problems [13].

While there exists a rich literature devoted to obtaining bounds on the generalization gap for conventional single-task learning, the analysis of the meta-generalization gap is not as well understood. Most notably, Baxter [9] proved the first theoretical probably approximate correct (PAC) bound on meta-generalization gap in the framework of Vapnik-Chervonenkis (VC) dimensions; and Maurer [11] employed the concept of algorithmic stability [14], [15] to obtain meta-generalization gap bounds. A recent line of work extends PAC-Bayesian bounds to meta-learning, including the bounds introduced by Pentina and Lambert [12], the tighter bound of Amit and Meir [16], and most recently, by Rothfuss et al [17].

B. Main Contributions

In light of these developments, the main contribution of this paper is the introduction of novel information-theoretic upper bounds on the expected meta-generalization gap. To the best of our knowledge, this work is the first to derive meta-generalization gap bounds within an information-theoretic framework. We specifically extend the line of work initiated by Russo and Zou [18] and Xu and Raginsky [19] for conventional learning to meta-learning. Information-theoretic bounds concern the average of the meta-generalization gap, and they depend explicitly on the task and per-task data distributions, on the loss function, and on the meta-training algorithm. The high probability PAC-Bayesian bounds [12], [16] closely resemble information-theoretic bounds given their dependence on information-theoretic divergence measures, but they are agnostic to task and data distributions. In fact, a variational formulation of information-theoretic bounds can recover the general form of PAC-Bayesian bounds [19]. A technical advantage of the information-theoretic bounds is their ability to account for unbounded loss functions, which is not the case for traditional PAC-Bayes approaches.

The derivation of meta-generalization gap bounds differs from conventional learning owing to two levels of uncertainties – *environment-level* uncertainty and *within-task* uncertainty. While within-task uncertainty results from observing a finite number m of data samples per task as in conventional learning, environment-level uncertainty results from observing a finite number N of tasks from the task-environment. The relative importance of these two forms of uncertainty depend on the use made by the meta-learner of the meta-training data. In fact, there are two main classes of meta-training algorithms – with *separate within-task training and test sets*, and *joint within-task training and test sets*. The former class includes the state-of-the-art meta-learning algorithms such as Model Agnostic Meta-Learning (MAML) [20] that split the training data corresponding to each task into training and test sets, with the latter reserved for within-task validation. In contrast, the second class of algorithms, such as Reptile [21], use the entire per-task data both for training and testing. Our main contributions are as follows.

- For the case with separate within-task training and test sets, we show that the average meta-generalization gap contains only the contribution of environment-level uncertainty, which is captured by a ratio of the mutual information (MI) between the output of the meta-learner and the meta-training set and the number of tasks N – a direct parallel of the MI-based bounds for single-task learning [19].
- For the case with joint within-task training and test sets, we prove that the bound on the meta-generalization gap also contains a contribution due to the within-task uncertainty via the ratio of the MI between the output of the base learner and within task training data and the per-task data sample size m.
- We then extend the notion of individual sample MI (ISMI) of [22] to obtain novel Individual Task MI (ITMI)-based bounds on the meta-generalization gap for both separate and within-task training and test sets. Under given conditions, these bounds are tighter than MI-based bounds.
- Finally, we study the applications of the derived bounds to two meta-learning problems. The first is a parameter estimation setup that involves one-shot meta-learning and base-learning procedures, for which a closed form expression for meta-generalization gap can be derived. The second application covers a broad range of noisy iterative meta-learning algorithms and is inspired by the work of Pensia et al [23] for conventional learning.

C. Related Work

For conventional learning, there exists a rich literature on diverse frameworks for deriving upper bounds on the generalization gap, i.e. on the difference between generalization and training losses. Classical bounds from statistical learning theory quantify the generalization gap in terms of measures of complexity of the model class, most notably VC dimension [24] and Radmacher complexity [25]. This approach obtains high-probability probably approximate correct (PAC) bounds on the generalization gap. An alternate line of high-probability bounding techniques relies on the notion of *algorithmic stability*, which measures the sensitivity of the output of a learning algorithm to the replacement of individual samples from the training data set. The pioneering work [26] has been extended to include various notions of algorithmic stability [27]–[29]. As notable examples, a distributional notion of stability in terms of *differential privacy*, which quantifies the sensitvity of the distribution of algorithm's output to data set, has been studied in [30], [31], while PAC-Bayesian bounds rely on change of measure arguments [32]–[34].

Following the initial work of Russo and Zou [18], information-theoretic bounds on the average generalization gap for conventional learning have been widely investigated in recent years. Xu and Raginsky [19] showed that the MI between the output of the learning algorithm and its training data set yields an upper bound bound in expectation on the generalization gap. The bound has been shown to offer computable generalization gaurentees for noisy iterative algorithms including Stochastic Gradient Langevin Dynamics (SGLD) in [23]. Various refinements of the MI-based bound have since been analyzed to obtain tighter bounds. In particular, the bounds in [35] employ chaining mutual information techniques to tighten the bounds in [19], while the bound in [22] depend on the MI between the output of the algorithm and an individual data sample. The MI between the output of the algorithm and a random subset of the data set appears in the bounds introduced in [36]. The total variation information between the joint distribution of the training data and algorithmic output and the product of marginals was shown in [37] to yield a bound on the generalization gap for any bounded loss function. Subsequent works in [38]–[40] consider other information-theoretic measures, such as maximum leakage and lautum information. Most recently, a conditional mutual information (CMI)-based approach has been proposed in [41] as a unifying framework to develop generalization bounds.

D. Notation

Throughout this paper, upper case letters, e.g. X, denote random variables and lower case letters, e.g. x, their realizations. We use $\mathcal{P}(\cdot)$ to denote the set of all probability distributions on the argument set or vector space. For a discrete or continuous random variable X taking values in a set or vector space $\mathcal{X}$, $P_X \in \mathcal{P}(\mathcal{X})$ denotes its probability distribution, with $P_X(x)$ being the probability mass or density value at $x \in \mathcal{X}$. We denote as P_{X^n} the n-fold product distribution induced by P_X. The conditional distribution of a random variable X given random variable Y is similarly defined as $P_{X|Y}$, with $P_{X|Y}(x|y)$ representing the probability mass or density at $X = x$ conditioned on the event $Y = y$. We use $||\cdot||_2$ to denote the Euclidean norm of the argument vector, and I_d to denote a d-dimensional identity matrix.

II. Problem Definition

In this section, we define the problem of interest by introducing the key definitions of generalization gap for conventional, or single-task, learning and for meta-learning.

A. Generalization Gap for Single-Task Learning

Consider first the conventional problem of learning for a single task indexed by an integer k. As illustrated in Figure 1, each task k is associated with an underlying *unknown* data distribution, $P_{Z|k} \in \mathcal{P}(\mathcal{Z})$, defined in a subset or vector space $\mathcal{Z}$. As a preliminary step, an *inductive bias* is selected that consists of a model class $\mathcal{W}$, parameterized by a vector $w \in \mathcal{W}$, and of a training

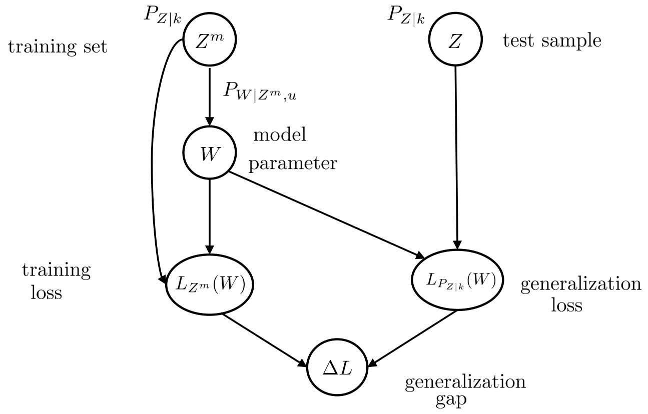

Fig. 1: Directed graph representing the variables involved in the definition of generalization gap (5) for single-task learning.

procedure. The training procedure, which is referred to as the *base learner*, has access to a training data set $Z^m = (Z_1, Z_2, \dots, Z_m) \sim P_{Z^m|k}$ of m independent and identically distributed (i.i.d.) samples drawn from distribution $P_{Z|k}$. The base learner uses this data set to choose a model, or hypothesis, W from the model class $\mathcal{W}$ by using a *randomized* training procedure defined by a conditional distribution $P_{W|Z^m,u}$ as

$$W \sim P_{W|Z^m,u}. \tag{2}$$

The conditional distribution $P_{W|Z^m,u}$ defines a stochastic mapping from the training data set Z^m to the model class $\mathcal{W}$. The training procedure (2) is parameterized by a vector $u \in \mathcal{U}$ of *hyperparameters*, which is considered to be part of the inductive bias along with the model class $\mathcal{W}$. As an example, the base learner $P_{W|Z^m,u}$ may follow Stochastic Gradient Descent (SGD) updates with hyperparameters u including the learning rate and the initialization point.

The performance of a parameter vector $w \in \mathcal{W}$ on a data sample $z \in \mathcal{Z}$ is measured by a loss function $l(w, z)$. The *generalization loss* for a model parameter vector $w \in \mathcal{W}$ is the average

$$L_{P_{Z|k}}(w) = \mathbb{E}_{P_{Z|k}}[l(w, Z)], \tag{3}$$

over a test example Z independently drawn from the data distribution $P_{Z|k}$. The generalization loss cannot be computed by the learner, given that the data distribution $P_{Z|k}$ is unknown. Instead, the learner can evaluate the *training loss* on the data set Z^m, which is defined as the empirical average

$$L_{Z^m}(w) = \frac{1}{m}\sum_{i=1}^{m} l(w, Z_i). \tag{4}$$

The difference between generalization loss (3) and training loss (4), known as *generalization gap*, is a key metric that quantifies the level of uncertainty[1] at the learner regarding the data distribution $P_{Z|k}$. The average generalization gap for the data distribution $P_{Z|k}$ and base learner $P_{W|Z^m,u}$ is defined as

$$\Delta L(P_{Z|k}, P_{W|Z^m,u}) = \mathbb{E}_{P_{Z^m,W|k,u}}\Big[L_{P_{Z|k}}(W) - L_{Z^m}(W)\Big], \tag{5}$$

where the expectation is taken with respect to the joint distribution $P_{Z^m,W|k,u} = P_{Z^m|k}P_{W|Z^m,u}$. A summary of the variables involved in the definition of the generalization gap (5) can be found in Figure 1.

[1] This type of uncertainty is known as epistemic.

Intuitively, if the generalization gap is small, on average or with high probability, then the base learner can take the performance (4) on the training set Z^m as a reliable measure of the generalization loss (3) of the trained model W. Furthermore, data-dependent bounds on the generalization gap can be used as regularization terms to avoid overfitting, yielding generalized Bayesian inference problems [13], [42].

B. Generalization Gap for Meta-Learning

As discussed, in single-task learning, the inductive bias $(\mathcal{W}, u)$, defining model class and hyperparameters of the training procedure, must be selected a priori, i.e., without having access to task-specific data. The inductive bias determines the training data set size m needed to ensure a small generalization loss (3), since, generally speaking, richer models require more data to be trained [1]. The sample complexity can be generally reduced if one selects a suitable inductive bias based on prior information. Such prior information is typically obtained from domain knowledge on the problem under study. In contrast, meta-learning aims at automatically inferring an effective inductive bias based on data from related tasks.

To elaborate, we follow the setting of [9], in which a meta-learner observes data from a number of tasks, known as *meta-training tasks*, from the same *task environment*. A task environment is defined by a task distribution $P_K \in \mathcal{P}(\mathcal{K})$, supported on a subset $\mathcal{K}$ of the integers, and by a per-task data distribution $P_{Z|k}$ for each task $k \in \mathcal{K}$. Using the meta-training data drawn from a randomly selected subset of tasks, the meta-learner infers a hyperparameter vector $u \in \mathcal{U}$ defining the inductive bias. This is done with the goal of ensuring that, using hyperparameter u, the base learner $P_{W|Z^m,u}$ can efficiently learn on a new task, referred to as *meta-test task*, drawn independently from the same task environment distribution P_K.

To elaborate, the meta-training data consists of N data sets $Z^m_{1:N} = (Z^m_1, \ldots, Z^m_N)$. Each ith data set is generated independently by first drawing a task $K_i \sim P_K$ from the task environment and then a task-specific training data set $Z^m_i \sim P_{Z^m|K_i}$. The meta-learner uses the meta-training data set $Z^m_{1:N}$ to infer a hyperparameter vector $u \in \mathcal{U}$. To this end, we consider a *randomized meta-learner*

$$U \sim P_{U|Z^m_{1:N}}, \tag{6}$$

where $P_{U|Z^m_{1:N}}$ is a stochastic mapping from the meta-training set $Z^m_{1:N}$ to the space $\mathcal{U}$ of hyperparameters. We distinguish two different formulations of meta-learning that are often considered in the literature. In the first, the per-task data set Z^m is split into training, or support, and test, or query subsets [10], [20]; while, in the second, the entire data set Z^m is used for both within-task training and testing [9], [12], [16].

