DOI:
10.1109/TIT.2021.3119605

Document Version
Peer reviewed version

Transfer Meta-Learning: Information-Theoretic Bounds and Information Meta-Risk Minimization

Sharu Theresa Jose, Osvaldo Simeone, Giuseppe Durisi

Abstract

Meta-learning automatically infers an *inductive bias* by observing data from a number of related tasks. The inductive bias is encoded by hyperparameters that determine aspects of the model class or training algorithm, such as initialization or learning rate. Meta-learning assumes that the learning tasks belong to a *task environment*, and that tasks are drawn from the same task environment both during meta-training and meta-testing. This, however, may not hold true in practice. In this paper, we introduce the problem of *transfer meta-learning*, in which tasks are drawn from a *target task environment* during meta-testing that may differ from the *source task environment* observed during meta-training. Novel information-theoretic upper bounds are obtained on the *transfer meta-generalization gap*, which measures the difference between the meta-training loss, available at the meta-learner, and the average loss on meta-test data from a new, randomly selected, task in the target task environment. The first bound, on the average transfer meta-generalization gap, captures the *meta-environment shift* between source and target task environments via the KL divergence between source and target data distributions. The second, PAC-Bayesian bound, and the third, single-draw bound, account for this shift via the log-likelihood ratio between source and target task distributions. Furthermore, two transfer meta-learning solutions are introduced. For the first, termed *Empirical Meta-Risk Minimization* (EMRM), we derive bounds on the average optimality gap. The second, referred to as *Information Meta-Risk Minimization* (IMRM), is obtained by minimizing the PAC-Bayesian bound. IMRM is shown via experiments to potentially outperform EMRM.

S. T. Jose and O. Simeone are with King's Communications, Learning, and Information Processing (KCLIP) lab at the Department of Engineering of King's College London, UK (emails: sharu.jose@kcl.ac.uk, osvaldo.simeone@kcl.ac.uk). They have received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 725731). G. Durisi is with the Department of Electrical Engineering, Chalmers Institute of Technology, Sweden (email: durisi@chalmers.se).

Index Terms

Transfer meta-learning, information-theoretic generalization bounds, PAC-Bayesian bounds, single-draw bounds, information risk minimization

I. Introduction

Any machine learning algorithm makes assumptions on the task of interest, which are collectively referred to as the inductive bias. In parametric machine learning, the inductive bias is encoded in the choice of a model class and of a training algorithm used to identify a model parameter vector based on training data. The inductive bias is fixed a priori, ideally with the help of domain expertise, and it can be refined via validation. As a typical example, an inductive bias may consist of a class of neural networks parameterized by synaptic weights and of an optimization procedure such as stochastic gradient descent (SGD). Hyperparameters including number of layers and SGD learning rate schedule can be selected by optimizing the validation error on an held-out data set.

Meta-learning or *learning to learn* aims to automatically infer some aspects of the inductive bias based on the observation of data from related tasks [1]–[3]. For example, the choice of an inductive bias—model class and training algorithm—for the problem of classifying images of animals may be based on labelled images of vehicles or faces. As formalized in [4], meta-learning assumes the presence of a *task environment* consisting of related learning tasks. A task environment is defined by a distribution on the set of tasks and by per-task data distributions. A meta-learner observes data sets from a finite number of tasks drawn from the task environment to infer the inductive bias, while its performance is evaluated on a new, previously unseen, task drawn from the same task environment.

As discussed, a key assumption in the standard formulation of meta-learning is that the tasks encountered during meta-learning are from the same task environment that generates the new "meta-test" task on which the performance of the hyperparameter is evaluated. This assumption may not be realistic in some applications [5]. For example, a personalized health application may be meta-trained by using data from a population of users that is not fully representative of the distribution of the health profiles expected in a different population on which the application is deployed and meta-tested. In this paper, we introduce the problem of *transfer meta-learning*, wherein the performance of a meta-learner that uses data sets drawn from a *source task environment* is tested on a new task drawn from a generally different *target task*

environment. In the proposed formulation, highly popular, or more frequently observed, tasks during meta-training may have a small probability in the target task environment, while other tasks may have a higher chance of being encountered.

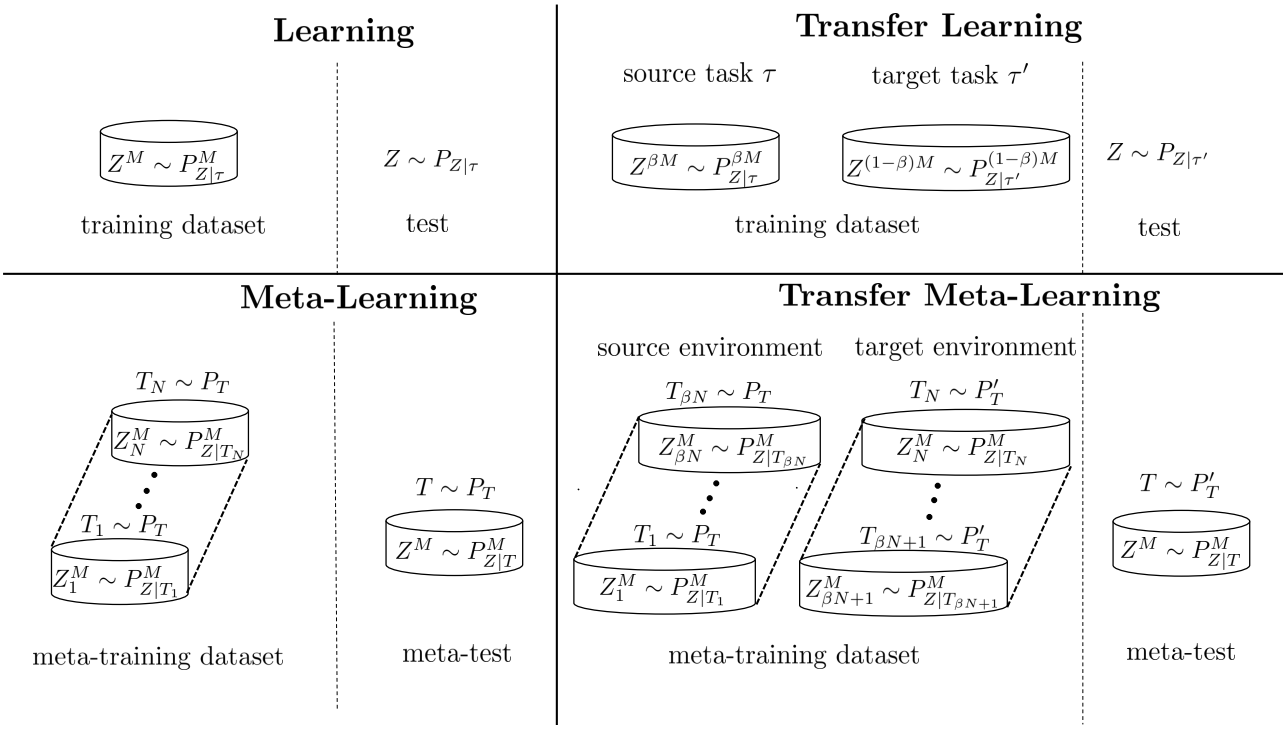


Fig. 1: Illustration of conventional learning, transfer learning, conventional meta-learning and transfer meta-learning with $P_{Z|\tau}$ denoting the distribution $P_{Z|T=\tau}$.

As illustrated in Figure 1, we consider a general formulation of transfer meta-learning where the meta-learner observes a meta-training set of N datasets $Z_1^M, \ldots, Z_N^M$, each of M samples, of which βN, with $\beta \in (0, 1]$, datasets correspond to tasks drawn from the *source task environment* and $(1-\beta)N$ datasets correspond to tasks from the *target task environment*. Under source and target task environments, tasks are drawn according to distinct distributions P_T and P'_T, respectively. Based on the meta-training set $Z_{1:N}^M = (Z_1^M, \ldots, Z_N^M)$, the meta-learner infers the vector of hyper-parameters $u \in \mathcal{U}$. The hyperparameters u determine the base learning algorithm through a conditional distribution $P_{W|Z^M,U=u}$, that maps a training set Z^M to a model parameter W given u. The performance of the inferred hyperparameter u is evaluated in terms of the *transfer meta-generalization loss* $\mathcal{L}'_g(u)$, which is the expected loss over a data set $Z^M \sim P^M_{Z|T}$ sampled from a task T randomly selected from the target task distribution P'_T. The subscript g of $\mathcal{L}'_g(u)$ indicates that the considered loss is the generalization loss and the superscript $'$

indicates that the generalization loss is evaluated with respect to the target task distribution P'_T. This objective function is not available at the meta-learner since the target task distribution P'_T and the per-task distributions $P_{Z|T=\tau}$ for every task τ are unknown. Instead, the meta-learner can evaluate the empirical performance of the inferred hyperparameter on the meta-training set $Z_{1:N}^M$ in terms of the *meta-training loss* $\mathcal{L}_t(u|Z_{1:N}^M)$. The subscript t of $\mathcal{L}_t(u|Z_{1:N}^M)$ indicates that the loss considered is training loss.

The difference between the transfer meta-generalization loss and the meta-training loss, referred to as the *transfer meta-generalization gap* $\Delta\mathcal{L}'(u|Z_{1:N}^M)$, is a key metric to evaluate the generalization performance of the meta-learner. If the transfer meta-generalization gap is small, on average or with high probability, the meta-learner can take the performance on the meta-training set as a reliable measure of accuracy of the inferred hyperparameter in terms of the transfer meta-generalization loss. In this paper, we first study information-theoretic upper bounds on the transfer meta-generalization gap of three different flavours – bounds on the average transfer meta-generalization gap, high-probability probably-approximately-correct (PAC)-Bayesian bounds, and high-probability single-draw bounds– and, we introduce two transfer meta-learning algorithms based on *Empirical Meta-Risk Minimization* (EMRM) and *Information Meta-Risk Minimization* (IMRM).

The transfer meta-learning setting considered in this paper generalizes conventional transfer learning [6]–[8], as well as meta-learning (see Figure 1). Specifically, when the source and target task distributions are delta functions centered at source domain task τ and target domain task τ' respectively, with $\tau \neq \tau'$, and the hyperparameter u to be inferred coincides with the model parameter, the transfer meta-learning setting reduces to transfer learning. While there exists a rich literature on generalization bounds and algorithms for transfer learning, this work is, to the best of our knowledge, the first one to extend the notion of transfer to meta-learning, to derive information-theoretic upper bounds on the transfer meta-generalization gap, and to propose transfer meta-learning design criteria.

A. Related Work

Three distinct kinds of bounds on generalization gap, i.e., the difference between training and generalization losses, have been studied in literature for conventional learning—bounds on average generalization gap, high-probability PAC-Bayesian bounds and high-probability single-draw bounds [9]. For learning algorithms described as a stochastic mapping from the input

training set to the model parameter, the average generalization gap evaluates the average difference between the training and generalization losses over the learning algorithm and its input training set. Information-theoretic upper bounds on the average generalization gap have been studied first by Russo *et al.* [10] and Xu *et al.* [11], and variants of the bounds have been investigated in [12]–[14]. Of particular relevance to our work is the individual sample mutual information (ISMI) based bound [12], which captures the sensitivity of the learning algorithm to the input training set, and thus the generalization ability, via the mutual information (MI) between the model parameter output of the algorithm and individual data sample of the input training set. These bounds have the distinction that they depend explicitly on the data distribution, the learning algorithm, and the loss function. Moreover, for deterministic algorithms, the ISMI approach yields a finite upper bound as compared to the MI bounds in [11]. The ISMI bound has been extended to obtain bounds on generalization gap for transfer learning in [15] and for meta-learning in [16], where, in the latter, the MI between the hyperparameter and per-task data of the meta-training set captures the sensitivity of the meta-learner to the meta-training data set. The results in this paper can be seen as a natural extension of these lines of work to transfer meta-learning.

Apart from bounds on average generalization gap, PAC bound on the generalization gap which holds with high probability over the training set have been studied in the literature. Classical PAC bounds for conventional learning assume deterministic learners and employ measures of complexity of the model class like Vapnik-Chervonenkis (VC) dimension [17] or Radmacher complexity [18] to characterize the generalization gap. For stochastic learning algorithms, McAllester [19] developed a PAC-Bayesian upper bound on the average of the generalization gap over the learning algorithm, which holds with probability at least $1-\delta$, with $\delta \in (0,1)$, over the input training set. These bounds employ a reference data-independent 'prior' distribution on the model parameter space, and the sensitivity of the learning algorithm to the training set is captured by the Kullback-Leibler (KL) divergence between the posterior distribution of the learning algorithm and the prior. As such, the PAC-Bayesian bounds are independent of data distributions. We note that the recent line of work in [20] suggests tightening the PAC-Bayesian bounds by choosing a data-dependent prior distribution evaluated on an heldout data set, which is not part of the training data.

Various refinements of PAC-Bayesian bounds have been studied for conventional learning [21]–[24], and for meta-learning [25]–[27] where for the latter PAC-Bayesian bounds employ a hyper-prior distribution on the space of hyperparameters in addition to the prior. A PAC-

Bayesian approach to domain adaptation specialized to linear classifiers has been considered in [28]. Furthermore, PAC-Bayesian bounds can be employed to design learning algorithms that ensure generalization through the principle of *Information Risk Minimization* (IRM) [29]. For conventional learning, the IRM principle finds a randomized learning algorithm that minimizes the PAC-Bayesian upper bound on the generalization loss, which is given by the empirical training loss regularized by the KL divergence between the posterior learning algorithm and the prior. In Section IV-B, we resort to the IRM principle and propose a novel learning algorithm for transfer meta-learning.

PAC-Bayesian bounds apply to the scenario when a model parameter is drawn every time the learning algorithm is used, and the performance of the learner is evaluated with respect to the average of the generalization gap over these draws. In contrast, high-probability *single-draw bounds* are relevant in scenarios when a model parameter is drawn only once from the stochastic learner, and the goal is to evaluate the generalization performance with respect to this parameter. Precisely, single-draw probability bounds yield upper bounds on the generalization gap which holds with probability at least $1-\delta$, with $\delta \in (0,1)$, over the training set and the model parameter. For conventional learning, MI-based single-draw bounds have been obtained in [30], [31], while information-theoretic quantities like Rényi divergence, α-mutual information, and information leakage have been used in [32]. To the best of our knowledge, single-draw bounds have not been studied in the context of meta-learning or transfer meta-learning before.

In comparison to the generalization bounds for conventional learning, the generalization bounds for transfer learning have to account for the *domain shift* between source domain and target domain. For conventional transfer learning, upper bound on the generalization loss on target domain is obtained in terms of generalization loss on the source domain, together with a divergence measure that captures the domain shift [6], [33], [34]. Various distance and divergence measures have been explored in the literature to quantify the domain shift. These measures have the advantage that they can be empirically estimated from finite data sets from source and target domains. For example, [6] studies transfer learning for classification tasks and obtains high-probability upper bounds on the target domain generalization loss based on the $\mathscr{H}$-divergence, or $d_{\mathcal{A}}$-distance, in terms of VC dimensions or Radmacher complexity [33]. The $d_{\mathcal{A}}$ distance has been generalized to the discrepancy distance so as to account for loss functions beyond the detection loss in [34], and to integral probability metric in [35]. Estimates of these distance measures yield generalization bounds in terms of Radmacher complexity. The $\mathscr{H}$-divergence has

been further extended to define the $\mathscr{H}\Delta\mathscr{H}$ divergence in [36]. While these distance measures are tailored to given loss functions and model class, general statistical divergence measures, such as Rényi divergence and Wasserstein distance have been considered in [37]–[39] and [40] respectively. The information-theoretic generalization bound in [15] captures the domain shift in terms of the KL divergence between source and target domain. Our work draws inspiration from this line of research.

B. Main Contributions

Building on the lines of work on transfer learning outlined above, we introduce the problem of transfer meta-learning, in which data from both source and target task environments are available for meta-training. Extending the methods in [33], [35], [36] for transfer learning, we measure the meta-training loss as a weighted average of the training losses on source and target task environment data sets. This weighted average includes as special cases methods that use only data from source or target task environments. We refer to the resulting design criterion as EMRM. We derive information-theoretic upper bounds on the average transfer meta-generalization gap, i.e., on the average difference between transfer meta-generalization loss $\mathcal{L}'_g(u)$ and meta-training loss $\mathcal{L}_t(u|Z_{1:N}^M)$. The bounds generalize prior works on transfer learning [15] and meta-learning [16]. We also present novel PAC-Bayesian and single-draw probability bounds. Central to the derivation of these generalization bounds is the information-density based exponential inequality approach of [9]. We detail the main contributions as follows.

1) We extend the individual task mutual information (ITMI) based approach of [16] for meta-learning to obtain novel upper bounds on the average transfer meta-generalization gap that holds for any meta-learner. The resulting bound captures the *meta-environment shift* from source to target task distributions via the KL divergence between source environment data distribution and target environment data distribution.
2) We specialize the obtained generalization bound on the average transfer meta-generalization gap to study the performance of the EMRM algorithm that minimizes the empirical average meta-training loss, and obtain a novel upper bound on the average transfer excess meta-risk for EMRM. The average transfer excess meta-risk is the optimality gap between the average transfer meta-generalization loss of EMRM and the optimal transfer meta-generalization loss.

3) We derive novel PAC-Bayesian bounds for transfer meta-learning that quantify the impact of the meta-environment shift through the log-likelihood ratio of the source and target task distributions. We use these bounds to introduce a novel meta-training algorithm, termed IMRM, based on the principle of information risk minimization [29].
4) We obtain new single-draw probability bounds for transfer meta-learning in terms of information densities and a log-likelihood ratio between source and target task distribution. Single-draw bounds captures the performance under a single realization of the hyperparameter drawn by a stochastic meta-learner. Furthermore, the resulting bounds can be specialized to obtain novel single-draw bounds for conventional meta-learning.
5) Finally, we compare the performance of EMRM and IMRM algorithms on a transfer meta-learning example, and show that IMRM can outperform EMRM in terms of transfer meta-generalization loss for sufficiently small number of tasks and per-task data samples. As the number of tasks and per-task data samples grow, IMRM reduces to EMRM.