C. Separate Within-Task Training and Test Sets

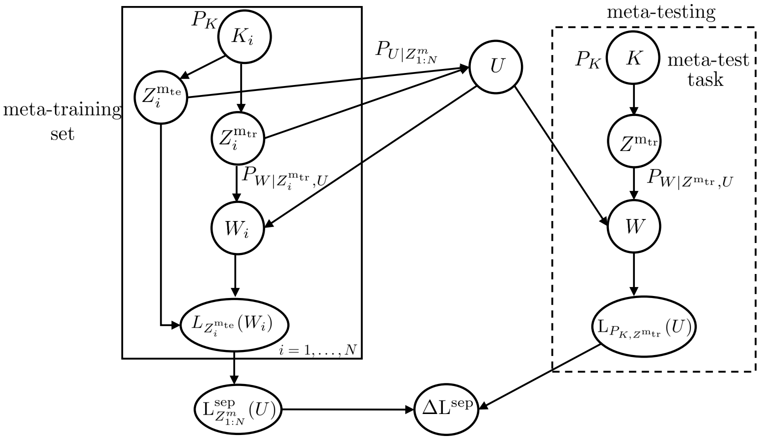
meta-training set
P_K
K_i
$Z_i^{\mathrm{m_{te}}}$
$Z_i^{\mathrm{m_{tr}}}$
$P_{W|Z_i^{\mathrm{m_{tr}}},U}$
W_i
$L_{Z_i^{\mathrm{m_{te}}}}(W_i)$
$i=1,\ldots,N$
$P_{U|Z_{1:N}^m}$
U
meta-testing
P_K
K
meta-test task
$Z^{\mathrm{m_{tr}}}$
$P_{W|Z^{\mathrm{m_{tr}}},U}$
W
$\mathrm{L}_{P_{K,Z^{\mathrm{m_{tr}}}}}(U)$
$\mathrm{L}^{\mathrm{sep}}_{Z_{1:N}^m}(U)$
$\Delta\mathrm{L}^{\mathrm{sep}}$

Fig. 2: Directed graph representing the variables involved in the definition of meta-generalization gap (10) for separate within-task training and testing sets.

As seen in Figure 2, in this first approach to meta-learning, each meta-training sub data set Z_i^m is split into a training set and a test set as $Z_i^m = (Z_i^{\mathrm{m_{tr}}}, Z_i^{\mathrm{m_{te}}})$, where $Z_i^{\mathrm{m_{tr}}}$ contains m_{tr} i.i.d. training examples and $Z_i^{\mathrm{m_{te}}}$ contains m_{te} i.i.d. test examples, with $m = m_{\mathrm{tr}} + m_{\mathrm{te}}$. The *within-task base learner* $P_{W|Z_i^{\mathrm{m_{tr}}},u} \in \mathcal{P}(\mathcal{W})$ maps the per-task training subset $Z_i^{\mathrm{m_{tr}}}$ to random model parameter $W_i \sim P_{W|Z_i^{\mathrm{m_{tr}}},u}$ for a given hyperparameter $U = u$. The test subset is used to evaluate the empirical loss of a model w for task K_i as

$$L_{Z_i^{\mathrm{m_{te}}}}(w) = \frac{1}{m_{\mathrm{te}}} \sum_{j=1}^{m_{\mathrm{te}}} l(w, Z_{i,j}^{m_{\mathrm{te}}}), \tag{7}$$

where $Z_{i,j}^{m_{\mathrm{te}}}$ denote the jth example of the test subset $Z_i^{m_{\mathrm{te}}}$. Furthermore, the overall empirical *meta-training loss* for a hyperparameter u is computed by summing over all meta-training tasks as

$$\mathrm{L}^{\mathrm{sep}}_{Z_{1:N}^m}(u) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{P_{W|Z_i^{\mathrm{m_{tr}}},u}} \Big[L_{Z_i^{\mathrm{m_{te}}}}(W) \Big]. \tag{8}$$

We emphasize that the meta-training loss (8) can be computed by the meta-learner and used as a criterion to select the meta-learning procedure (6) since it is obtained from the meta-training data $Z^m_{1:N}$. We also note that the rationale of splitting training and test sets is that the average empirical loss $\mathbb{E}_{P_{W|Z_i^{\mathrm{m_{tr}}},u}}[L_{Z_i^{\mathrm{m_{te}}}}(W)]$ is an unbiased estimate of the corresponding average test loss $\mathbb{E}_{P_{W|Z_i^{\mathrm{m_{tr}}},u}}[L_{P_{Z|K_i}}(W)]$.

The true goal of the meta-learner is to minimize the *meta-generalization loss*,

$$\mathrm{L}_{P_{K,Z^{\mathrm{m_{tr}}}}}(u) = \mathbb{E}_{P_{K,Z^{\mathrm{m_{tr}}}}}\mathbb{E}_{P_{W|Z^{\mathrm{m_{tr}}},u}}\big[L_{P_{Z|K}}(W)\big]. \tag{9}$$

Unlike the meta-training loss (8), the meta-generalization loss is evaluated on a new, meta-test task K and on the corresponding training data $Z^{\mathrm{m_{tr}}}$. The difference between the meta-generalization loss (9) and the meta-training loss (8) is known as the *meta-generalization gap* and is defined as

$$\Delta\mathrm{L}^{\mathrm{sep}}(P_{K,Z^{\mathrm{m_{tr}}}}, P_{U|Z^m_{1:N}}, P_{W|Z^{\mathrm{m_{tr}}},U}) = \mathbb{E}_{P_{Z^m_{1:N},U}}\bigg[\mathrm{L}_{P_{K,Z^{\mathrm{m_{tr}}}}}(U) - \mathrm{L}^{\mathrm{sep}}_{Z^m_{1:N}}(U)\bigg], \tag{10}$$

where the expectation is with respect to the joint distribution $P_{Z^m_{1:N},U} = P_{Z^m_{1:N}}P_{U|Z^m_{1:N}}$, of the meta-training set $Z^m_{1:N}$ and of the hyperparameter U.

Intuitively, if the meta-generalization gap is small, on average or with high probability, the meta learner can take the performance (8) on the meta-training data as a reliable measure of the accuracy of the inferred hyperparameter vector in terms of the meta-generalization loss (9). Furthermore, data-dependant bounds on the meta-generalization gap can be used as regularization terms to avoid meta-overfitting. Meta-overfitting occurs when the meta-trained hyperparameter yields a small meta-training loss but a large meta-test loss due to an excessive dependence on the meta-training set [9].

D. Joint Within-Task Training and Test Sets

In the second formulation of meta-learning, as illustrated in Figure 3, the entire data set Z^m_i is used for within-task training and testing. Accordingly, the meta-learner computes the *meta-training loss*

$$\mathrm{L}^{\text{joint}}_{Z^m_{1:N}}(u) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{P_{W|Z^m_i,u}}[L_{Z^m_i}(W)] \tag{11}$$

by using data set Z^m_i to infer model parameters W and to evaluate the per-task empirical loss. The expectation in (11) is taken over the output of the base learner W for each task K_i given the hyperparameter vector u. As discussed, the *meta-generalization loss* for hyperparameter $u \in \mathcal{U}$ is computed by randomly selecting a novel task $K \sim P_K$ as

$$\mathrm{L}_{P_{K,Z^m}}(u) = \mathbb{E}_{P_{K,Z^m}}\mathbb{E}_{P_{W|Z^m,u}}\big[L_{P_{Z|K}}(W)\big]. \tag{12}$$

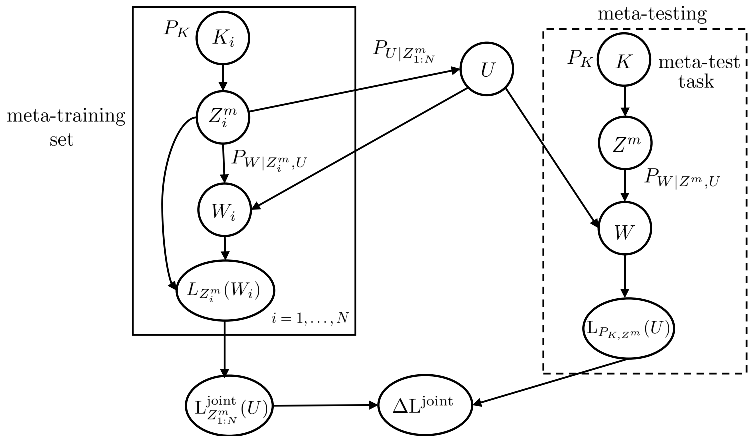

Fig. 3: Directed graph representing the variables involved in the definition of meta-generalization gap (13) for joint within-task training and testing sets.

In a manner similar to (10), the *meta-generalization gap* for a task distribution P_K, data distribution $P_{Z^m|K}$, meta-learning algorithm $P_{U|Z^m_{1:N}}$, and base learner $P_{W|Z^m,U}$ is defined as

$$\Delta\mathrm{L}^{\text{joint}}(P_{K,Z^m}, P_{U|Z^m_{1:N}}, P_{W|Z^m,U}) = \mathbb{E}_{P_{Z^m_{1:N},U}}\Big[\mathrm{L}_{P_{K,Z^m}}(U) - \mathrm{L}^{\text{joint}}_{Z^m_{1:N}}(U)\Big], \tag{13}$$

where the expectation is taken over all meta-training sets and over the output of the meta-learner.

III. Preliminaries

In this section, we cover some technical background and notations that will be useful in the following sections. Since the generalization and meta-generalization gaps measure the deviation of empirical-mean random variables representing training and meta-training losses from reference values, we will make use of tools and definitions from large-deviation theory (see, e.g, [43]). To start, the cumulant generating function (CGF) of a random variable $X \sim P_X \in \mathcal{P}(\mathcal{X})$ is defined as $\varLambda_X(\lambda) = \log \mathbb{E}_{P_X}[e^{\lambda(X - \mathbb{E}_{P_X}[X])}]$. If it is well-defined, the CGF $\varLambda_X(\lambda)$ is convex and it satisfies the equalities $\varLambda_X(0) = \varLambda'_X(0) = 0$. A random variable X with finite mean, i.e., with $\mathbb{E}_{P_X}[X] < \infty$, is said to σ^2-sub-Gaussian if its CGF is bounded as

$$\varLambda_X(\lambda) \leq \frac{\lambda^2\sigma^2}{2}, \quad \text{for all } \lambda \in \mathbb{R}. \tag{14}$$

As a special case, if X is $[a, b]$-bounded almost surely, i.e., if the inequality $-\infty < a \leq X \leq b < \infty$ holds for some constants a and b, then X is $(b - a)^2/4$-sub-Gaussian.

The definition of sub-Gaussianity can be extended by introducing the notion of a generalized-sub-Gaussian random variable X that relaxes the upper bound condition (14) on the CGF $\varLambda_X(\lambda)$ as follows [44], [22].

Definition 3.1: A random variable X is said to be $(\varPsi_+, \varPsi_-, b_+, b_-)$-*generalized sub-Gaussian* if there exists convex functions $\varPsi_+ : \mathbb{R}_+ \rightarrow \mathbb{R}$ and $\varPsi_- : \mathbb{R}_+ \rightarrow \mathbb{R}$ that satisfy the equalities $\varPsi_+(0) = \varPsi_-(0) = \varPsi'_+(0) = \varPsi'_-(0) = 0$ and bound the CGF of X as

$$\varLambda_X(\lambda) \leq \varPsi_+(\lambda), \qquad \text{for } \lambda \in [0, b_+) \tag{15a}$$

$$\varLambda_X(\lambda) \leq \varPsi_-(-\lambda), \quad \text{for } \lambda \in (b_-, 0], \tag{15b}$$

for some constants $0 < b_+ \leq \infty$ and $-\infty \leq b_- < 0$.

For a $(\varPsi_+, \varPsi_-, b_+, b_-)$-generalized sub-Gaussian random variable, we also introduce the following standard definitions. First, the Legendre dual of function $\varPsi_+(\lambda)$ is defined as

$$\varPsi^*_+(x) = \sup_{\lambda \in [0, b_+)} (\lambda x - \varPsi_+(\lambda)). \tag{16}$$

It can be easily seen that $\varPsi^*_+(\cdot)$ is a non-negative, convex, and non-decreasing function on $[0, \infty)$ with $\varPsi^*_+(0) = 0$. Second, the inverse Legendre dual of function $\varPsi_+(\lambda)$ is defined as $\varPsi^{*-1}_+(y) = \inf\{x \geq 0 : \varPsi^*_+(x) \geq y\}$. This function is concave, and it can be equivalently written as [22]

$$\varPsi^{*-1}_+(y) = \inf_{\lambda \in [0, b_+)} \frac{y + \varPsi_+(\lambda)}{\lambda}. \tag{17}$$

Similar definitions and results apply for $\varPsi_-(\cdot)$.

A σ^2-sub-Gaussian random variable X is a generalized sub-Gaussian variable with $\varPsi_+(\lambda) = \varPsi_-(\lambda) = \lambda^2\sigma^2/2$, $b_+ = \infty$ and $b_- = -\infty$. Furthermore, the Legendre dual functions are given as $\varPsi^*_+(x) = \varPsi^*_-(x) = x^2/(2\sigma^2)$, and the inverse Legendre dual functions evaluate to

$$\varPsi^{*-1}_+(y) = \varPsi^{*-1}_-(y) = \sqrt{2\sigma^2 y}. \tag{18}$$

IV. Information-Theoretic Generalization Bounds for Single-Task Learning

In this section, we review two information-theoretic bounds on the generalization gap (5) for conventional learning derived in [19] and [22]. The material covered in this section provides the necessary background for the analysis of the meta-generalization gap to be studied in the rest of the paper. Throughout this section, the task index k is fixed. Finally, as a point of notation, we will write inequalities in the form $\pm A \leq \varPsi^{*-1}_{\mp}(B)$ to indicate the conditions $-\varPsi^{*-1}_+(B) \leq A \leq \varPsi^{*-1}_-(B)$.

A. Mutual Information (MI) Bound

We first present the Mutual Information (MI)-based upper bound obtained in [19]. Key to this result is the following assumption.

Assumption 4.1: For all $w \in \mathcal{W}$, the loss function $l(w, Z)$ is a $(\varPsi_+, \varPsi_-, \infty, -\infty)$-generalized sub-Gaussian random variable under $Z \sim P_{Z|k}$.

The main result is as follows.