C. Notation

Throughout this paper, we use upper case letters, e.g. X, to denote random variables and lower case letters, e.g. x to represent 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 $X = x$. 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 define the Kronecker delta $\delta(x - x_0) = 1$ if $x = x_0$ and $\delta(x - x_0) = 0$ otherwise, and use $\mathbb{I}_E$ to denote the indicator function which equals one when the event E is true and equals zero otherwise.

II. Problem Formulation

A. Conventional Transfer Learning

We review first the conventional transfer learning problem [33], [35], [36] in order to define important notation and provide the necessary background for the introduction of transfer meta-learning. We refer to Figure 1 for an illustration comparing conventional learning and transfer learning. In transfer learning, we are given a data set that consists of: (i) data points from a *source*

task τ drawn from an underlying *unknown* data distribution, $P_{Z|T=\tau} \in \mathcal{P}(\mathcal{Z})$, defined in a subset or vector space $\mathcal{Z}$; as well as (ii) data from a target task τ', drawn from a generally different distribution $P_{Z|T=\tau'} \in \mathcal{P}(\mathcal{Z})$. The goal is to infer a machine learning model that generalizes well on the data from the *target task* τ'. For notational convenience, in the following, we use $P_{Z|\tau}$ to denote source data distribution $P_{Z|T=\tau}$, and $P_{Z|\tau'}$ to denote the target data distribution $P_{Z|T=\tau'}$.

The learner has access to a training data set $Z^M = (Z_1, Z_2, \ldots, Z_M)$, which consists of βM, for some fixed $\beta \in (0, 1]$, independent and identically distributed (i.i.d.) samples $(Z_1, \ldots, Z_{\beta M}) \sim P_{Z|\tau}^{\beta M}$ drawn from the source data distribution $P_{Z|\tau}$, and $(1-\beta)M$ i.i.d. samples $(Z_{\beta M+1}, \ldots Z_M) \sim P_{Z|\tau'}^{(1-\beta)M}$ from the target data distribution $P_{Z|\tau'}$. Since the learner instead does not know the distributions $P_{Z|\tau}$ and $P_{Z|\tau'}$, it uses the data set Z^M to choose a model, or hypothesis, W from the model class $\mathcal{W}$ by using a *randomized* training procedure defined by a conditional distribution $P_{W|Z^M} \in \mathcal{P}(\mathcal{W})$ as $W \sim P_{W|Z^M}$. The conditional distribution $P_{W|Z^M}$ defines a stochastic mapping from the training data set Z^M to the model class $\mathcal{W}$.

The performance of a model parameter vector $w \in \mathcal{W}$ on a data sample $z \in \mathcal{Z}$ is measured by a loss function $l(w, z)$ where $l : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}_+$. The *generalization loss*, or *population loss*, for a model parameter vector $w \in \mathcal{W}$ is evaluated on the target task τ', and is defined as

$$L_g(w|\tau') = \mathbb{E}_{P_{Z|\tau'}}[l(w, Z)], \tag{1}$$

where the average is taken over a test example Z drawn independently of Z^M from the target task data distribution $P_{Z|\tau'}$.

The generalization loss cannot be computed by the learner, given that the data distribution $P_{Z|\tau'}$ is unknown. A typical solution is for the learner to minimize instead the *weighted average training loss* on the data set Z^M, which is defined as the empirical average

$$L_t(w|Z^M) = \frac{\alpha}{\beta M} \sum_{i=1}^{\beta M} l(w, Z_i) + \frac{1-\alpha}{(1-\beta)M} \sum_{i=\beta M+1}^{M} l(w, Z_i), \tag{2}$$

where $\alpha \in [0, 1]$ is a hyperparameter [36]. Note that this formulation assumes that the learner knows which training data comes from the source task and which are from the target task. We distinguish the generalization loss and the training loss via the subscripts g and t of $L_g(w|\tau')$ and $L_t(w|Z^M)$ respectively. The difference between generalization loss (1) and training loss (2), known as *transfer generalization gap*, is a key metric that relates to the performance of the learner. This is because a small transfer generalization gap ensures that the training loss (2) is a reliable estimate of the generalization loss (1). An information theoretic study of the transfer

generalization gap and of the excess risk gap of a learner that minimizes (2) was presented in [15].

B. Meta-Learning

We now review the meta-learning setting [41]. To start, let us fix a class of *within-task base learners* $P_{W|Z^M,U=u}$ mapping a data set Z^M to a model parameter vector W, where each base learner is identified by a hyperparameter $u \in \mathcal{U}$. Meta-learning aims to automatically infer the hyperparameter u using data from related tasks, thereby "learning to learn". Towards this goal, a *meta-learner* observes data from tasks drawn from a *task environment*. A task environment is defined by a *task distribution* P_T supported over the set of tasks $\mathcal{T}$, as well as by a per-task data distribution $P_{Z|T=\tau}$ for each $\tau \in \mathcal{T}$. Using the meta-training data drawn from a randomly selected subset of tasks, the meta-learner infers the hyperparameter $u \in \mathcal{U}$ with the goal of ensuring that the base learner $P_{W|Z^M,u}$ generalize well on a new, previously unobserved *meta-test task* $T \sim P_T$ drawn independently from the same task environment.

To elaborate, as seen in Figure 1, the meta-training data set consists of N data sets $Z^M_{1:N} = (Z^M_1, \ldots, Z^M_N)$, where each ith sub-data set Z^M_i is generated independently by first drawing a task $T_i \sim P_T$ and then generating a task specific data set $Z^M_i \sim P^M_{Z|T=T_i}$. The meta-learner does not know the distributions P_T and $\{P_{Z|T=\tau}\}_{\tau\in\mathcal{T}}$. We consider a *randomized* meta-learner [16]

$$U \sim P_{U|Z^M_{1:N}}, \tag{3}$$

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. As discussed, for a given hyperparameter $U = u$ and given a data set Z^M, the *within-task base learner* $P_{W|Z^M,u} \in \mathcal{P}(\mathcal{W})$ maps the per-task training subset Z^M to random model parameter $W \sim P_{W|Z^M,u}$. The average per-task test loss for a given task T is obtained as

$$L_g(u|T, Z^M) = \mathbb{E}_{P_{W|Z^M,u}}[L_g(W|T)], \tag{4}$$

where the per-task generalization loss $L_g(w|T)$ is defined in (1). The goal of meta-learning is to minimize the *meta-generalization loss* defined as

$$\mathcal{L}_g(u) = \mathbb{E}_{P_T P^M_{Z|T}}[L_g(u|T, Z^M)]. \tag{5}$$

The meta-generalization loss is averaged over new, meta-test tasks $T \sim P_T$ drawn from the task environment P_T and on the corresponding training data Z^M drawn i.i.d from the data distribution $P^M_{Z|T}$.

The meta-generalization loss cannot be computed by the meta-learner, given that the task distribution P_T and per-task data distribution $P_{Z|T}$ are unknown. The meta-learner relies instead on the *empirical meta-training loss*

$$\mathcal{L}_t(u|Z_{1:N}^M) = \frac{1}{N}\sum_{i=1}^{N} L_t(u|Z_i^M), \tag{6}$$

where $L_t(u|Z_i^M)$ is the average per-task training loss,

$$L_t(u|Z_i^M) = \mathbb{E}_{P_{W|Z_i^M,u}}[L_t(W|Z_i^M)], \tag{7}$$

with $L_t(w|Z^M)$ defined in (2) (with $\alpha = \beta = 1$). The difference between the meta-generalization loss (5) and meta-training loss (6) is known as the *meta-generalization gap*, and is a measure of performance of the meta-learner.

C. Transfer Meta-Learning

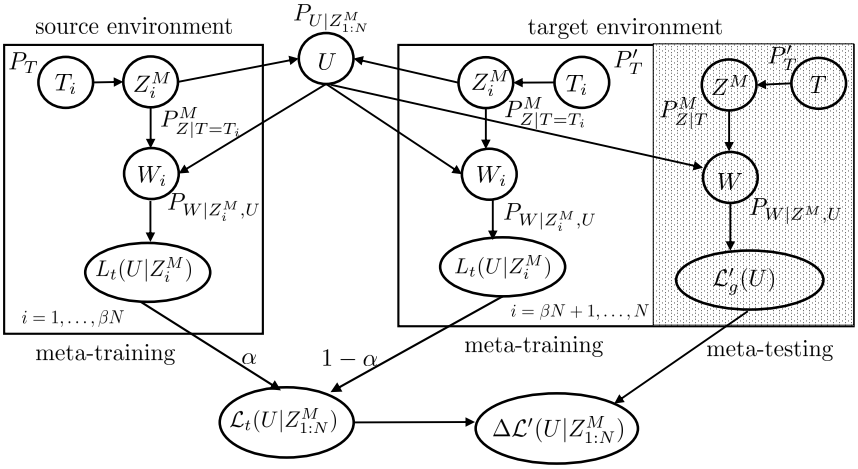


Fig. 2: A Bayesian network representation of the variables involved in the definition of transfer meta-learning.