Lemma 4.1 ([44]): Under Assumption 4.1, for any base learner $W \sim P_{W|Z^m,u}$ with fixed hyperparameter vector $u \in \mathcal{U}$ such that the inequality $I(W; Z^m) < \infty$ holds, we have the following bounds on the generalization gap (5)

$$\pm \Delta L(P_{Z|k}, P_{W|Z^m,u}) \leq \varPsi_{\mp}^{*-1}\left(\frac{1}{m} I(W; Z^m)\right). \tag{19}$$

The proof of Lemma 4.1 is based on a decoupling estimate lemma, which is reported for completeness in Lemma A.1. The bound in Lemma 4.1 simplifies when specialized to the example of σ^2-sub-Gaussian loss functions $l(w, z)$.

Corollary 4.1 ([19]): If the loss function $l(w, Z)$ is a σ^2-sub-Gaussian random variable for all $w \in \mathcal{W}$ under $Z \sim P_{Z|k}$, then for any base learner $W \sim P_{W|Z^m,u}$, the following bound holds on the generalization gap

$$|\Delta L(P_{Z|k}, P_{W|Z^m,u})| \leq \sqrt{\frac{2\sigma^2}{m} I(W; Z^m)}. \tag{20}$$

The bounds (19) and (20) on the generalization gap are in terms of the mutual information $I(W; Z^m)$, which quantifies the overall dependence between the base learner output W and the input training data set Z^m. The mutual information in (20) is hence a measure of the *sensitivity* of the base learner output to the data set. Using the terminology in [19], if $I(W; Z^m) \leq \epsilon$, the base learner $P_{W|Z^m,u}$ is said to be $(\epsilon, P_{Z|k})$-MI stable, in which case the bound in (20) evaluates to $\sqrt{2\sigma^2\epsilon/m}$. The relationship between generalization and stability of a training algorithm is well-established [1], and the result (20), or more generally (19), amounts to a formulation of this link in information-theoretic terms.

The traditional notion of algorithmic stability measures how much the base learner output changes with the replacement of an individual training sample [26], [45]. In the next section, we review the bound in [22] that translates this per-sample stability concept within an information-theoretic framework.

B. Individual Sample MI (ISMI) Bound

The MI-based bound in Lemma 4.1 has the disadvantage of being vacuous, i.e., $I(W;Z^m)=\infty$, for deterministic algorithms and a continuous parameter space $\mathcal{W}$. An *individual sample MI* (ISMI)-based bound that address this shortcoming was introduced in [22]. The ISMI bound borrows the standard algorithmic stability notion of sensitivity of the base learner output to the replacement of any individual training sample [14], [15]. Accordingly, the resulting bound is in terms of the MI between the trained parameter W and each data point Z_i of the training data set Z^m. The bound, summarized in Lemma 4.2 and Corollary 4.2, applies under the following assumption.

Assumption 4.2: For fixed hyperparameter $u\in\mathcal{U}$, the loss function $l(W,Z)$ is a $(\varPsi_+,\varPsi_-,b_+,b_-)$-generalized sub-Gaussian random variable when variables W and Z are conditionally independent as $(W,Z)\sim P_{W|u,k}P_{Z|k}$, where $P_{W|u,k}\in\mathcal{P}(\mathcal{W})$ is the marginal of the joint distribution $P_{W,Z^m|k,u}$.

Under Assumption 4.2, we have the following bound on the generalization gap (5).

Lemma 4.2 ([22]): Under Assumption 4.2, for any base learner $W\sim P_{W|Z^m,u}$ with fixed hyperparameter vector $u\in\mathcal{U}$, the following bounds hold on the generalization gap (5)

$$\pm\,\Delta L(P_{Z|k},P_{W|Z^m,u})\leq\frac{1}{m}\sum_{i=1}^{m}\varPsi_{\mp}^{*-1}\Big(I(W;Z_i)\Big). \tag{21}$$

Corollary 4.2 ([22]): If the loss function $l(W,Z)$ is a σ^2-sub-Gaussian random variable when $(W,Z)\sim P_{W|u,k}P_{Z|k}$, we have the inequalities

$$|\Delta L(P_{Z|k},P_{W|Z^m,u})|\leq\frac{1}{m}\sum_{i=1}^{m}\sqrt{2\sigma^2 I(W;Z_i)}. \tag{22}$$

We note that, in general, Assumption 4.1 does not imply Assumption 4.2 (see [36, Appendix C]), and vice versa (see [22]). There are, however, loss functions $l(w,z)$ and relevant distributions for which both the assumptions hold, including the case of loss functions $l(\cdot,\cdot)$ which almost surely takes values in a bounded interval $[a,b]$. In such cases, it can be seen that the ISMI bound (22) is tighter than (20), i.e.,

$$\frac{1}{m}\sum_{i=1}^{m}\sqrt{2\sigma^2 I(W;Z_i)}\leq\sqrt{\frac{2\sigma^2}{m}I(W;Z^m)}. \tag{23}$$

The inequality in (23) follows from the chain rule of mutual information and Jensen's inequality [22].

V. Information-Theoretic Generalization Bounds for Meta-Learning

In this section, we first derive novel MI-based bounds on the meta-generalization gap with separate within-task training and test sets, as introduced in Section V-A, and then we consider joint within-task training and test sets, as described in Section V-B.

A. Bounds on Meta-Generalization Gap with Separate Within-Task Training and Test Sets

In this section, we present two novel MI-based bounds on the meta-generalization gap (10) for the setup with separate within-task training and testing sets. The first is an MI-based bound, which is akin to Lemma 4.1, and the second is an Individual Task MI (ITMI) bound, which resembles Lemma 4.2 for conventional learning. We start by defining the training loss for the meta-training sub-data set on average with respect to the training procedure as a function of the hyperparameter u as

$$L^{\mathrm{sep}}_{Z^m}(u) = \mathbb{E}_{P_{W|Z^{m_{\mathrm{tr}}},u}}\big[L_{Z^{m_{\mathrm{te}}}}(W)\big]. \tag{24}$$

1) MI-Based Bound: In order to derive the MI-based bound, we make the following assumption on $L^{\mathrm{sep}}_{Z^m}(u)$ in (24).

Assumption 5.1: For all $u \in \mathcal{U}$, the average training loss $L^{\mathrm{sep}}_{Z^m}(u)$ is a $(\Psi_+, \Psi_-, \infty, -\infty)$-generalized sub-Gaussian random variable under $Z^m \sim P_{Z^m}$, where P_{Z^m} is the marginal of the joint distribution P_{K,Z^m}.

A sufficient condition for Assumption 5.1 to hold, which is easier to check, is given next.

Lemma 5.1: If the loss function $l(\cdot,\cdot)$ is $[a,b]-$bounded almost surely, then $L^{\mathrm{sep}}_{Z^m}(\cdot)$ is also $[a,b]$ bounded for all $Z^m \in \mathcal{Z}^m$ with probability one. Consequently, $L^{\mathrm{sep}}_{Z^m}(u)$ is a $(\lambda^2(b-a)^2/8, \lambda^2(b-a)^2/8, +\infty, -\infty)$-generalized sub-Gaussian random variable under $Z^m \sim P_{Z^m}$ for all $u \in \mathcal{U}$.

Theorem 5.1: Let Assumption 5.1 hold for the base learner $P_{W|Z^{m_{\mathrm{tr}}},u}$. Then, for any meta learner $P_{U|Z^m_{1:N}}$ such that the inequality $I(U; Z^m_{1:N}) < \infty$ holds, we have the following bounds on the meta-generalization gap

$$\pm\Delta \mathrm{L}^{\mathrm{sep}}(P_{K,Z^m}, P_{U|Z^m_{1:N}}, P_{W|Z^{m_{\mathrm{tr}}},U}) \leq \Psi^{*-1}_{\mp}\left(\frac{1}{N} I(U; Z^m_{1:N})\right). \tag{25}$$

Proof: See Appendix B. ■

Specializing to the case when average training loss $L^{\mathrm{sep}}_{Z^m}(u)$ is σ^2-sub-Gaussian for all $u \in \mathcal{U}$ under $Z^m \sim P_{Z^m}$, the following upper bound on (10) holds.

Corollary 5.2: If $L^{\mathrm{sep}}_{Z^m}(u)$ is σ^2-sub Gaussian for all $u \in \mathcal{U}$ under $Z^m \sim P_{Z^m}$, we have the following bound

$$\left|\Delta \mathrm{L}^{\mathrm{sep}}(P_{K,Z^m}, P_{U|Z^m_{1:N}}, P_{W|Z^{m_{\mathrm{tr}}},U})\right| \leq \sqrt{\frac{2\sigma^2}{N} I(U; Z^m_{1:N})}. \tag{26}$$

In order to prove Theorem 5.1, one needs to overcome an additional challenge as compared to the derivation of bounds for learning reviewed in Section IV. In fact, the meta-generalization gap is caused by two distinct sources of uncertainty: (a) *environment-level uncertainty* due to finite number N of observed tasks, and (b) *within-task uncertainty* resulting from the finite number m of per-task data samples. Our proof approach involves applying the single-task MI-based bound in Lemma 4.1 to bound the effect of both sources of uncertainties.

Towards this, we start by introducing the average training loss for the randomly selected meta-test task as

$$\mathrm{L}^{\mathrm{sep}}_{P_{K,Z^m}}(u) = \mathbb{E}_{P_{K,Z^m}}[L^{\mathrm{sep}}_{Z^m}(u)]. \tag{27}$$

Note that this differs from the meta-test loss $\mathrm{L}_{P_{K,Z^{m_{\mathrm{tr}}}}}$ in (9) in that the per-task loss is evaluated in (27) on the training set. With this definition the meta-generalization gap can be decomposed as

$$\begin{aligned}&\Delta \mathrm{L}^{\mathrm{sep}}(P_{K,Z^m}, P_{U|Z^m_{1:N}}, P_{W|Z^{m_{\mathrm{tr}}},U})\\ &= \mathbb{E}_{P_{Z^m_{1:N},U}}\left[(\mathrm{L}^{\mathrm{sep}}_{P_{K,Z^m}}(U) - \mathrm{L}^{\mathrm{sep}}_{Z^m_{1:N}}(U)) + (\mathrm{L}_{P_{K,Z^{m_{\mathrm{tr}}}}}(U) - \mathrm{L}^{\mathrm{sep}}_{P_{K,Z^m}}(U))\right].\end{aligned} \tag{28}$$

In (28), the difference $\mathrm{L}^{\mathrm{sep}}_{P_{K,Z^m}}(u) - \mathrm{L}^{\mathrm{sep}}_{Z^m_{1:N}}(u)$, arises from the observation of a finite number N of tasks. In fact, as N increases, the meta-training loss $\mathrm{L}^{\mathrm{sep}}_{Z^m_{1:N}}(u)$ almost surely tends to $\mathrm{L}^{\mathrm{sep}}_{P_{K,Z^m}}(u)$ by the law of large numbers. However, the average $\mathbb{E}_{P_{Z^m_{1:N},U}}\left[\mathrm{L}^{\mathrm{sep}}_{P_{K,Z^m}}(U) - \mathrm{L}^{\mathrm{sep}}_{Z^m_{1:N}}(U)\right]$ is not equal to zero in general for finite values of N. The within-task generalization gap is instead measured by the difference $\mathrm{L}_{P_{K,Z^{m_{\mathrm{tr}}}}}(u) - \mathrm{L}^{\mathrm{sep}}_{P_{K,Z^m}}(u)$. In the setup under study with separate within-task training and test sets, this term equals zero since as we discussed, $\mathrm{L}^{\mathrm{sep}}_{P_{K,Z^m}}(u)$ is an unbiased estimate of $L_{P_{K,Z^{m_{\mathrm{tr}}}}}(u)$ (cf. (27)). This is no longer true for joint within-task training and test sets, as we discuss in Section V-B.

We note that this approach follows the main steps of the bounding techniques introduced in [11, equation (6)]. In contrast, the PAC-Bayesian bounds in [16], [17] rely on a nested application of the single-task PAC-Bayesian bounds [32], [34] combined via a union bound argument.

The bounds (25) and (26) relate the meta-generalization gap to the information-theoretic stability of the meta-training procedure. As first introduced here, this stability is measured by the MI $I(U;Z^m_{1:N})$ between the hyperparameter U and the meta-training data set $Z^m_{1:N}$, in a manner similar to the MI-based bounds in Lemma 4.1 and Corollary 4.1 for conventional learning. Importantly, as we will discuss in Section V-B, this direct parallel between learning and meta-learning no longer applies with joint within-task training and test data sets.

2) ITMI Bound: We now present the ITMI bound, which holds under the following assumption.

Assumption 5.2: The average training loss $L^{\rm sep}_{Z^m}(U)$ is a $(\varPsi_+, \varPsi_-, b_+, b_-)$-generalized sub-Gaussian random variable when variables U and Z^m are conditionally independent as $(U, Z^m) \sim P_U P_{Z^m}$, where P_U is the marginal of the joint distribution $P_{Z^m_{1:N},U}$ and P_{Z^m} is the marginal of the joint distribution P_{K,Z^m}.

Assumption 5.2 can be seen to be implied by the sufficient conditions in Lemma 5.1.

Theorem 5.3: Let Assumption 5.2 hold for the base learner $P_{W|Z^{m_{\rm tr}},U}$. Then, for any meta learner $P_{U|Z^m_{1:N}}$, the following bounds hold on the meta-generalization gap (10)

$$\pm\Delta{\rm L}^{\rm sep}(P_{K,Z^m}, P_{U|Z^m_{1:N}}, P_{W|Z^{m_{\rm tr}},U}) \leq \frac{1}{N}\sum_{i=1}^{N} \varPsi^{*-1}_{\mp}\bigg(I(U;Z^m_i)\bigg), \tag{29}$$

where the MI $I(U;Z^m_i)$ is computed with respect to the joint distribution $P_{Z^m_i,U}$ obtained by marginalizing the probability distribution $P_{Z^m_{1:N},U}$.