In this section, we introduce the problem of transfer meta-learning. As we will explain, it generalizes both transfer learning and meta-learning. In transfer meta-learning, as seen in Figure 2, a meta-learner observes meta-training data from two different environments: (i) a *source task environment* which is defined by a *source task distribution* $P_T \in \mathcal{P}(\mathcal{T})$ and a per-task data distribution $P_{Z|T}$; and (ii) a *target task environment* which is defined by *target task distribution* $P'_T \in \mathcal{P}(\mathcal{T})$ and per-task data distribution $P_{Z|T}$. For a given family of per-task base learner $P_{W|Z^M,u}$, the goal of transfer meta-learning is to infer a hyperparameter $u \in \mathcal{U}$ from the

meta-training data such that the base learner $P_{W|Z^M,u}$ generalize well to a new task $T \sim P'_T$ drawn independently from the target task distribution P'_T.

The source and target task distributions P_T and P'_T model the likelihood of observing a given set of tasks during meta-training and meta-testing, respectively. Highly "popular", or more frequently observed, tasks in the source task environment may have a smaller chance of being observed, or they may even not appear, in the target task environment, while new tasks may only be encountered during meta-testing. For example, a personalized health application may be meta-trained by using data from a population of users that is not fully representative of the distribution of the health profiles expected in a different population on which the application is deployed and meta-tested.

The meta-training data set consists of N data sets $Z^M_{1:N} = (Z^M_1, \ldots, Z^M_N)$, where $(Z^M_1, \ldots, Z^M_{\beta N}) \triangleq Z^M_{1:\beta N}$, for some fixed $\beta \in (0, 1]$, constitutes the source environment data set, with each ith sub-data set Z^M_i being generated independently by first drawing a task $T_i \sim P_T$ from the source task distribution P_T and then a task-specific data set $Z^M_i \sim P^M_{Z|T=T_i}$. The sub-data sets $(Z^M_{\beta N+1}, \ldots, Z^M_N) \triangleq Z^M_{\beta N+1:N}$ belong to the target environment with each ith data set generated independently by first drawing a task $T_i \sim P'_T$ and then task specific data set $Z^M_i \sim P^M_{Z|T=T_i}$. All distributions P_T, P'_T and $\{P_{Z|T=\tau}\}_{\tau\in\mathcal{T}}$ are unknown to the meta-learner. Note that $\beta = 1$ corresponds to the extreme scenario in which only data from source task environment is available for meta-training.

Considering a randomized meta-learner $U \sim P_{U|Z^M_{1:N}} \in \mathcal{P}(\mathcal{U})$ as in (3), the goal of the meta-learner is to minimize the *transfer meta-generalization loss*

$$\mathcal{L}'_g(u) = \mathbb{E}_{P'_T P^M_{Z|T}}\big[L_g(u|Z^M, T)\big], \tag{8}$$

evaluated on a new meta-test task $T \sim P'_T$ drawn from the target task distribution P'_T and on the corresponding training data Z^M drawn i.i.d. from the data distribution $P_{Z|T}$.

In analogy with the weighted average training loss (2) used for transfer learning, we propose that the meta-learner aims at minimizing the *weighted average meta-training loss* on the meta-training set $Z^M_{1:N}$, which is defined as

$$\mathcal{L}_t(u|Z^M_{1:N}) = \frac{\alpha}{\beta N}\sum_{i=1}^{\beta N} L_t(u|Z^M_i) + \frac{1-\alpha}{(1-\beta)N}\sum_{i=\beta N+1}^{N} L_t(u|Z^M_i), \tag{9}$$

for some hyper-hyperparameter $\alpha \in [0, 1]$. We note that this formulation assumes that the meta-learner knows which data comes from the source task environment and which are from the target

task environment. We distinguish the transfer meta-generalization loss and the meta-training loss via the subscripts g, t of $\mathcal{L}'_g(u)$ and $\mathcal{L}_t(u|Z_{1:N}^M)$ respectively, with the superscript $'$ of $\mathcal{L}'_g(u)$ denoting that the generalization loss is evaluated with respect to the target task distribution P'_T. The meta-training loss (9) can be computed by the meta-learner based on the meta-training data $Z_{1:N}^M$ and it can be used as a criterion to select the hyperparameter u (for a fixed α). We refer to the meta-training algorithm that outputs the hyperparameter that minimizes (9) as Empirical Meta-Risk Minimization (EMRM). Note that EMRM is deterministic with $P_{U|Z_{1:N}^M} = \delta(U - U^{\mathrm{EMRM}}(Z_{1:N}^M))$ where

$$U^{\mathrm{EMRM}}(Z_{1:N}^M) = \arg\min_{u\in\mathcal{U}} \mathcal{L}_t(u|Z_{1:N}^M). \tag{10}$$

Here, and hence forth, we take arg min to output any one of the optimal solutions of the problem at hand and we assume that the set of optimal solutions is not empty. In the following sections, we also use loss functions with double subscript. For example, $\mathcal{L}'_{g,t}(u) = \mathbb{E}_{P'_T P_{Z|T}^M}[L_t(u|Z^M)]$, defined in (21), with subscripts g, t denote that it accounts for the generalization loss ('g') at the environment level (with average over $T \sim P'_T$ and Z^M), and the empirical training loss ('t') at the task level ($L_t(u|Z^M)$). We conclude this section with the following remark.

Remark II.1. The transfer meta-learning setting introduced here generalizes conventional learning, transfer learning and meta-learning:

1) When $\beta = 1$, only data from source task environment is available for meta-training. If, in addition, source and target task distributions are equal, i.e., if $P_T = P'_T$, we recover the conventional meta-generalization problem reviewed in Section II-B.
2) Consider now the special case where source and target task distributions are concentrated around two specific tasks τ and τ' respectively, that is, we have $P_T = \delta(T - \tau)$ and $P'_T = \delta(T - \tau')$ for some $\tau, \tau' \in \mathcal{T}$. With $N = 2$, the meta-training set $Z_{1:N}^M = (Z_\tau^{\beta NM}, Z_{\tau'}^{(1-\beta)NM})$ with $Z_\tau^{\beta NM} \sim P_{Z|T=\tau}^{\beta NM}$ and $Z_{\tau'}^{(1-\beta)NM} \sim P_{Z|T=\tau'}^{(1-\beta)NM}$ contains samples that are generated i.i.d. from the source data distribution $P_{Z|T=\tau}$ and target data distribution $P_{Z|T=\tau'}$. Assume that the base learner neglects data from the task to output always the hyperparameter U, i.e., $P_{W|Z^M,U} = \delta(W - U)$. Upon fixing $W = U$, we then have the meta-learner $P_{U|Z_{1:N}^M} = P_{W|Z_{1:N}^M}$. With these choices, the problem of transfer meta-learning reduces to the conventional transfer learning reviewed in Section II-A by mapping the transfer meta-generalization loss $\mathcal{L}'_g(u)$ to the generalization loss $L_g(w|\tau') = L_g(u|\tau')$ and the meta-training loss $\mathcal{L}_t(u|Z_{1:N}^M)$ to the training loss $L_t(w|Z^{NM}) = L_t(u|Z^{NM})$.

□

III. Information-Theoretic Analysis of Empirical Meta-Risk Minimization

In this section, we focus on the information-theoretic analysis of empirical meta-risk minimization (EMRM), which is defined by the optimization (10). To this end, we will first study bounds on the average transfer meta-generalization gap for *any* meta-learner $P_{U|Z_{1:N}^M}$, where the average is taken with respect to $P_{Z_{1:N}^M} P_{U|Z_{1:N}^M}$. Since our goal is to specialize the derived bound to a deterministic algorithm like EMRM, we obtain individual task based bounds [16], which yield non-vacuous bounds for deterministic mappings from the space of $Z_{1:N}^M$ to $\mathcal{U}$. We then apply the results to analyze the average transfer excess meta-risk for EMRM. We refer to Section IV-B for PAC-Bayesian bounds and Section V for single-draw bounds on transfer meta-generalization gap. We start with a formal definition of the performance criteria of interest.

The *transfer meta-generalization gap* is the difference between the transfer meta-generalization loss (8) and the meta-training loss (9). For any given hyperparameter $u \in \mathcal{U}$, it is defined as

$$\Delta\mathcal{L}'(u|Z_{1:N}^M) = \mathcal{L}'_g(u) - \mathcal{L}_t(u|Z_{1:N}^M). \tag{11}$$

For a general stochastic meta-learner $P_{U|Z_{1:N}^M}$, the *average transfer meta-generalization gap* is obtained as

$$\mathbb{E}_{P_{Z_{1:N}^M} P_{U|Z_{1:N}^M}}[\Delta\mathcal{L}'(U|Z_{1:N}^M)] \tag{12}$$

with the expectation taken over the meta-training data set $Z_{1:N}^M$ and hyperparameter $U \sim P_{U|Z_{1:N}^M}$. Note that $P_{Z_{1:N}^M}$ is the marginal of the product distribution $\prod_{i=1}^{\beta N} P_{T_i} P_{Z|T=T_i}^M \prod_{i=\beta N+1}^{N} P'_{T_i} P_{Z|T=T_i}^M$, as described in the previous section. The average transfer meta-generalization gap (12) quantifies how close the meta-training loss is to the transfer meta-generalization loss, which is the desired, but unknown, meta-learning criterion. If the transfer meta-generalization gap is sufficiently small, the meta-training loss can be taken as a reliable measure of the transfer meta-generalization loss. In this case, one can expect EMRM (10), which relies on the minimum of the weighted meta-training loss $\mathcal{L}_t(u|Z_{1:N}^M)$, to perform well.

The *average transfer excess meta-risk* evaluates the performance of a meta-training algorithm with respect to the optimal hyperparameter u^* that minimizes the transfer meta-generalization

loss (8). For a fixed class of base learners $P_{W|Z^M,u}$, the optimal hyperparameter minimizing (8) is given by

$$u^* = \arg\min_{u\in\mathcal{U}} \mathcal{L}'_g(u). \tag{13}$$

The *average transfer excess meta-risk of the EMRM algorithm* is hence computed as

$$\mathbb{E}_{P_{Z^M_{1:N}}}[\mathcal{L}'_g(U^{\mathrm{EMRM}}(Z^M_{1:N})) - \mathcal{L}'_g(u^*)]. \tag{14}$$

In the next subsection, we present the technical assumptions underlying the analysis, as well as some exponential inequalities that will play a central role in the derivations. In Section III-B, we obtain upper bounds on the average transfer meta-generalization gap (12) for any meta-learner, while Section III-C focuses on bounding the average transfer excess meta-risk (14) for EMRM.

A. Assumptions and Exponential Inequalities

We start by defining σ^2-sub-Gaussian random variables.

Definition 3.1: A random variable $X \sim P_X$ with finite mean, i.e., $\mathbb{E}_{P_X}[X] < \infty$, is said to be σ^2-sub-Gaussian if its moment generating function satisfies

$$\mathbb{E}_{P_X}[\exp(\lambda(X - \mathbb{E}_{P_X}[X]))] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right), \quad \text{for all } \lambda \in \mathbb{R}. \tag{15}$$

Moreover, if X_i, $i = 1, \ldots, n$ are independent σ^2-sub-Gaussian random variables, then the average $\sum_{i=1}^n X_i/n$ is σ^2/n-sub-Gaussian.

Throughout, we denote as P_U the marginal of the joint distribution $P_{Z^M_{1:N}} P_{U|Z^M_{1:N}}$ induced by the meta-learner. We also use P_{Z^M} to denote the marginal of the joint distribution $P_T P^M_{Z|T}$ of the data under the source environment and, in a similar manner, P'_{Z^M} to denote the marginal of the joint distribution $P'_T P^M_{Z|T}$ of the data under the target environment. In the rest of this section, we make the following assumptions on the loss function.

Assumption 3.1: The environment distributions P_T, P'_T, and $\{P_{Z|T=\tau}\}_{\tau\in\mathcal{T}}$, the base learner $P_{W|Z^M,U}$, and the meta-learner $P_{U|Z^M_{1:N}}$ satisfy the following assumptions.

- (a) For each task $\tau \in \mathcal{T}$, the loss function $l(W, Z)$ is δ^2_τ-sub-Gaussian when $(W, Z) \sim P_{W|T=\tau} P_{Z|T=\tau}$, where $P_{W|T=\tau}$ is the marginal of the model parameter trained for task τ, which is obtained by marginalizing the joint distribution $P_U P^M_{Z|T=\tau} P_{W|Z^M,U}$;
- (b) The per-task average training loss $L_t(U|Z^M)$ is σ^2-sub-Gaussian when $(U, Z^M) \sim P_U P'_{Z^M}$.

We note that the sub-Gaussianity properties in Assumption 3.1(a) and Assumption 3.1(b) are with respect to different distributions. As such, satisfying Assumption 3.1(a) does not guarantee

sub-Gaussianity in the sense of Assumption 3.1(b). However, if the loss function is bounded, i.e., $l(\cdot,\cdot) \in [a,b]$ for $0 \leq a \leq b < \infty$, it can be verified that both of these assumptions hold with $\delta_\tau^2 = (b-a)^2/4 = \sigma^2$ for any $\tau \in \mathcal{T}$.

Definition 3.2: The information density between two discrete or continuous random variables $(A,B) \sim P_{A,B}$ with well-defined joint probability mass or density function $P_{A,B}(a,b)$, and marginals $P_A(a)$ and $P_B(b)$ is the random variable

$$\imath(A,B) = \log \frac{P_{A,B}(A,B)}{P_A(A)P_B(B)} = \log \frac{P_{A|B}(A|B)}{P_A(A)}. \tag{16}$$

The information density quantifies the evidence for the hypothesis that A is produced from B via the stochastic mechanism $P_{A|B}$ rather than being drawn from the marginal P_A. The average of the information density is given by the mutual information (MI) $I(A;B) = \mathbb{E}_{P_{A,B}}[\imath(A,B)]$.

In the analysis, the information densities $\imath(U, Z_i^M)$ for $i = 1,\ldots,N$, and $\imath(W, Z_j|T=\tau)$ for $j = 1,\ldots,M$ will play a key role. The information density $\imath(U, Z_i^M)$ is defined for random variables $(U, Z_i^M) \sim P_{U,Z_i^M}$, where P_{U,Z_i^M} is obtained by marginalizing the joint distribution $P_{Z_{1:N}^M} P_{U|Z_{1:N}^M}$ over the subsets Z_j^M of the meta-training set $Z_{1:N}^M$ for all $j \neq i, j = 1,\ldots,N$. Similarly, the information density $\imath(W, Z_j|T=\tau)$ is defined for random variables $(W, Z_j) \sim P_{W,Z_j|T=\tau}$, where $P_{W,Z_j|T=\tau}$ is obtained by marginalizing the joint distribution $P_U P_{W|Z^M,U} P_{Z|T=\tau}^M$ over U and over data samples Z_k of the training set Z^M for all $k \neq j$ with $k = 1,\ldots,M$. The information density $\imath(U, Z_i^M)$ quantifies the evidence for the hyperparameter U to be generated by the meta-learner $P_{U|Z_{1:N}^M}$ based on meta-training data that includes the data set Z_i^M. Similarly, the evidence for the model parameter W to be produced by the base learner $P_{W|Z^M}$ (which is the marginal of the joint distribution $P_U P_{W|Z^M,U}$) based on the training set for task τ that includes the data sample Z_j is captured by the information density $\imath(W, Z_j|T=\tau)$. All these measures can also be interpreted as the sensitivity of hyperparameter and model parameter to per-task data set Z_i^M (from source or target environment) and data sample Z_j within per-task data set, respectively. Moreover, the average of these information densities yield the following MI terms

$$I(U; Z_i^M) = \begin{cases} \mathbb{E}_{P_{Z_i^M} P_{U|Z_i^M}}[\imath(U, Z_i^M)] & \text{for } i = 1,\ldots,\beta N, \\ \mathbb{E}_{P'_{Z_i^M} P_{U|Z_i^M}}[\imath(U, Z_i^M)] & \text{for } i = \beta N+1,\ldots,N, \end{cases}$$

$$I(W; Z_j|T=\tau) = \mathbb{E}_{P_{W,Z_j|T=\tau}}[\imath(W, Z_j|T=\tau)] \text{ for } j = 1,\ldots,M. \tag{17}$$

Assumption 3.2: The source environment data distribution satisfies $P_{Z^M}(z^M) = 0$ almost surely for all $z^M = (z_1,\ldots,z_M) \in \mathcal{Z}^M$ such that $P'_{Z^M}(z^M) = 0$.

We are now ready to present two important inequalities that will underlie the analysis in the rest of the section. We note that a similar unified approach was presented in [9] to study generalization in conventional learning, and our methodology is inspired by this work. The proofs for these inequalities can be found in Appendix A.

Lemma 3.1: Under Assumption 3.1(a), the following inequality holds

$$\mathbb{E}_{P_{W,Z_j|T=\tau}}\left[\exp\left(\lambda(l(W,Z_j)-\mathbb{E}_{P_{W|T=\tau}P_{Z_j|T=\tau}}[l(W,Z_j)]-\frac{\lambda^2\delta_\tau^2}{2}-\imath(W,Z_j|T=\tau)\right)\right]\leq 1, \quad (18)$$

for all $j=1,\ldots,M$, $\lambda\in\mathbb{R}$ and for each task $\tau\in\mathcal{T}$.