Proof: See Appendix B. ■

Corollary 5.4: If the average training loss $L^{\rm sep}_{Z^m}(U)$ is σ^2-sub-Gaussian when $(U, Z^m) \sim P_U P_{Z^m}$, the following bounds hold on the meta-generalization gap (10)

$$\bigg|\Delta{\rm L}^{\rm sep}(P_{K,Z^m}, P_{U|Z^m_{1:N}}, P_{W|Z^{m_{\rm tr}},U})\bigg| \leq \frac{1}{N}\sum_{i=1}^{N} \sqrt{2\sigma^2 I(U;Z^m_i)}. \tag{30}$$

As can be seen from (29) and (30), the ITMI bound on the meta-generalization gap is in terms of the MI $I(U;Z^m_i)$ between the output U of the meta learner and each per-task data set Z^m_i. This, in turn, quantifies the sensitivity of the meta learner output to the replacement of a single per-task data set. Moreover, under the sufficient conditions in Lemma 5.1, the ITMI bound

(30) yields a tighter bound than the MI-based bound (26). This can be seen from the following sequence of relations

$$\sqrt{\frac{1}{N}I(U;Z^m_{1:N})} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} I(U;Z^m_i|Z^m_{(i-1)})} \tag{31a}$$

$$\overset{(a)}{\geq} \sqrt{\frac{1}{N}\sum_{i=1}^{N} I(U;Z^m_i)} \tag{31b}$$

$$\overset{(b)}{\geq} \frac{1}{N}\sum_{i=1}^{N} \sqrt{I(U;Z^m_i)}, \tag{31c}$$

where $Z^m_{(i-1)} = (Z^m_1, \ldots, Z^m_{i-1})$; (a) follows since Z^m_i is independent of $Z^m_{(i-1)}$; and (b) follows from Jensen's inequality.

B. Bounds on Generalization Gap with Joint Within-Task Training and Test Sets

We now derive MI and ITMI-based bounds on the meta-generalization gap in (13) for the case with joint within-task training and test sets. As we will see, the key difference with respect to the case with separate within-task training and test sets is that the uncertainty due to finite number of per-task samples, measured by the second term in the decomposition (28), contributes in a non-negligible way to the meta-generalization gap. Since there is no split into separate within-task training and test sets, the average training loss with respect to the training procedure is given as (cf. (24))

$$L^{\text{joint}}_{Z^m}(u) = \mathbb{E}_{P_{W|Z^m,u}}\big[L_{Z^m}(W)\big]. \tag{32}$$

1) MI-based Bound: In order to derive the MI-based bound, we make the following assumptions.

Assumption 5.3:

(1) For each task $k \in \mathcal{K}$, the loss function $l(w,Z)$ is $(\mathit{\Psi}_{k,+}, \mathit{\Psi}_{k,-}, \infty, -\infty)$-generalized sub-Gaussian for all $w \in \mathcal{W}$ under $Z \sim P_{Z|k}$.

(2) The average training loss $L^{\text{joint}}_{Z^m}(u)$ in (32) is $(\mathit{\Gamma}_+, \mathit{\Gamma}_-, \infty, -\infty)$-generalized sub-Gaussian for all $u \in \mathcal{U}$ when $Z^m \sim P_{Z^m}$.

An easily verifiable sufficient condition for the above assumption to hold is the boundedness of loss function $l(w,z)$, which follows in a manner similar to Lemma 5.1.

Theorem 5.5: Let Assumption 5.3 hold for a base learner $W \sim P_{W|Z_k^m,U}$. Then, for any meta learner $P_{U|Z_{1:N}^m}$, we have the following bound on the meta-generalization gap (13)

$$\begin{aligned} &\pm \Delta \mathrm{L}^{\text{joint}}(P_{K,Z^m}, P_{U|Z_{1:N}^m}, P_{W|Z^m,U}) \\ &\leq \mathit{\Gamma}_{\mp}^{*-1}\biggl(\frac{1}{N} I(U; Z_{1:N}^m)\biggr) + \mathbb{E}_{P_K}\biggl[\mathit{\Psi}_{K,\mp}^{*-1}\biggl(\frac{1}{m} I(W; Z^m|K=k)\biggr)\biggr] \\ &\leq \mathit{\Gamma}_{\mp}^{*-1}\biggl(\frac{1}{N} I(U; Z_{1:N}^m)\biggr) + \sup_{k \in \mathcal{K}}\biggl[\mathit{\Psi}_{k,\mp}^{*-1}\biggl(\frac{1}{m} I(W; Z^m|k)\biggr)\biggr], \end{aligned} \tag{33}$$

where the MI $I(W; Z^m|k)$ is evaluated with respect to the distribution $P_{Z^m,W|k}$ obtained by marginalizing the joint distribution $P_{W|Z^m,U} P_{Z_{1:N}^m,U} P_{Z^m|k}$.

Proof: See Appendix C. ■

In (33), the second looser bound based on the maximization over tasks $k \in \mathcal{K}$ follows the bounding argument in [11, (6)]. In order to gain insight into the significance of the bound in Theorem 5.5, it is useful to consider the following special case, which encompasses the setup in which the loss function $l(\cdot, \cdot)$ is bounded.

Corollary 5.6: If for each task $k \in \mathcal{K}$, the loss function $l(w, Z)$ is δ_k^2-sub-Gaussian for all $w \in \mathcal{W}$ under $Z \sim P_{Z|k}$, and $L_{Z^m}^{\text{joint}}(u)$ is σ^2-sub-Gaussian for all $u \in \mathcal{U}$ under $Z^m \sim P_{Z^m}$, the following bounds on the meta-generalization gap (13) holds

$$\biggl|\Delta \mathrm{L}^{\text{joint}}(P_{K,Z^m}, P_{U|Z_{1:N}^m}, P_{W|Z^m,U})\biggr| \leq \sqrt{\frac{2\sigma^2}{N} I(U; Z_{1:N}^m)} + \sup_{k \in \mathcal{K}} \sqrt{\frac{2\delta_k^2}{m} I(W; Z^m|k)}. \tag{34}$$

With joint within-task training and test sets, the bounds (33) and (34) on the meta-generalization gap contain the contributions of two mutual informations. The first, $I(U; Z_{1:N}^m)$, quantifies the sensitivity of the meta learner output U to the meta-training data set $Z_{1:N}^m$. This term also appeared in the bounds (25) and (26) with separate within-task training and test sets. Decomposing the meta-generalization gap in a manner analogous to (28), it corresponds to a bound on the average of the first difference. The second contribution, $I(W; Z^m|k)$, quantifies the sensitivity of the output of the base learner $P_{W|Z^m,U}$ to the per-task data set Z^m, when the hyperparameter is randomly selected by the meta-learner $P_{U|Z_{1:N}^m}$ using the meta-training set $Z_{1:N}^m$. This second term is in line with the single-task generalization gap bounds (19) and (20), and it bounds the corresponding second difference in the decomposition (28). Similar meta-generalization bounds with two contributions- one applying across tasks and one within-task were derived in [12], [16], [17] using PAC-Bayesian arguments.

2) ITMI Bound on (13)*:* For deriving the ITMI bound on the meta-generalization gap (13), we assume the following.

Assumption 5.4:

(1) For each task $k \in \mathcal{K}$, the loss function $l(W,Z)$ is $(\varPsi_{k,+},\varPsi_{k,-},b_+,b_-)$-generalized sub-Gaussian when $(W,Z) \sim P_{W|k}P_{Z|k}$, where $P_{W|k}$ is the marginal of the joint distribution $P_{W|Z^m,U}P_{Z^m_{1:N},U}P_{Z^m|k}$.

(2) The function $L^{\rm joint}_{Z^m}(U)$ is $(\varGamma_+,\varGamma_-,b_+,b_-)$-generalized sub-Gaussian when $(U,Z^m) \sim P_U P_{Z^m}$.

As in Section V-A2, Assumption 5.4 can be seen to be implied by the sufficient conditions in Lemma 5.1.

Theorem 5.7: Under Assumption 5.4, for any meta learner $P_{U|Z^m_{1:N}}$, the following bounds hold on the meta-generalization gap

$$\begin{aligned}
&\pm \Delta \mathrm{L}^{\rm joint}(P_{K,Z^m},P_{U|Z^m_{1:N}},P_{W|Z^m,U})\\
&\leq \frac{1}{N}\sum_{i=1}^N \varGamma^{*-1}_{\mp}\Big(I(U;Z^m_i)\Big)+\mathbb{E}_{P_K}\left[\sum_{j=1}^m \frac{1}{m}\varPsi^{*-1}_{K,\mp}\Big(I(W;Z_j|K=k)\Big)\right]\\
&\leq \frac{1}{N}\sum_{i=1}^N \varGamma^{*-1}_{\mp}\Big(I(U;Z^m_i)\Big)+\sup_{k\in\mathcal{K}}\left[\sum_{j=1}^m \frac{1}{m}\varPsi^{*-1}_{k,\mp}\Big(I(W;Z_j|k)\Big)\right],
\end{aligned} \tag{35}$$

where the MI $I(U;Z^m_i)$ is evaluated with respect to $P_{Z^m_i,U}$ obtained by marginalizing $P_{Z^m_{1:N},U}$, and the MI $I(W;Z_j|k)$ is with respect to $P_{Z_j,W|k}$ obtained by marginalizing $P_{Z^m,W|k}$.

Proof: See Appendix C. ■

We have the following special case.

Corollary 5.8: If for each task $k \in \mathcal{K}$, $l(W,Z)$ is δ_k^2-sub-Gaussian when $(W,Z) \sim P_{W|k}P_{Z|k}$, and $L^{\rm joint}_{Z^m}(U)$ is σ^2-sub-Gaussian when $(U,Z^m) \sim P_U P_{Z^m}$, the following bound holds

$$\begin{aligned}
\Big|\Delta \mathrm{L}^{\rm joint}(P_{K,Z^m},P_{U|Z^m_{1:N}},P_{W|Z^m,U})\Big| &\leq \frac{1}{N}\sum_{i=1}^N \sqrt{2\sigma^2 I(U;Z^m_i)}\\
&\quad+\sup_{k\in\mathcal{K}}\frac{1}{m}\sum_{j=1}^m \sqrt{2\delta_k^2 I(W;Z_j|k)}.
\end{aligned} \tag{36}$$

Similar to the bounds in (33) and (34), the bounds on meta-generalization gap in (35) and (36) are in terms of two types of mutual informations, the first describing the sensitivity of the meta-learner and the second the sensitivity of the base learner. Specifically, the MI $I(U;Z^m_i)$ quantifies the sensitivity of the output of the meta learner to per-task data set Z^m_i, and the MI $I(W;Z_j|k)$ measures the sensitivity of the output of the base learner, $P_{W|Z^m,U}$ to each data sample Z_i within its training set. Moreover, it can be shown, in a manner similar to (31c), that, under the sufficient conditions of Lemma 5.1, the ITMI bound in (36) is tighter than the MI bound in (34).

VI. Applications

In this section, we consider two applications of the information-theoretic bounds proposed in Section V-A. The first, simpler, example concerns a parameter estimation problem for which an optimized meta-learner can be obtained in closed form. In contrast, the second application covers a broad class of iterative meta-training schemes.

A. Parameter Estimation

To illustrate the bounds on the meta-generalization gap derived in Section V-A, we first consider the problem of prediction for a Bernoulli process with a 'soft' predictor that uses only a few samples from the process, as well as meta-training data. The data distribution $P_{Z|k}$ for each task $k \in \mathcal{K}$ is given as $\text{Bernoulli}(\mu_k)$ with mean parameter μ_k. The task distribution P_K is defined over an arbitrary discrete finite set of mean parameters $\{\mu_1, \ldots, \mu_M\}$. The base learner uses training data, distributed i.i.d. from $\text{Bernoulli}(\mu_k)$, to determine the parameter W, which is used as a predictor of new observation $Z \sim \text{Bernoulli}(\mu_k)$ at test time. The loss function is defined as $l(w, z) = (w - z)^2$, measuring the quadratic error between prediction and realized test input z. Note that the optimal (Bayes) predictor, computable in the ideal case of known distribution $P_{Z|k}$, is given as $W = \mu_k$. We now distinguish the two cases with separate and joint within-task training and test sets.

1) Separate within-task training and test sets: The base learner $P_{W|Z_k^{\text{m}_{\text{tr}}},u}$ for task $k \in \mathcal{K}$ deterministically selects the prediction

$$W_k = \alpha D_k^{\text{m}_{\text{tr}}} + (1-\alpha)u, \tag{37}$$

where $D_k^{\text{m}_{\text{tr}}} = \frac{1}{m_{\text{tr}}} \sum_{j=1}^{m_{\text{tr}}} Z_{k,j}^{\text{m}_{\text{tr}}}$, is an empirical average over the training set, u is a hyperparameter defining a bias that can be meta-trained, and $\alpha \in [0, 1]$ is a fixed scalar. Here, $Z_{k,j}^{\text{m}_{\text{tr}}}$ denote the jth data sample in the training set of task k. The bias term in (37) may help approximate the ideal Bayes predictor in the presence of limited data $Z_k^{\text{m}_{\text{tr}}}$.

The objective of the meta-learner is to infer the hyperparameter u. For a given meta-training data set $Z_{1:N}^m$, the meta-learner can compute the empirical meta-training loss as

$$\text{L}_{Z_{1:N}^m}^{\text{sep}}(u) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{m_{\text{te}}} \sum_{j=1}^{m_{\text{te}}} (W_i - Z_{i,j}^{\text{m}_{\text{te}}})^2, \tag{38}$$

where $Z_{i,j}^{\mathrm{m_{te}}}$ denote the jth data sample in the test set of Z_i^m, the ith sub-data set of $Z_{1:N}^m$. The meta-learner $P_{U|Z_{1:N}^m}$ then deterministically selects the minimizing hyperparameter u of the meta-training empirical loss function in (38). This optimization yields

$$U = \frac{(1-\alpha)^{-1}}{N}\biggl(\sum_{i=1}^{N} D_i^{\mathrm{m_{te}}} - \alpha D_i^{\mathrm{m_{tr}}}\biggr), \quad \text{for} \quad 0 \le \alpha < 1, \tag{39}$$

where $D_i^{\mathrm{m_{te}}} = \sum_{j=1}^{m_{\mathrm{te}}} Z_{i,j}^{\mathrm{m_{te}}}/m_{\mathrm{te}}$. Note that, by (39), we can take without loss of optimality the domain $\mathcal{U}$ to be the interval $\mathcal{U} = [-\alpha(1-\alpha)^{-1}, (1-\alpha)^{-1}]$. The meta-test loss can be explicitly computed as

$$\mathrm{L}_{P_{K,Z^m}}(u) = (1-\alpha)^2\bigl(u^2 - 2u\mathbb{E}_{P_K}[\mu_K]\bigr) + \mathbb{E}_{P_K}\biggl[\alpha^2\bigl(\mu_K^2 + \frac{\mu_K\bar{\mu}_K}{m_{\mathrm{tr}}}\bigr) + \mu_K - 2\alpha\mu_K^2\biggr], \tag{40}$$

where $\bar{\mu}_K = 1-\mu_K$, and the meta-generalization gap evaluates to

$$\begin{aligned}\Delta\mathrm{L}^{\mathrm{sep}}(P_{K,Z^m}, P_{U|Z_{1:N}^m}, P_{W|Z^{\mathrm{m_{tr}}},U}) = \frac{2(1-\alpha)^2}{N}\biggl(\mathbb{E}_{P_K}[\mu_K^2] - (\mathbb{E}_{P_K}[\mu_K])^2\biggr)\\ + \frac{2\mathbb{E}_{P_K}[\mu_K\bar{\mu}_K]}{N}\biggl(\frac{1}{m_{\mathrm{te}}} + \frac{\alpha^2}{m_{\mathrm{tr}}}\biggr).\end{aligned} \tag{41}$$

To compute the MI and ITMI-based bounds on the meta-generalization gap (41), it is easy to verify that the average training loss $L_{Z^m}^{\mathrm{sep}}(\cdot)$ is almost surely bounded, i.e., $0 \le L_{Z^m}^{\mathrm{sep}}(u) \le (1+\alpha)^2$ for all $u \in \mathcal{U}$ and $Z^m \in \mathcal{Z}^m$. Thus, Assumption 5.1 for the MI bound and also Assumption 5.2 for the ITMI bound hold. Since the meta-learner is deterministic, we have the equality $I(U; Z_{1:N}^m) = H(U)$, whereby the MI-based bound (26) is evaluated as

$$|\Delta\mathrm{L}^{\mathrm{sep}}(P_{K,Z^m}, P_{U|Z_{1:N}^m}, P_{W|Z^{\mathrm{m_{tr}}},U})| \le \sqrt{\frac{(1+\alpha)^4}{4N}H(U)}, \tag{42}$$

and the ITMI bound (30) is given as

$$|\Delta\mathrm{L}^{\mathrm{sep}}(P_{K,Z^m}, P_{U|Z_{1:N}^m}, P_{W|Z^{\mathrm{m_{tr}}},U})| \le \frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{(1+\alpha)^4}{4}I(U; Z_i^m)}. \tag{43}$$

The information-theoretic measures in (42) and (43) can be evaluated numerically as discussed in Appendix D. For a numerical illustration, Figure 4 plots the average of the meta-test loss (40) and average meta-training loss (63) along with the MI-based bound in (42) and the ITMI bound in (43). It can be seen that the ITMI bound is tighter than the MI-based bound. Furthermore, both bounds correctly predict the decrease in the meta-generalization gap as the number N of tasks increases.

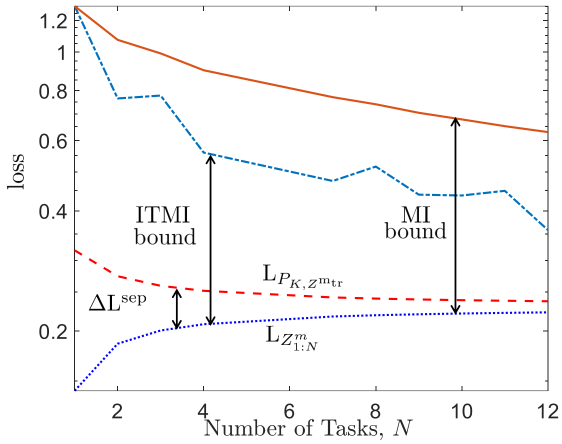

Fig. 4: Comparison of the MI and ITMI based bounds obtained in (42) and (43) with the meta-generalization gap for meta-learning with separate within-task training and test sets. The task environment is defined by $M = 12$ tasks distributed according to $P_K = [0.1136; 0.0999; 0.0138; 0.0810; 0.0644; 0.0825; 0.1148; 0.0044; 0.1513; 0.0086; 0.1517; 0.1140]$. Other parameters are set as $\alpha = 0.15$, $m_{\mathrm{tr}} = 15$, $m_{\mathrm{te}} = 5$.

2) Joint Within-Task Training and Testing sets: We now consider the case with joint within-task training and test sets. The base learner $P_{W|Z_k^m,U}$ for task $k \in \mathcal{K}$ still uses the predictor (37), but now the empirical average over the training set is given as $D_k = \sum_{j=1}^m Z_{k,j}^m/m$. As before, the meta-learner $P_{U|Z_{1:N}^m}$ deterministically selects the minimizing hyperparameter u of the meta-training empirical loss function, $L_{Z_{1:N}^m}(u) = (1/N)\sum_{i=1}^N (1/m)\sum_{j=1}^m (W_i - Z_{i,j}^m)^2$, yielding $U = \frac{1}{N}\sum_{i=1}^N D_i$. As discussed in Appendix D, the meta-test loss for this example can also be explicitly computed and the meta-generalization gap bounds in (34) and (36) can be evaluated numerically. Figure 5 plots the average meta-test loss and average meta-training loss along with the MI-based bound in (65) and the ITMI bound in (66), as a function of per-task data samples m. The ITMI bound is seen not only to be tighter than the MI bound, but also to better reflect the decrease of the meta-training loss as a function of m.

B. Noisy Iterative Meta-Learning Algorithms

Most meta-learning algorithms are built around a nested loop structure, with the inner loop applying the base learner on the meta-training set and the outer loop updating the hyperparameters U. In this section, we focus on a vast class of such meta-learning algorithms in which the

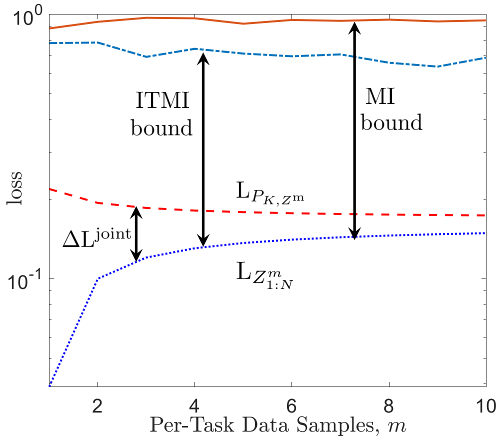

Fig. 5: Comparison of the MI and ITMI based bounds obtained in (65) and (66) with the meta-generalization gap for meta-learning with joint within-task training and test sets, as a function of the per-task data samples m for $N = 5$ and $\alpha = 0.55$. The task environment is defined by $M = 9$ tasks distributed according to $P_K = [0.1699; 0.1807; 0.1318; 0.1157; 0.1243; 0.1326; 0.0394; 0.0107; 0.0949]$.

inner loop applies training procedures dependent on the current iterate of the hyperparameter, while the outer loop updates the hyperparameter using a stochastic rule. This class includes stochastic variants of state-of-the-art algorithms such as MAML [20] and Reptile [21]. We apply the derived information-theoretic bounds to study the meta-generalization performance of the mentioned class of meta-training iterative stochastic rules by focusing on the case of separate within-task training and test sets here, which is assumed e.g., by MAML. The analysis for the setup with joint within-task training and test sets can also be carried out at the cost of a more cumbersome notation.

To start, let $U^t \in \mathbb{R}^d$ denote the hyperparameter vector at outer iteration t, with $U^0 \in \mathbb{R}^d$ being an arbitrary initialization. For example, in MAML, the hyperparameter U defines the initial iterate used by each base learner $k \in \mathcal{K}$ in the inner loop to update the model parameter W_k. At each iteration $t \geq 1$, we sample a mini-batch of tasks $K_t \subseteq [1, \ldots, N]$ from the meta-training data $Z^m_{1:N}$, obtaining the corresponding data set $Z^m_{K_t} = (Z^{\mathrm{m_{tr}}}_{K_t}, Z^{\mathrm{m_{te}}}_{K_t}) \subseteq Z^m_{1:N}$, where $Z^{\mathrm{m_{tr}}}_{K_t} = \{Z^{\mathrm{m_{tr}}}_k\}_{k \in K_t}$ and $Z^{\mathrm{m_{te}}}_{K_t} = \{Z^{\mathrm{m_{te}}}_k\}_{k \in K_t}$ are the separate training and test sets for the selected tasks. For each task $k \in K_t$, in the inner loop, the base learner selects the model parameter W_k^t as a, possibly stochastic, function

$$W_k^t = g(U^{t-1}, Z_k^{\mathrm{m_{tr}}}). \tag{44}$$

For instance, in MAML, the function $g(U^{t-1}, Z_k^{\mathrm{m_{tr}}}) \in \mathbb{R}^d$ in (44) represents the output of an SGD procedure that starts from initialization U^{t-1} and uses the task training data $Z_k^{\mathrm{m_{tr}}}$ to iteratively update the model parameters, producing the final iterate W_k^t. We denote as $W_{K_t} = \{W_k^t\}_{k \in K_t}$ the collection of the base learners' outputs for all tasks $k \in K_t$ at outer iteration t.

In the outer loop, the meta learner uses the task-specific adapted parameters W_{K_t} from the inner loop and the meta-test set $Z_{K_t}^{\mathrm{m_{te}}}$ to update the past iterate U^{t-1} according to the general update rule

$$U^t = F(U^{t-1}) + \beta_t G(U^{t-1}, W_{K_t}, Z_{K_t}^{\mathrm{m_{te}}}) + \xi_t, \tag{45}$$

where $F(\cdot)$ and $G(\cdot, \cdot, \cdot)$ are arbitrary deterministic functions; β_t is the step-size; and $\xi_t \sim \mathcal{N}(0, \gamma_t^2 I_d)$ is an isotropic Gaussian noise, independently drawn for $t = 1, 2, \ldots,$. As an example, in MAML, the function $F(\cdot)$ is the identity function and function $G(\cdot, \cdot, \cdot)$ equals the gradient of the empirical loss $1/|K_t| \sum_{i \in K_t} L_{Z_i^{\mathrm{m_{te}}}}(W_i^t)$ in (8) with respect to U^{t-1}. Note, however, that MAML does not add noise, i.e., $\gamma_t^2 = 0$ for all t.

The final output of the meta-learning algorithm is then defined as an arbitrary function $U = f(U^1, \ldots, U^T)$, of all iterates. Examples of function f include the last update $f(U^1, \ldots, U^T) = f(U^T)$ and average of the updates $f(U^1, \ldots, U^T) = 1/T \sum_{t=1}^T U^t$. A graphical model representation of the variables involved is shown in Figure 6.

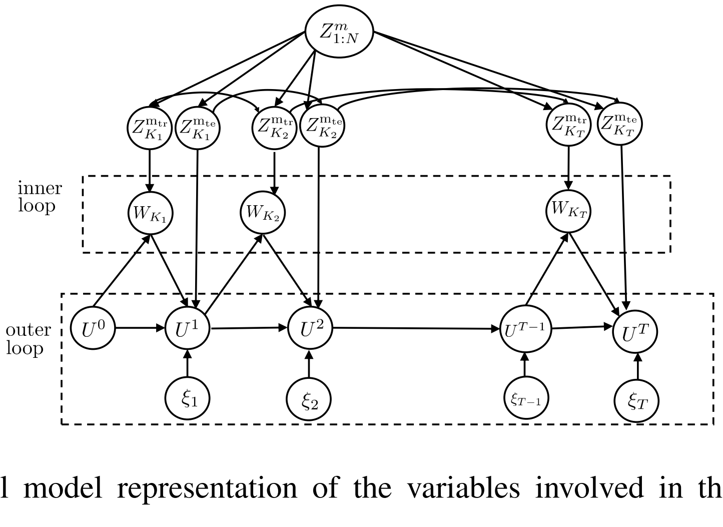
$Z_{1:N}^m$
$Z_{K_1}^{\mathrm{m_{tr}}}$
$Z_{K_1}^{\mathrm{m_{te}}}$
$Z_{K_2}^{\mathrm{m_{tr}}}$
$Z_{K_2}^{\mathrm{m_{te}}}$
$Z_{K_T}^{\mathrm{m_{tr}}}$
$Z_{K_T}^{\mathrm{m_{te}}}$
inner loop
W_{K_1}
W_{K_2}
W_{K_T}
outer loop
U^0
U^1
U^2
U^{T-1}
U^T
ξ_1
ξ_2
ξ_{T-1}
ξ_T

Fig. 6: A graphical model representation of the variables involved in the definition of noisy iterative algorithms.