Lemma 3.2: Under Assumption 3.1(b) and Assumption 3.2, we have the following inequalities

$$\mathbb{E}_{P'_{Z_i^M}P_{U|Z_i^M}}\left[\exp\left(\lambda(L_t(U|Z_i^M)-\mathbb{E}_{P_UP'_{Z_i^M}}[L_t(U|Z_i^M)]-\frac{\lambda^2\sigma^2}{2}-\imath(U,Z_i^M)\right)\right]\leq 1, \quad (19)$$

for $i=\beta N+1,\ldots N$ and

$$\mathbb{E}_{P_{Z_i^M}P_{U|Z_i^M}}\left[\exp\left(\lambda(L_t(U|Z_i^M)-\mathbb{E}_{P_UP'_{Z_i^M}}[L_t(U|Z_i^M)]-\frac{\lambda^2\sigma^2}{2}\right.\right.$$
$$\left.\left.-\log\frac{P_{Z_i^M}(Z_i^M)}{P'_{Z_i^M}(Z_i^M)}-\imath(U,Z_i^M)\right)\right]\leq 1, \quad (20)$$

for $i=1,\ldots,\beta N$, which holds for all $\lambda\in\mathbb{R}$.

Inequalities (18)–(20) relate the per-task training and meta-training loss functions to the corresponding ensemble averages and information densities, and will be instrumental in deriving information theoretic bounds on average transfer meta-generalization gap and average transfer excess meta-risk.

B. Bounds on the Average Transfer Meta-Generalization Gap

In this section, we derive upper bounds on the average transfer meta-generalization gap (12) for a general meta-learner $P_{U|Z_{1:N}^M}$. The results will be specialized to the EMRM meta-learner in Section III-C.

To start, we decompose the average transfer meta-generalization gap (12) as

$$\mathbb{E}_{P_{Z_{1:N}^M,U}}[\Delta\mathcal{L}'(U|Z_{1:N}^M)]=\mathbb{E}_{P_{Z_{1:N}^M,U}}\left[\left(\mathcal{L}'_g(U)-\mathcal{L}'_{g,t}(U)\right)+\left(\mathcal{L}'_{g,t}(U)-\mathcal{L}_t(U|Z_{1:N}^M)\right)\right], \quad (21)$$

where we have used the notation $P_{Z_{1:N}^M,U}=P_{Z_{1:N}^M}P_{U|Z_{1:N}^M}$, and $\mathcal{L}'_{g,t}(u)$ is the average training loss when data is drawn from the distribution $P_{Z|T}$ of a task T sampled from the target task distribution P'_T, i.e.

$$\mathcal{L}'_{g,t}(u)=\mathbb{E}_{P'_T}\mathbb{E}_{P_{Z|T}^M}[L_t(u|Z^M)]. \quad (22)$$

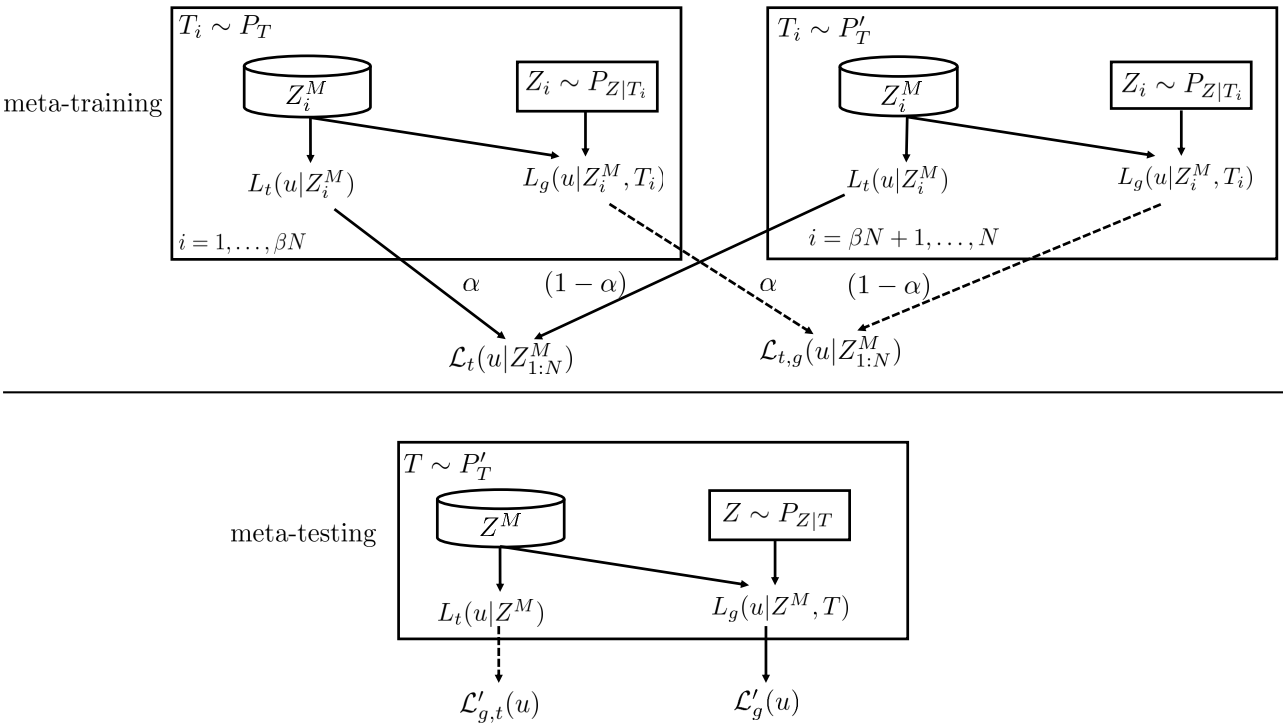


Fig. 3: Illustration of the variables involved in the definition of transfer meta-generalization gap (11).

A summary of all definitions for transfer meta-learning can be found in Figure 3.

The decomposition (21) captures two distinct contributions to the meta-generalization gap in transfer meta-learning. The first difference in (21) accounts for the *within-task generalization gap* that is caused by the observation of a finite number M of data samples for the meta-test task. In contrast, the second difference accounts for the *environment-level generalization gap* that results from the finite number of observed tasks (βN from the source environment and $(1-\beta)N$ from the target environment), as well as from the *meta-environment shift* in task distributions from P_T to P'_T. To upper bound the average transfer meta-generalization gap, we proceed by separately bounding the two differences in (21) via the exponential inequalities (18)–(20). This results in the following information-theoretic upper bound for transfer meta-learning that extends the individual sample mutual information based approach in [12] for conventional learning.

Theorem 3.1: Under Assumption 3.1 and Assumption 3.2, the following upper bound on the

average transfer meta-generalization gap holds for $\beta \in (0,1)$

$$
\begin{aligned}
&|\mathbb{E}_{P_{Z_{1:N}^M,U}}[\Delta\mathcal{L}'(U|Z_{1:N}^M)]| \\
&\leq \frac{\alpha}{\beta N}\sum_{i=1}^{\beta N}\sqrt{2\sigma^2\left(D(P_{Z^M}||P'_{Z^M})+I(U;Z_i^M)\right)}+\frac{1-\alpha}{(1-\beta)N}\sum_{i=\beta N+1}^{N}\sqrt{2\sigma^2 I(U;Z_i^M)} \\
&+\mathbb{E}_{P'_T}\left[\frac{1}{M}\sum_{j=1}^{M}\sqrt{2\delta_T^2 I(W;Z_j|T=\tau)}\right], \qquad (23)
\end{aligned}
$$

with the MI terms defined in (17).

Proof: See Appendix B. ■

The upper bound (23) on the average transfer meta-generalization gap is expressed in terms of three distinct contributions (i) *source environment-level generalization gap*: the MI $I(U;Z_i^M)$, for $i=1,\ldots,\beta N$, captures the sensitivity of the meta-learner U to the per-task data Z_i^M of the source environment data set, while the meta-environment shift between the source and target environment per-task data is captured by the KL divergence $D(P_{Z^M}||P'_{Z^M})$; (ii) *target environment-level generalization gap*: the MI $I(U;Z_i^M)$, for $i=\beta N+1,\ldots,N$, accounts for the sensitivity of the meta-learner to the per-task data sample Z_i^M from the target task environment; and lastly (iii) *within-task generalization gap*: the MI $I(W;Z_j|T=\tau)$ captures the sensitivity of the base learner to the data sample Z_j of the meta-test task data $Z^M \sim P_{Z|T=\tau}^M$.

As N increases, the dependence of a well-designed meta-learner output on each individual task-data set is expected to decrease, yielding a vanishing MI $I(U;Z_i^M)$. Similarly, with an increase in number M of per-task data samples, the MI $I(W;Z_j|T=\tau)$ is expected to decrease to zero. An interesting observation from (23) is that, even if these conditions are satisfied, as N, $M \to \infty$, the meta-environment shift between source and target task distributions results in a non-vanishing bound on the transfer meta-generalization gap, which is quantified by the KL divergence $D(P_{Z^M}||P'_{Z^M})$. Futhermore, when no data from target environment is available for meta-training, the bound in (23) can be specialized as follows.