We now derive an upper bound on the meta-generalization gap for the general class of iterative meta-learning algorithm satisfying (44) - (45) under the following assumptions.

Assumption 6.1:

(1) For the base-learner given in (44), the average training loss $L^{\mathrm{sep}}_{Z^m}(u)$ in (24) is σ^2-sub-Gaussian for all $u \in \mathcal{U}$ when $Z^m \sim P_{Z^m}$;

(2) The meta-training data set $Z^m_{K_t}$ sampled at each iteration t is conditionally independent of the history of model-parameter vectors $\{W_{K_j}\}_{j=1}^{t-1}$ and hyperparameter $U^{(t-1)} = (U^1, U^2, \dots, U^{t-1})$, i.e.,

$$P_{Z^m_{K_t}|\{Z^m_{K_j}\}_{j=1}^{t-1}, Z^m_{1:N}, U^{(t-1)}, \{W_{K_j}\}_{j=1}^{t-1}} = P_{Z^m_{K_t}|\{Z^m_{K_j}\}_{j=1}^{t-1}, Z^m_{1:N}}; \tag{46}$$

(3) The meta-parameter update function $G(\cdot,\cdot,\cdot)$ is uniformly bounded, i.e., $||G(\cdot,\cdot,\cdot)||_2 \le L$ for some $L > 0$.

Lemma 6.1: Under Assumption 6.1, the following upper bound on the meta-generalization gap (10) holds for the class of noisy iterative meta-training algorithms (44)-(45)

$$\Delta \mathrm{L}^{\mathrm{sep}}(P_{K,Z^m}, P_{U|Z^m_{1:N}}, P_{W|Z^{m\mathrm{tr}},U}) \le \sqrt{\frac{2\sigma^2}{N}\sum_{t=1}^{T}\frac{d}{2}\log\left(1+\frac{\beta_t^2 L^2}{d\gamma_t^2}\right)}. \tag{47}$$

Proof: See Appendix E. ■

The bound in (47) has the same form as the generalization gap derived in [23] for conventional learning. From (47), the generalization gap can be reduced by increasing the variance γ_t^2 of the injected Gaussian noise. In particular, the meta-generalization gap depends on the ratios β_t^2/γ_t^2 between squared step size β_t^2 and variance γ_t^2. For example, SGLD sets $\gamma_t = \sqrt{\beta_t}$, and a step size β_t decaying over time according to the standard Robbins-Monro conditions in order to ensure convergence of the output samples to the generalized posterior distribution of the hyperparameters [46].

Example: To illustrate bound (47), we now consider a simple logistic regression problem that generalizes the example studied in Section VI-A. Accordingly, each data point Z corresponds to labelled data $Z = (X, Y)$, where $X \in \{0,1\}^d$ represents the input vector and $Y \in \{0,1\}$ represents the corresponding binary label. The data distribution $P_{Z|k} = P_{X|k}P_{Y|X,k}$ for each task $k \in \mathcal{K}$ is such that $X \sim P_{X|k}$ is a d-dimensional Bernoulli vector obtained via d independent draws from $\mathrm{Bernoulli}(\nu)$ and Y is distributed as $Y \sim \mathrm{Bernoulli}(\phi(\mu_k^T X))$, where $\phi(a) = 1/(1+$

$\exp(-a))$ is the sigmoid function and $\mu_k \in \mathbb{R}^d$, with $||\mu_k||_2 \leq 1$. The base-learner uses training data generated i.i.d. from $P_{Z|k}$ to obtain a prediction w of the parameter vector μ_k for task $k \in \mathcal{K}$. The loss function is taken as the quadratic error $l(w,z) = (\phi(w^T x) - y)^2$. The task environment P_K defines a distribution over the parameter vectors μ_k, which is assumed to be an arbitrary discrete probability distribution over M parameter vectors $\mu_1, \ldots, \mu_M$.

At each iteration t, starting from initialization point U^{t-1}, the base-learner in (44) uses a one-step projected gradient descent algorithm on the training data set $Z_k^{m_{\mathrm{tr}}}$ to obtain the prediction W_k^t as

$$W_k^t = \mathrm{proj}_{\mathcal{W}}\bigg(U^{t-1} - \alpha \nabla_w L_{Z_k^{m_{\mathrm{tr}}}}(w)\big|_{w=U^{t-1}}\bigg), \tag{48}$$

where $\alpha > 0$ is the step-size, $\mathcal{W} = \{w \in \mathbb{R}^d \big| \; ||w||_2 \leq 1\}$ is the set of feasible model parameters and $\mathrm{proj}_{\mathcal{A}}(b) = \frac{1}{2}\min_{a \in \mathcal{A}} ||a - b||_2^2$ is the projection operator. The meta-learner (45) updates the initialization vector according to the noisy gradient descent rule

$$U^t = U^{t-1} - \beta_t \bigg(\frac{1}{|K_t|}\sum_{i=1}^{|K_t|} \nabla_w L_{Z^{m_{\mathrm{te}}}}(w)\big|_{w=W_i^t}\bigg) + \xi_t, \tag{49}$$

where β_t is the step-size; and $\xi_t \sim \mathcal{N}(0, \gamma_t^2 I_d)$ is isotropic Gaussian noise. This update rule corresponds to performing an First Order MAML (FOMAML) [20] with the addition of noise.

For this problem, it is easy to verify that Assumption 6.1 is satisfied, since the average training loss $L^{\mathrm{sep}}_{Z^m}(u)$ is bounded almost surely, i.e., $0 \leq L^{\mathrm{sep}}_{Z^m}(u) \leq 1$ for all $u \in \mathbb{R}^d$, and we have the inequality

$$\Bigg|\Bigg|\frac{1}{|K_t|}\sum_{i=1}^{|K_t|} \nabla_w L_{Z^{m_{\mathrm{te}}}}(w)\big|_{w=W_i^t}\Bigg|\Bigg|_2 \leq 2\sqrt{d}e^{\sqrt{d}} \triangleq L. \tag{50}$$

The MI bound in (47) then evaluates to

$$\Delta \mathrm{L}^{\mathrm{sep}}(P_{K,Z^m}, P_{U|Z^m_{1:N}}, P_{W|Z^{m_{\mathrm{tr}}},U}) = \sqrt{\frac{1}{2N}\sum_{t=1}^{T} \frac{d}{2}\log\bigg(1 + \frac{4\beta_t^2 e^{2\sqrt{d}}}{\gamma_t^2}\bigg)}. \tag{51}$$

We now evaluate the meta-training and meta-test loss, along with the bound (51) as a function of the ratio γ_t^2/β_t^2 in Figure 7. For the experiment, we considered a task environment of $M = 20$ tasks with $\nu = 0.4$, $d = 3$, $N = 4$ meta-training tasks with $m_{\mathrm{tr}} = 10$ training data samples and $m_{\mathrm{te}} = 5$ test data samples. For the inner-loop (48), we fixed step-size $\alpha = 10^{-4}$ and for the outer-loop (49), we set $|K_t| = N$, $\beta_t = 0.25$ and $T = 200$ iterations.

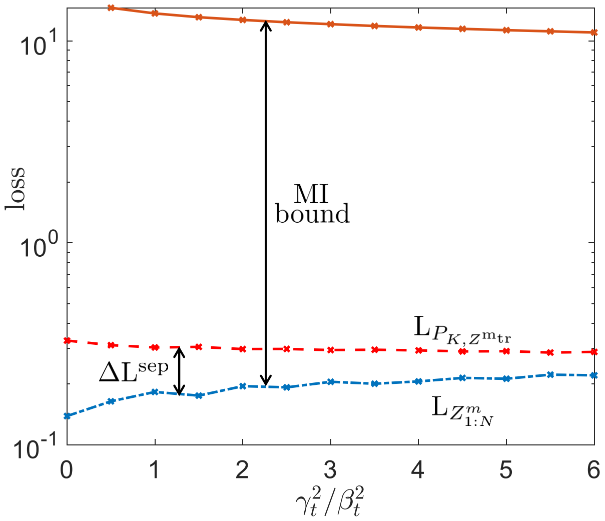

Fig. 7: Comparison of the meta-generalization gap with the MI-based bound in (51) as function of the ratio γ_t^2/β_t^2. The task distribution is given as $P_K = [0.0224; 0.0961; 0.0895; 0.0247; 0.0136; 0.0608; 0.0593; 0.0711; 0.0945; 0.0503; 0.0585; 0.0450; 0.0505; 0.0518; 0.0133; 0.0337; 0.0049; 0.0483; 0.0381; 0.0736]$.

As suggested by Lemma 6.1, the meta-generalization gap decreases with addition of noise. While the MI bound (47) is generally loose, it correctly quantifies the dependence of the meta-generalization loss and the ratio γ_t^2/β_t^2, and it can hence serve as a useful meta-training criterion [10], [16].

VII. Conclusions

This work has presented novel information-theoretic upper bounds on the generalization gap of meta-learning algorithms, thereby extending the well-studied information-theoretic approaches in conventional learning to meta-learning. The proposed bounds capture two sources of uncertainty – environment-level uncertainty and within-task uncertainty – and bound them via separate mutual information terms. Applications were also discussed with the aim of elucidating the use of the bounds to quantify meta-overfitting and guide the choice of the meta-inductive bias, i.e., the class of inductive biases. The derived bounds are amenable to further refinements such as those along the lines of [35], [36], [41]. It would also be interesting to study the meta-generalization bounds on noisy iterative meta-learning algorithms using the tighter information-theoretic bounds such as [22], [36].

Appendix A
Decoupling Estimate Lemmas

The proofs of the main results rely on the following decoupling estimate lemmas, which bound the difference in expectations under a change of measure from the joint $P_{X,Y}$ to the product of the marginals $P_X P_Y$.

Lemma A.1 (Decoupling Estimate [44]): Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be two jointly distributed random variables with joint distribution $P_{X,Y}$, and let $f(X,Y)$ be a real valued function such that $f(x,Y)$ is $(\varPsi_+, \varPsi_-, \infty, -\infty)$-generalized sub-Gaussian for all $x \in \mathcal{X}$ when $Y \sim P_Y$. Then we have the following inequalities

$$\pm\biggl(\mathbb{E}_{P_X P_Y}[f(\widetilde{X},\widetilde{Y})] - \mathbb{E}_{P_{X,Y}}[f(X,Y)]\biggr) \leq \varPsi_{\mp}^{*-1}(I(X;Y)), \tag{52}$$

where $(\widetilde{X},\widetilde{Y}) \sim P_X P_Y$.

Lemma A.2 (General Decoupling Estimate [22]): Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be two jointly distributed random variables with joint distribution $P_{X,Y}$, and let $f(X,Y)$ be a real valued function such that $f(X,Y)$ is a $(\varPsi_+, \varPsi_-, b_+, b_-)$-generalized sub-Gaussian when $(X,Y) \sim P_X P_Y$. Then, we have the inequality (52).

Appendix B
Proofs of Theorem 5.1 and Theorem 5.3

For the proof of Theorem 5.1, we use the decomposition (28) of the meta-generalization gap, which yields $\Delta\mathrm{L}^{\mathrm{sep}}(P_{K,Z^m}, P_{U|Z^m_{1:N}}, P_{W|Z^{m}{\mathrm{tr}},U}) =$

$$\mathbb{E}_{P_{Z^m_{1:N},U}}\biggl[\mathrm{L}^{\mathrm{sep}}_{P_{K,Z^m}}(U) - \mathrm{L}^{\mathrm{sep}}_{Z^m_{1:N}}(U)\biggr] + \mathbb{E}_{P_K}\biggl[\Delta L(P_{Z|K}, P_{W|Z^{m}{\mathrm{tr}}})\biggr] \tag{53}$$

where average per-task generalization gap, $\Delta L(P_{Z|K}, P_{W|Z^{m}{\mathrm{tr}}})$, is defined as

$$\Delta L(P_{Z|K}, P_{W|Z^{m}{\mathrm{tr}}}) = \mathbb{E}_{P_{Z^m|K}}\biggl[\mathbb{E}_{P_{W,U,Z^m_{1:N}|Z^{m}{\mathrm{tr}}}}[L_{P_{Z|K}}(W) - L_{Z^{m}{\mathrm{te}}}(W)]\biggr], \tag{54}$$

with $P_{W,U,Z^m_{1:N}|Z^{m}{\mathrm{tr}}} = P_{W|Z^{m}{\mathrm{tr}},U}P_{Z^m_{1:N},U}$ and $P_{W|Z^{m}{\mathrm{tr}}}$ being its marginal distribution. The meta-generalization gap $\Delta\mathrm{L}^{\mathrm{sep}}(P_{K,Z^m}, P_{U|Z^m_{1:N}}, P_{W|Z^{m}{\mathrm{tr}},U})$ in (53) can be then bounded as

$$\begin{aligned}\pm\Delta\mathrm{L}^{\mathrm{sep}}(P_{K,Z^m}, P_{U|Z^m_{1:N}}, P_{W|Z^{m}{\mathrm{tr}},U}) \leq \pm\mathbb{E}_{P_{Z^m_{1:N},U}}\biggl[\mathrm{L}^{\mathrm{sep}}_{P_{K,Z^m}}(U) - \mathrm{L}^{\mathrm{sep}}_{Z^m_{1:N}}(U)\biggr]\\ + \sup_{k\in\mathcal{K}}\biggl[\pm\Delta L(P_{Z|k}, P_{W|Z^{m}{\mathrm{tr}}})\biggr].\end{aligned} \tag{55}$$