Corollary 3.2: Under Assumption 3.1 and Assumption 3.2, when only data from the source environment is available for meta-training, i.e., when $\beta = 1$ and $\alpha = 1$, the following upper

bound on average transfer meta-generalization gap holds

$$|\mathbb{E}_{P_{Z_{1:N}^M,U}}[\Delta\mathcal{L}'(U|Z_{1:N}^M)]| \\ \leq \frac{1}{N}\sum_{i=1}^{N}\sqrt{2\sigma^2\Big(D(P_{Z^M}||P'_{Z^M})+I(U;Z_i^M)\Big)}+\mathbb{E}_{P'_T}\left[\frac{1}{M}\sum_{j=1}^{M}\sqrt{2\delta_T^2 I(W;Z_j|T=\tau)}\right]. \quad (24)$$

If, in addition, the source and target task distributions coincide, i.e., if $P_T = P'_T$, the bound (23) recovers the following result presented in [16, Cor. 5.8].

Corollary 3.3: When the source and task environment data distributions coincide, i.e., when $P_T = P'_T$, for $\beta = 1$ and $\alpha = 1$, we have the following upper bound on average meta-generalization gap

$$|\mathbb{E}_{P_{Z_{1:N}^M,U}}[\Delta\mathcal{L}(U|Z_{1:N}^M)]| \\ \leq \frac{1}{N}\sum_{i=1}^{N}\sqrt{2\sigma^2 I(U;Z_i^M)}+\mathbb{E}_{P_T}\left[\frac{1}{M}\sum_{j=1}^{M}\sqrt{2\delta_T^2 I(W;Z_j|T=\tau)}\right]. \quad (25)$$

Finally, the upper bound in (23) on average transfer meta-generalization gap can be specialized to recover the following upper bound [15, Cor. 2] on average generalization gap in conventional transfer learning (see Remark II.1).

Corollary 3.4: Consider the setting of Theorem 3.1 with $P_T = \delta(T-\tau)$ and $P'_T = \delta(T-\tau')$ for some $\tau, \tau' \in \mathcal{T}$. For $N = 2$, let the meta-training set be $Z_{1:N}^M = (Z_\tau^{\beta\bar{M}}, Z_{\tau'}^{(1-\beta)\bar{M}}) := Z^{\bar{M}}$ where $\bar{M} = NM$ and $Z_\tau^{\beta\bar{M}} \sim P_{Z|\tau}^{\beta\bar{M}}$ and $Z_{\tau'}^{(1-\beta)\bar{M}} \sim P_{Z|\tau'}^{(1-\beta)\bar{M}}$. Assume that $P_{W|Z^M,U} = \delta(W-U)$ and fix $W = U$. Then, the following upper bound on the average generalization gap for transfer learning holds for $\beta \in (0,1)$

$$|\mathbb{E}_{P_{Z^{\bar{M}},W}}[L_g(W|\tau') - L_t(W|Z^{\bar{M}})]| \leq \frac{\alpha}{\beta\bar{M}}\sum_{i=1}^{\beta\bar{M}}\sqrt{2\delta_{\tau'}^2\Big(D(P_{Z|\tau}||P_{Z|\tau'})+I(W;Z_i)\Big)} \\ + \frac{1-\alpha}{(1-\beta)\bar{M}}\sum_{i=\beta\bar{M}+1}^{\bar{M}}\sqrt{2\delta_{\tau'}^2 I(W;Z_i)}. \quad (26)$$

where the MI $I(W;Z_i)$ is evaluated with respect to the joint distribution $P_{W,Z_i|\tau}$ for $i = 1,\ldots,\beta\bar{M}$ and is evaluated with respect to the joint distribution $P_{W,Z_i|\tau'}$ for $i = \beta\bar{M}+1,\ldots,\bar{M}$.

Proof: See Appendix C. ■

Finally, we remark that, as proved in Appendix D, all the upper bounds in this section, starting from (23), can be also obtained under the following different assumption analogous to the one considered in the work of Xu and Raginsky [11].

Assumption 3.3: For every task $\tau \in \mathcal{T}$, the loss function $l(w, Z)$ is δ_τ^2-sub-Gaussian when $Z \sim P_{Z|T=\tau}$ for all $w \in \mathcal{W}$. Similarly, the per-task average training loss $L_t(u|Z^M)$ is σ^2-sub-Gaussian when $Z^M \sim P'_{Z^M}$ for all $u \in \mathcal{U}$.

As discussed in [13], this assumption does not imply Assumption 3.1, and vice versa. This is unless the loss function $l(\cdot, \cdot)$ is bounded in the interval $[a, b]$, in which case both these assumptions hold.

C. Bounds on Transfer Excess Meta-Risk of EMRM

In this section, we obtain an upper bound on the average transfer meta-excess risk (14) for the EMRM meta-learner (10). We will omit the dependence of U^{EMRM} on $Z_{1:N}^M$ to simplify notation. We start by decomposing the average transfer excess meta-risk (14) of EMRM as

$$\begin{aligned}
&\mathbb{E}_{P_{Z_{1:N}^M}}[\mathcal{L}'_g(U^{\mathrm{EMRM}}) - \mathcal{L}'_g(u^*)] \\
&= \mathbb{E}_{P_{Z_{1:N}^M}}\Big[\Big(\mathcal{L}'_g(U^{\mathrm{EMRM}}) - \mathcal{L}_t(U^{\mathrm{EMRM}}|Z_{1:N}^M)\Big) + \Big(\mathcal{L}_t(U^{\mathrm{EMRM}}|Z_{1:N}^M) - \mathcal{L}_t(u^*|Z_{1:N}^M)\Big) \\
&+ \Big(\mathcal{L}_t(u^*|Z_{1:N}^M) - \mathcal{L}'_g(u^*)]\Big)\Big]. \qquad (27)
\end{aligned}$$

We first observe that we have the inequality $\mathcal{L}_t(U^{\mathrm{EMRM}}|Z_{1:N}^M) \leq \mathcal{L}_t(u^*|Z_{1:N}^M)$ which is by the definition of EMRM (10). Therefore, from (27), the average transfer meta-excess risk is upper bounded by the sum of average transfer meta-generalization gap studied above, which is the first difference in (27), and of the average difference $\mathbb{E}_{P_{Z_{1:N}^M}}[\mathcal{L}_t(u^*|Z_{1:N}^M) - \mathcal{L}'_g(u^*)]$, the last difference in (27). Combining a bound on this term with the bound (23) on the transfer meta-generalization gap yields the following upper bound on the average transfer excess meta-risk.

Theorem 3.5: Under Assumption 3.3 and Assumption 3.2, and for $\beta \in (0, 1)$, the following upper bound on the average transfer meta-excess risk holds for the EMRM meta-learner (10)

$$\begin{aligned}
&\mathbb{E}_{P_{Z_{1:N}^M}}[\mathcal{L}'_g(U^{\mathrm{EMRM}}) - \mathcal{L}'_g(u^*)] \\
&\leq \frac{\alpha}{\beta N}\sum_{i=1}^{\beta N}\sqrt{2\sigma^2\Big(D(P_{Z^M}||P'_{Z^M}) + I(U^{\mathrm{EMRM}}; Z_i^M)\Big)} \\
&+ \frac{1-\alpha}{(1-\beta)N}\sum_{i=\beta N+1}^{N}\sqrt{2\sigma^2 I(U^{\mathrm{EMRM}}; Z_i^M)} + \mathbb{E}_{P'_T}\left[\frac{1}{M}\sum_{j=1}^{M}\sqrt{2\delta_T^2 I(W; Z_j|T=\tau)}\right] \\
&+ \alpha\sqrt{2\sigma^2 D(P_{Z^M}||P'_{Z^M})} + \mathbb{E}_{P'_T}\left[\frac{1}{M}\sum_{j=1}^{M}\sqrt{2\delta_T^2 I(W; Z_j|T=\tau, u^*)}\right], \qquad (28)
\end{aligned}$$

where the MI terms are defined in (17) with $U = U^{\mathrm{EMRM}}$.

Proof: See Appendix E. ■

Comparing (28) with (23) reveals that, in addition to the terms contributing to the average transfer meta-generalization gap, the excess meta-risk of EMRM meta-learner also includes the KL divergence between source and target environment per-task data $D(P_{Z^M}||P'_{Z^M})$ and the MI $I(W; Z_j|u^*, \tau)$. The latter captures the sensitivity of the base learner $P_{W|Z^M,u^*}$ under the optimal hyperparameter u^* to a training sample Z_j of the meta-test task data $Z^M \sim P'_{Z^M}$. Since u^* is unknown in general, once can further upper bound this mutual information by the supremum value $\sup_{u\in\mathcal{U}} I(W; Z_j|T = \tau, u)$.

All the bounds obtained in this section depend on the distributions of source and target task environments, namely P_T, and P'_T, and per-task data distributions $\{P_{Z|T=\tau}\}_{\tau\in\mathcal{T}}$, all of which are generally unknown. In the next section, we obtain high-probability PAC-Bayesian bounds on the transfer meta-generalization gap, which are in general independent of these distributions except for the quantity that captures the meta-environment shift. We further build on this bound to define a novel meta-learner inspired by the principle of information risk minimization [29].