We first bound the term $\mathbb{E}_{P_{Z^m_{1:N},U}}\big[\mathrm{L}^{\mathrm{sep}}_{P_{K,Z^m}}(U)-\mathrm{L}^{\mathrm{sep}}_{Z^m_{1:N}}(U)\big]$, which represents the expected environment-level uncertainty measured with respect to the average training loss $L^{\mathrm{sep}}_{Z^m}(u)$ defined in (24). To this end, we extend the single-task learning generalization bound of Lemma 4.1 by resorting to the decoupling estimate in Lemma A.1 with $X=U$, $Y=Z^m_{1:N}$ and $f(X,Y)=\mathrm{L}^{\mathrm{sep}}_{Z^m_{1:N}}(U)$, so that $\mathbb{E}_{P_{X,Y}}[f(X,Y)]=\mathbb{E}_{P_{Z^m_{1:N},U}}[\mathrm{L}^{\mathrm{sep}}_{Z^m_{1:N}}(U)]$ and $\mathbb{E}_{P_XP_Y}[f(\widetilde{X},\widetilde{Y})]=\mathbb{E}_{P_{Z^m_{1:N},U}}[\mathrm{L}^{\mathrm{sep}}_{P_{K,Z^m}}(U)]$. Using Assumption 5.1, we then get for all $u\in\mathcal{U}$

$$\begin{aligned}\log\mathbb{E}_{P_{Z^m_{1:N}}}\Bigg[e^{\pm\lambda\left(\mathrm{L}^{\mathrm{sep}}_{Z^m_{1:N}}(u)-\mathbb{E}_{P_{Z^m_{1:N}}}[\mathrm{L}^{\mathrm{sep}}_{Z^m_{1:N}}(u)]\right)}\Bigg]&=\log\Bigg(\mathbb{E}_{P_{Z^m}}\Bigg[e^{\pm\lambda/N\left(L^{\mathrm{sep}}_{Z^m}(u)-\mathbb{E}_{P_{Z^m}}[L^{\mathrm{sep}}_{Z^m}(u)]\right)}\Bigg]\Bigg)^N\\&\le N\Psi_{\pm}(\lambda/N).\end{aligned}$$

Applying Lemma A.1 together with (17), we get the inequality

$$\mathbb{E}_{P_{Z^m_{1:N},U}}\Bigg[\mathrm{L}^{\mathrm{sep}}_{P_{K,Z^m}}(U)-\mathrm{L}^{\mathrm{sep}}_{Z^m_{1:N}}(U)\Bigg]\le\inf_{\lambda>0}\frac{I(U;Z^m_{1:N})+N\Psi_{-}(\lambda/N)}{\lambda}\tag{56}$$

$$=\Psi^{*-1}_{-}\Bigg(\frac{I(U;Z^m_{1:N})}{N}\Bigg).\tag{57}$$

Similarly, it can be shown that $-\mathbb{E}_{P_{Z^m_{1:N},U}}\big[\mathrm{L}^{\mathrm{sep}}_{P_{K,Z^m}}(U)-\mathrm{L}^{\mathrm{sep}}_{Z^m_{1:N}}(U)\big]\le\Psi^{*-1}_{+}\big(I(U;Z^m_{1:N})/N\big)$.

We now evaluate the second term in the right hand side of (55). It can be seen that for a fixed task $k\in\mathcal{K}$, the average within-task uncertainty evaluates to

$$\Delta L(P_{Z|k},P_{W|Z^{\mathrm{m}_{\mathrm{tr}}}})=\mathbb{E}_{P_{Z^{\mathrm{m}_{\mathrm{tr}}}|k}}\mathbb{E}_{P_{W|Z^{\mathrm{m}_{\mathrm{tr}}}}}[L_{P_{Z|k}}(W)-\mathbb{E}_{P_{Z^{\mathrm{m}_{\mathrm{te}}}|k}}L_{Z^{\mathrm{m}_{\mathrm{te}}}}(W)]\overset{(a)}{=}0,\tag{58}$$

where (a) follows since W and $Z^{\mathrm{m}_{\mathrm{te}}}$ are independent conditioned on task $k\in\mathcal{K}$ which implies that $\mathbb{E}_{P_{Z^{\mathrm{m}_{\mathrm{te}}}|k}}L_{Z^{\mathrm{m}_{\mathrm{te}}}}(W)=L_{P_{Z|k}}(W)$. Substituting (57) and (58) in (55) then concludes the proof.

For Theorem 5.3, the proof follows along the same line, and bounds the average environment-level uncertainty $\mathbb{E}_{P_{Z^m_{1:N},U}}\big[\mathrm{L}^{\mathrm{sep}}_{P_{K,Z^m}}(U)-\mathrm{L}^{\mathrm{sep}}_{Z^m_{1:N}}(U)\big]$ using the general decoupling estimate in Lemma A.2. This is done by decomposing the generalization gap $\mathbb{E}_{P_{Z^m_{1:N},U}}\big[\mathrm{L}^{\mathrm{sep}}_{P_{K,Z^m}}(U)-\mathrm{L}^{\mathrm{sep}}_{Z^m_{1:N}}(U)\big]$ across different tasks as

$$\mathbb{E}_{P_{Z^m_{1:N},U}}\Bigg[\mathrm{L}^{\mathrm{sep}}_{P_{K,Z^m}}(U)-\mathrm{L}^{\mathrm{sep}}_{Z^m_{1:N}}(U)\Bigg]=\frac{1}{N}\sum_{i=1}^{N}\Bigg(\mathbb{E}_{P_{Z^m}P_U}[L^{\mathrm{sep}}_{Z^m}(U)]-\mathbb{E}_{P_{Z^m_i,U}}[L^{\mathrm{sep}}_{Z^m_i}(U)]\Bigg),\tag{59}$$

where Z^m and U in the first term are conditionally independent random variables distributed as $(Z^m,U)\sim P_{Z^m}P_U$, while, in the second term, they are jointly distributed according to $P_{Z^m_i,U}$, which is obtained by marginalizing the joint distribution $P_{Z^m_{1:N},U}$. Now using Assumption 5.2, we apply the general decoupling estimate in Lemma A.2 to each term inside the summation in (59) with $X=U$, $Y=Z^m_i$ and $f(X,Y)=L^{\mathrm{sep}}_{Z^m_i}(U)$ so that $\mathbb{E}_{P_XP_Y}[f(\widetilde{X},\widetilde{Y})]=\mathbb{E}_{P_{Z^m}P_U}[L^{\mathrm{sep}}_{Z^m}(U)]$ and $\mathbb{E}_{P_{X,Y}}[f(X,Y)]=\mathbb{E}_{P_{Z^m_i,U}}[L^{\mathrm{sep}}_{Z^m_i}(U)]$. This yields the bound in (29).

Appendix C
Proofs of Theorem 5.5 and Theorem 5.7

For Theorem 5.5, we start from the relation in (55) with $\Delta\mathrm{L}^{\mathrm{sep}}$ replaced with $\Delta\mathrm{L}^{\mathrm{joint}}$; $\mathrm{L}^{\mathrm{sep}}_{Z^m_{1:N}}(U)$ with $\mathrm{L}^{\mathrm{joint}}_{Z^m_{1:N}}(U)$ and $\mathrm{L}^{\mathrm{sep}}_{P_{K,Z^m}}(u)$ with $\mathrm{L}^{\mathrm{joint}}_{P_{K,Z^m}}(u) = \mathbb{E}_{P_{K,Z^m}}[L^{\mathrm{joint}}_{Z^m}(u)]$. Bounds on the expected environment-level uncertainty can be obtained by using Lemma A.1 and the second assumption of Assumption 5.3 as in (57). While the average within-task uncertainty vanishes in the case of separate within-task training and test sets, this is not the case in the setup under study. Consequently, we have

$$\Delta L(P_{Z|k}, P_{W|Z^m}) = \mathbb{E}_{P_{Z^m|k}}\mathbb{E}_{P_{W|Z^m}}\biggl[L_{P_{Z|k}}(W) - L_{Z^m}(W)\biggr], \tag{60}$$

for $k \in \mathcal{K}$, where recall that $P_{W|Z^m}$ is the marginal of the joint distribution $P_{W|Z^m,U}P_{Z^m_{1:N},U}$. To obtain bounds on the generalization gap (60), we resort to Lemma A.1 with $X = W$, $Y = Z^m$ and $f(X,Y) = L_{Z^m}(W)$, so that $\mathbb{E}_{P_{X,Y}}[f(X,Y)] = \mathbb{E}_{P_{W,Z^m|k}}[L_{Z^m}(W)]$ and $\mathbb{E}_{P_X P_Y}[f(\widetilde{X},\widetilde{Y})] = \mathbb{E}_{P_{W,Z^m|k}}[L_{P_{Z|k}}(W)]$, where $P_{W,Z^m|k} = P_{W|Z^m}P_{Z^m|k}$ and using the first assumption in Assumption 5.3. Combining the resulting bound with the bounds on expected environment-level uncertainty, and plugging in (53) and (55) yield the two bounds of (33) respectively.

For Theorem 5.7, the proof follows along the same line. The ITMI bound on the expected environment-level uncertainty can be obtained along the lines of (59), using the second assumption in Assumption 5.4. To bound the average within-task uncertainty, we write

$$\Delta L(P_{Z|k}, P_{W|Z^m}) = \frac{1}{m}\sum_{j=1}^{m}\biggl(\mathbb{E}_{P_{W|k}P_{Z|k}}[l(W,Z)] - \mathbb{E}_{P_{W,Z_j|k}}[l(W,Z_j)]\biggr), \tag{61}$$

where W and Z_j in the second term are jointly distributed according to $P_{W,Z_j|k}$, which is the marginal of the joint distribution $P_{W,Z^m|k}$. In contrast, W and Z in the first term are conditionally independent random variables distributed as $(W,Z) \sim P_{W|k}P_{Z|k}$ where $P_{W|k}$ is the marginal distribution of $P_{W,Z_j|k}$. Consequently, we apply Lemma A.2 to each of the terms in the summation, with $X = W$, $Y = Z_j$ and $f(X,Y) = l(W,Z_j)$ together with the first assumption in Assumption 5.4. Combining the resulting bound with the ITMI bound on environment-level gap, and plugging in (53) and (55) yield the two bounds in (35) respectively.

Appendix D
Details of Example

We first give details of the derivation of meta-generalization gap for the case with separate within-task training and test sets. The average meta-generalization loss can be computed as

$$\mathbb{E}_{P_{Z^m_{1:N},U}}[\mathrm{L}^{\mathrm{sep}}_{P_{K,Z^{\mathrm{m_{tr}}}}}(U)] =$$

$$\mathbb{E}_{P_{Z^m_{1:N},U}}\left[(1-\alpha)^2U^2 + \mathbb{E}_{P_{K,Z^{\mathrm{m_{tr}}}}}\left[\alpha^2(D^{\mathrm{m_{tr}}}_K)^2 + \mathbb{E}_{P_{Z|K}}[Z^2] - 2\alpha D^{\mathrm{m_{tr}}}_k\mu_K + 2(1-\alpha)U(\alpha D^{\mathrm{m_{tr}}}_K - \mu_K)\right]\right]$$

$$\overset{(a)}{=} \mathbb{E}_{P_{Z^m_{1:N},U}}\left[(1-\alpha)^2\big(U^2 - 2U\mathbb{E}_{P_K}[\mu_K]\big)\right] + \mathbb{E}_{P_K}\left[\alpha^2\left(\mu_K^2 + \frac{\mu_K\bar{\mu}_K}{m_{\mathrm{tr}}}\right) + \mu_K - 2\alpha\mu_K^2\right], \tag{62}$$

where the equality in (a) follows since $\mathbb{E}_{P_{Z|K}}[Z^2] = \mu_K$, $\mathbb{E}_{P_{Z^{\mathrm{m_{tr}}}|K}}[D^{\mathrm{m_{tr}}}_K] = \mu_K$ and $\mathbb{E}_{P_{Z^{\mathrm{m_{tr}}}|K}}[(D^{\mathrm{m_{tr}}}_K)^2] = \mu_K^2 + \mu_K\bar{\mu}_K/m_{\mathrm{tr}}$. In a similar manner, the average meta-training loss can be computed as

$$\mathbb{E}_{P_{Z^m_{1:N},U}}[\mathrm{L}^{\mathrm{sep}}_{Z^m_{1:N}}(U)] = \mathbb{E}_{P_{Z^m_{1:N}}}\Bigg[-(1-\alpha)^2U^2 + \frac{1}{N}\sum_{i=1}^{N}\alpha^2(D^{\mathrm{m_{tr}}}_i)^2$$

$$+ \frac{1}{N}\sum_{i=1}^{N}\frac{1}{m_{\mathrm{te}}}\sum_{j=1}^{m_{\mathrm{te}}}(Z^{m_{\mathrm{te}}}_{i,j})^2 - 2\alpha\frac{1}{N}\sum_{i=1}^{N}D^{\mathrm{m_{tr}}}_iD^{\mathrm{m_{te}}}_i\Bigg], \tag{63}$$

with U defined as in (39). The meta-generalization gap in (41) then results by taking the difference of (62) and (63), and using that $\mathbb{E}_{P_{Z^m_{1:N}}}\left[(1-\alpha)^2U^2\right] = (1-\alpha)^2\big(\frac{1}{N}\mathbb{E}_{P_K}[\mu_K^2] + \big(1-\frac{1}{N}\big)(\mathbb{E}_{P_K}[\mu_K])^2\big) + \mathbb{E}_{P_K}[\mu_K\bar{\mu}_K]\big(\frac{1}{Nm_{\mathrm{te}}} + \frac{\alpha^2}{Nm_{\mathrm{tr}}}\big)$.

We now evaluate the mutual informations $I(U;Z^m_{1:N})$ and $I(U;Z^m_i)$. For the first MI, note that since the meta-learner is deterministic (see (39)), $H(U|Z^m_{1:N}) = 0$ and thus $I(U;Z^m_{1:N}) = H(U)$. For the second MI, we can write $I(U;Z^m_i) = H(U) - \mathbb{E}_{Z^m_i}[H(U|Z^m_i = z^m)]$. It can be seen that random variables U and $U|Z^m_i = z^m$ are mixtures of probability distributions, whose entropies can be evaluated following standard methods [47].

For the case with joint within-task training and test sets, the meta-generalization gap can be obtained in a similar way as $\Delta\mathrm{L}^{\mathrm{joint}}(P_{K,Z^m}, P_{U|Z^m_{1:N}}, P_{W|Z^m,U}) =$

$$\frac{2}{N}(1-\alpha)^2\bigg(\mathbb{E}_{P_K}\left[\mu_K^2\right] - (\mathbb{E}_{P_K}[\mu_K])^2\bigg) + 2\mathbb{E}_{P_K}[\mu_K\bar{\mu}_K]\bigg(\frac{\alpha}{m} + \frac{(1-\alpha)^2}{Nm}\bigg). \tag{64}$$

For the MI and ITMI-based bounds, note that with $\mathcal{W} = [0,1]$, the loss function $l(\cdot,\cdot)$ is $[0,1]$-bounded almost surely, and for the deterministic base-learner in (37) with $\mathcal{U} = [0,1]$, the average training loss $L^{\mathrm{joint}}_{Z^m}(\cdot)$ is also $[0,1]$-bounded almost surely for all $Z^m \in \mathcal{Z}^m$. Thus, Assumption 5.3 and Assumption 5.4 hold. The MI-based bound in (34) can be evaluated as

$$|\Delta\mathrm{L}^{\mathrm{joint}}(P_{K,Z^m}, P_{U|Z^m_{1:N}}, P_{W|Z^m,U})| \leq \sqrt{\frac{1}{2N}H(U)} + \sup_{k\in\mathcal{K}}\sqrt{\frac{1}{2m}I(W;Z^m|k)}. \tag{65}$$

For the ITMI bound (36), we similarly have

$$|\Delta\mathrm{L}^{\mathrm{joint}}(P_{K,Z^m}, P_{U|Z^m_{1:N}}, P_{W|Z^m,U})| \leq \frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{1}{2}I(U;Z^m_{1:N})} + \sup_{k\in\mathcal{K}}\frac{1}{m}\sum_{j=1}^{m}\sqrt{\frac{1}{2}I(W;Z_j|k)}. \tag{66}$$

All information measures can be easily evaluated numerically [47].

Appendix E
Proof of Lemma 6.1

From the update rule of the meta-learner in (45), we get the Markov dependency

$$P_{U^t|U^{(t-1)},\{W_{K_j}\}_{j=1}^t,\{Z^m_{K_j}\}_{j=1}^t,Z^m_{1:N}} = P_{U^t|U^{t-1},W_{K_t},Z^{\mathrm{m_{te}}}_{K_t}}, \tag{67}$$

where $U^{(t-1)} = \{U^1, \dots, U^{t-1}\}$ is the history vector of hyperparameters. The sampling strategy in (46) together with (67) then implies the following relation

$$P_{U^t|U^{(t-1)},\{W_{K_j}\}_{j=1}^t,\{Z^m_{K_j}\}_{j=1}^T,Z^m_{1:N}} = P_{U^t|U^{t-1},W_{K_t},Z^{\mathrm{m_{te}}}_{K_t}}. \tag{68}$$

Using $U^{(T)} = \{U^1, \dots, U^T\}$ to denote the set of all updates, we have the following relations

$$\begin{align} I(U;Z^m_{1:N}) &\overset{(a)}{\leq} I(U^{(T)};Z^m_{1:N}) \nonumber \\ &\overset{(b)}{\leq} I(U^{(T)};\{Z^m_{K_j}\}_{j=1}^T) = \sum_{t=1}^T I(U^t;\{Z^m_{K_j}\}_{j=1}^T|U^{(t-1)}) \tag{69} \\ &\leq \sum_{t=1}^T I(U^t;\{Z^m_{K_j}\}_{j=1}^T,\{W_{K_j}\}_{j=1}^t|U^{(t-1)}) \tag{70} \\ &= \sum_{t=1}^T h(U^t|U^{(t-1)}) - h\biggl(U^t|U^{(t-1)},\{Z^m_{K_j}\}_{j=1}^T,\{W_{K_j}\}_{j=1}^t\biggr) \tag{71} \\ &\overset{(c)}{=} \sum_{t=1}^T \biggl[h(U^t|U^{t-1}) - h(U^t|U^{t-1},W_{K_t},Z^{\mathrm{m_{te}}}_{K_t})\biggr], \tag{72} \end{align}$$

where, the inequality in (a) follows from data processing inequality on Markov chain $Z^m_{1:N} \rightarrow U^{(T)} \rightarrow U$; (b) follows from the Markov chain $Z^m_{1:N} \rightarrow \{Z^m_{K_j}\}_{j=1}^T \rightarrow U^{(T)}$; and the equality in (c) follows from $U^{(t-2)} \rightarrow U^{t-1} \rightarrow U^t$ and (68). Finally, the computation of bound in (72) follows similar to Lemma 5 in [23].

References

[1] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] O. Simeone, "A Brief Introduction to Machine Learning for Engineers," *Foundations and Trends® in Signal Processing*, vol. 12, no. 3-4, pp. 200–431, 2018.

[4] 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.

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

[6] S. Thrun, “Is Learning the N-th Thing Any Easier than Learning the First?” in *Proc. of Adv. in Neural Inf. Processing Sys. (NIPS)*, Dec. 1996, pp. 640–646.

[7] R. Vilalta and Y. Drissi, “A Perspective View and Survey of Meta-Learning,” *Artificial Intelligence Review*, vol. 18, no. 2, pp. 77–95, June 2002.

[8] O. Simeone, S. Park, and J. Kang, “From Learning to Meta-Learning: Reduced Training Overhead and Complexity for Communication Systems,” *arXiv preprint arXiv:2001.01227*, 2020.

[9] J. Baxter, “A Model of Inductive Bias Learning,” *Journal of Artificial Intelligence Research*, vol. 12, pp. 149–198, March 2000.

[10] M. Yin, G. Tucker, M. Zhou, S. Levine, and C. Finn, “Meta-Learning Without Memorization,” *arXiv preprint arXiv:1912.03820*, 2019.

[11] A. Maurer, “Algorithmic Stability and Meta-Learning,” *Journal of Machine Learning Research*, vol. 6, pp. 967–994, Jun 2005.

[12] A. Pentina and C. Lampert, “A PAC-Bayesian Bound for Lifelong Learning,” in *Proc. of Int. Conf. on Machine Learning (ICML)*, June 2014, pp. 991–999.

[13] J. Knoblauch, J. Jewson, and T. Damoulas, “Generalized Variational Inference,” *arXiv preprint arXiv:1904.02063*, 2019.

[14] L. Devroye and T. Wagner, “Distribution-Free Performance Bounds for Potential Function Rules,” *IEEE Trans. Inf. Theory*, vol. 25, no. 5, pp. 601–604, Sept. 1979.

[15] W. H. Rogers and T. J. Wagner, “A Finite Sample Distribution-Free Performance Bound for Local Discrimination Rules,” *The Annals of Statistics*, pp. 506–514, Mar 1978.

[16] R. Amit and R. Meir, “Meta-Learning by Adjusting Priors Based on Extended PAC-Bayes Theory,” in *Proc. of Int. Conf. Machine Learning (ICML)*, Jul 2018, pp. 205–214.

[17] J. Rothfuss, V. Fortuin, and A. Krause, “PACOH: Bayes-Optimal Meta-Learning with PAC-Guarantees,” *arXiv preprint arXiv:2002.05551*, 2020.

[18] D. Russo and J. Zou, “Controlling Bias in Adaptive Data Analysis Using Information Theory,” in *Proc. of Artificial Intelligence and Statistics (AISTATS)*, May 2016, pp. 1232–1240.

[19] A. Xu and M. Raginsky, “Information-Theoretic Analysis of Generalization Capability of Learning Algorithms,” in *Proc. of Adv. in Neural Inf. Processing Sys. (NIPS)*, Dec. 2017, pp. 2524–2533.

[20] C. Finn, P. Abbeel, and S. Levine, “Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks,” in *Proc. of Int. Conf. Machine Learning-Volume 70*, Aug. 2017, pp. 1126–1135.

[21] A. Nichol, J. Achiam, and J. Schulman, “On First-Order Meta-Learning Algorithms,” *arXiv preprint arXiv:1803.02999*, 2018.

[22] Y. Bu, S. Zou, and V. V. Veeravalli, “Tightening Mutual Information Based Bounds on Generalization Error,” in *Proc. of IEEE Int. Symp. Inf. Theory (ISIT)*, July 2019, pp. 587–591.

[23] A. Pensia, V. Jog, and P.-L. Loh, “Generalization Error Bounds for Noisy, Iterative Algorithms,” in *Proc. of IEEE Int. Symp. Inf. Theory (ISIT)*, June 2018, pp. 546–550.

[24] V. N. Vapnik and A. Y. Chervonenkis, “On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities,” in *Theory of Probability and its Applications*. SIAM, May 1971, vol. 16, no. 2, pp. 264–280.

[25] V. Koltchinskii and D. Panchenko, “Rademacher Processes and Bounding the Risk of Function Learning,” in *High Dimensional Probability II*. Springer, 2000, vol. 47, pp. 443–457.

[26] O. Bousquet and A. Elisseeff, “Stability and Generalization,” *Journal of Machine Learning Research*, vol. 2, pp. 499–526, Mar 2002.

[27] M. Kearns and D. Ron, “Algorithmic Stability and Sanity-Check Bounds for Leave-One-Out Cross-Validation,” *Neural Computation*, vol. 11, no. 6, pp. 1427–1453, Aug 1999.

[28] T. Poggio, R. Rifkin, S. Mukherjee, and P. Niyogi, “General Conditions for Predictivity In Learning Theory,” *Nature*, vol. 428, no. 6981, pp. 419–422, Mar 2004.

[29] S. Kutin and P. Niyogi, “Almost-Everywhere Algorithmic Stability and Generalization Error,” in *Proc. of Uncertainty in Artificial Intelligence (UAI)*, Dec 2002, pp. 275–282.

[30] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. L. Roth, “Preserving Statistical Validity in Adaptive Data Analysis,” in *Proc. of ACM Symp. Theory of Computing (STOC)*, June 2015, pp. 117–126.

[31] R. Bassily, K. Nissim, A. Smith, T. Steinke, U. Stemmer, and J. Ullman, “Algorithmic Stability for Adaptive Data Analysis,” in *Proc. of ACM Symp. Theory of Computing (STOC)*, June 2016, pp. 1046–1059.

[32] D. A. McAllester, “PAC-Bayesian Model Averaging,” in *Proc. of Annual Conf. Computational Learning Theory (COLT)*, July 1999, pp. 164–170.

[33] M. Seeger, “PAC-Bayesian Generalization Error Bounds for Gaussian Process Classification,” *Journal of Machine Learning Research*, vol. 3, pp. 233–269, Oct 2002.

[34] P. Alquier, J. Ridgway, and N. Chopin, “On the Properties of Variational Approximations of Gibbs Posteriors,” *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 8374–8414, Dec 2016.

[35] A. Asadi, E. Abbe, and S. Verdú, “Chaining Mutual Information and Tightening Generalization Bounds,” in *Proc. of Adv. in Neural Inf. Processing Sys. (NIPS)*, Dec 2018, pp. 7234–7243.

[36] 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. of Adv. Neural Inf. Processing Sys. (NIPS)*, Dec 2019, pp. 11 013–11 023.

[37] I. Alabdulmohsin, “Towards a Unified Theory of Learning and Information,” *Entropy*, vol. 22, no. 4, p. 438, April 2020.

[38] J. Jiao, Y. Han, and T. Weissman, “Dependence Measures Bounding the Exploration Bias for General Measurements,” in *Proc. of IEEE Int. Symp. Inf. Theory (ISIT)*, May 2017, pp. 1475–1479.

[39] I. Issa and M. Gastpar, “Computable Bounds On the Exploration Bias,” in *Proc. of IEEE Int. Symp. Inf. Theory (ISIT)*. IEEE, June 2018, pp. 576–580.

[40] I. Issa, A. R. Esposito, and M. Gastpar, “Strengthened Information-Theoretic Bounds on the Generalization Error,” in *Proc. of IEEE Int. Symp. Inf. Theory (ISIT)*, July 2019, pp. 582–586.

[41] T. Steinke and L. Zakynthinou, “Reasoning About Generalization via Conditional Mutual Information,” *arXiv preprint arXiv:2001.09122*, 2020.

[42] P. G. Bissiri, C. C. Holmes, and S. G. Walker, “A General Framework for Updating Belief Distributions,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 78, no. 5, pp. 1103–1130, Nov 2016.

[43] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019, vol. 48.

[44] M. Raginsky, A. Rakhlin, M. Tsao, Y. Wu, and A. Xu, “Information-Theoretic Analysis of Stability and Bias of Learning Algorithms,” in *Proc. of IEEE Inf. Theory Workshop (ITW)*, Sep. 2016, pp. 26–30.

[45] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, “Learnability, Stability and Uniform Convergence,” *Journal of Machine Learning Research*, vol. 11, pp. 2635–2670, Oct 2010.

[46] M. Welling and Y. W. Teh, “Bayesian Learning via Stochastic Gradient Langevin Dynamics,” in *Proc. Int. Conf. Machine Learning (ICML)*, June 2011, pp. 681–688.

[47] J. V. Michalowicz, J. M. Nichols, and F. Bucholtz, *Handbook of Differential Entropy*. CRC Press, 2013.