IV. Information Risk Minimization for Transfer Meta-Learning

In this section, we first obtain a novel PAC-Bayesian bound on the transfer meta-generalization gap which holds with high probability over the meta-training set. Based on the derived bound, we then propose a new meta-training algorithm, termed Information Meta Risk Minimization (IMRM), that is inspired by the principle of information risk minimization [29]. This will be compared to EMRM through a numerical example in Section VI.

We first discuss in the next sub-section some technical assumptions that are central to the derivation of PAC-Bayesian bound for transfer meta-learning. We then present the PAC-Bayesian bounds in Section IV-B, and we introduce IMRM in Section IV-C.

A. Assumptions

The derivation of the PAC-Bayesian bound relies on slightly different conditions than Assumption 3.3, which are stated next.

Assumption 4.1: The environment distributions P_T, P'_T and $\{P_{Z|T=\tau}\}_{\tau\in\mathcal{T}}$, the base learner $P_{W|Z^M,U}$ and the meta-learner $P_{U|Z^M_{1:N}}$ satisfy the following assumptions.

(a) For each task $\tau \in \mathcal{T}$, the loss function $l(w, Z)$ is δ_τ^2-sub-Gaussian under $Z \sim P_{Z|\tau}$ for all $w \in \mathcal{W}$.

(b) The average per-task generalization loss $L_g(u|T, Z^M)$ in (4) is σ^2-sub-Gaussian when $(T, Z^M) \sim P'_T P^M_{Z|T}$ for all $u \in \mathcal{U}$.

Assumption 4.1(a) on the loss function $l(w, Z)$ is the same as the one considered in Assumption 3.3. In contrast, while Assumption 4.1(b) is on the average per-task generalization loss, Assumption 3.3 considers average per-task training loss. This distinction is necessary in order to also bound the task-level generalization gap in high probability. If the loss function is bounded in the interval $[a, b]$, then both Assumption 3.3 and Assumption 4.1 are satisfied with $\sigma^2 = \delta_\tau^2$ for all $\tau \in \mathcal{T}$.

PAC-Bayes bounds depend on arbitrary reference data-independent "prior" distributions that allow the evaluation of sensitivity measures for base learners [42] and meta-learners [26]. Accordingly, in the following sections, we consider a hyper-prior $Q_U \in \mathcal{P}(\mathcal{U})$ for the hyperparameter and a family of priors $Q_{W|U=u} \in \mathcal{P}(\mathcal{W})$ for each $u \in \mathcal{U}$ satisfying the following assumption.

Assumption 4.2: The hyper-prior $Q_U \in \mathcal{P}(\mathcal{U})$ must satisfy that $P_{U|Z^M_{1:N}=z^M_{1:N}}(u) = 0$ almost surely for every $u \in \mathcal{U}$ such that $Q_U(u) = 0$, for all $z^M_{1:N} \in \mathcal{Z}^{MN}$. Similarly, for given $u \in \mathcal{U}$, the prior $Q_{W|U=u} \in \mathcal{P}(\mathcal{W})$ must satisfy that $P_{W|Z^M=z^M,U=u}(w) = 0$ almost surely for every $w \in \mathcal{W}$ such that $Q_{W|U=u}(w) = 0$, for all $z^M \in \mathcal{Z}^M$. Finally, $P_T(\tau) = 0$ almost surely for every $\tau \in \mathcal{T}$ such that $P'_T(\tau) = 0$.

The derivation of PAC-Bayesian bound is based on novel exponential inequalities that are derived in a similar manner as in the previous section and can be found in Appendix F. In the following, we use $T_{1:N} = (T_1, \ldots, T_N)$ to denote the N selected tasks for generating the meta-training data set $Z^M_{1:N}$ with $P_{T_{1:N}} = \prod_{i=1}^{\beta N} P_{T_i} \prod_{j=\beta N+1}^{N} P'_{T_i}$ and $P_{Z^M_{1:N}|T_{1:N}}$ denoting the product distribution $\prod_{i=1}^{N} P^M_{Z|T_i}$.

B. PAC-Bayesian Bound for Transfer Meta-Learning

In this section, we focus on obtaining PAC-Bayesian bounds of the following form: With probability at least $1 - \delta$ over the distribution of meta-training tasks and data $(T_{1:N}, Z^M_{1:N}) \sim P_{T_{1:N}} P_{Z^M_{1:N}|T_{1:N}}$, the transfer meta-generalization gap satisfies

$$\left| \mathbb{E}_{P_{U|Z^M_{1:N}}} [\Delta \mathcal{L}'(U|Z^M_{1:N})] \right| \leq \epsilon, \tag{29}$$

for $\delta \in (0,1)$. To start, we define the the empirical weighted average of the per-task test loss of the meta-training set as

$$\mathcal{L}_{t,g}(u|Z_{1:N}^M, T_{1:N}) = \frac{\alpha}{\beta N}\sum_{i=1}^{\beta N} L_g(u|Z_i^M, T_i) + \frac{1-\alpha}{(1-\beta)N}\sum_{i=\beta N+1}^{N} L_g(u|Z_i^M, T_i), \tag{30}$$

where $L_g(u|Z_i^M, T_i)$ is defined in (4). Then, the transfer meta-generalization gap can be decomposed as

$$\begin{aligned}&\mathbb{E}_{P_{U|Z_{1:N}^M}}\left[\Delta\mathcal{L}'(U|Z_{1:N}^M)\right]\\ &= \mathbb{E}_{P_{U|Z_{1:N}^M}}\Big[\Big(\mathcal{L}'_g(U) - \mathcal{L}_{t,g}(U|Z_{1:N}^M, T_{1:N})\Big) + \Big(\mathcal{L}_{t,g}(U|Z_{1:N}^M, T_{1:N}) - \mathcal{L}_t(U|Z_{1:N}^M)\Big)\Big].\end{aligned} \tag{31}$$

In (31), the first difference accounts for the *environment-level generalization gap* resulting from the observation of a finite number N of meta-training tasks and also from the meta-environment shift between source and target task distributions. The second difference accounts for the *within-task generalization gap* in each subset of the meta-training set $Z_{1:N}^M$ arising from observing a finite number M of per-task data samples. We note that the decomposition in (31) can also be used to obtain an upper bound on the average transfer meta-generalization gap. However, the resulting bound does not recover the bound in [16], or specialize to the case of conventional transfer learning. We leave a full investigation of this alternate bound to future work.

As in the bounds on average transfer meta-generalization gap presented in Section III-B, the idea is to separately bound the above two differences in high probability over $(T_{1:N}, Z_{1:N}^M) \sim P_{T_{1:N}} P_{Z_{1:N}^M|T_{1:N}}$ and then combine the results via union bound. This results in the following PAC-Bayesian bound.

Theorem 4.1: For a fixed base learner $P_{W|Z^M,U}$, let $Q_U \in \mathcal{P}(\mathcal{U})$ be an arbitrary hyper-prior distribution over the space of hyper-parameters and $Q_{W|U=u} \in \mathcal{P}(\mathcal{W})$ be an arbitrary prior distribution over the space of model parameters for each $u \in \mathcal{U}$ and $\beta \in (0,1)$. Then, under Assumption 4.1 and Assumption 4.2, the following inequality holds uniformly for any meta-

learner $P_{U|Z_{1:N}^M}$ with probability at least $1-\delta$, $\delta \in (0,1)$, over $(T_{1:N}, Z_{1:N}^M) \sim P_{T_{1:N}} P_{Z_{1:N}^M|T_{1:N}}$

$$\left|\mathbb{E}_{P_{U|Z_{1:N}^M}}[\Delta\mathcal{L}'(U|Z_{1:N}^M)]\right| \leq \sqrt{2\sigma^2\left(\frac{\alpha^2}{\beta N}+\frac{(1-\alpha)^2}{(1-\beta)N}\right)\left(\sum_{i=1}^{\beta N}\log\frac{P_T(T_i)}{P'_T(T_i)}+D(P_{U|Z_{1:N}^M}||Q_U)+\log\frac{2}{\delta}\right)}$$
$$+\frac{\alpha}{\beta N}\sum_{i=1}^{\beta N}\sqrt{\frac{2\delta_{T_i}^2}{M}\left(D(P_{U|Z_{1:N}^M}||Q_U)+\mathbb{E}_{P_{U|Z_{1:N}^M}}[D(P_{W|U,Z_i^M}||Q_{W|U})]+\log\frac{4\beta N}{\delta}\right)}$$
$$+\frac{1-\alpha}{(1-\beta)N}\sum_{i=\beta N+1}^{N}\sqrt{\frac{2\delta_{T_i}^2}{M}\left(D(P_{U|Z_{1:N}^M}||Q_U)+\mathbb{E}_{P_{U|Z_{1:N}^M}}[D(P_{W|U,Z_i^M}||Q_{W|U})]+\log\frac{4(1-\beta)N}{\delta}\right)}. \tag{32}$$

Proof: See Appendix G. ■

The first term in the upper bound (32) captures the environment-level generalization gap through the log-likelihood ratio $\log(P_T(T_i)/P'_T(T_i))$, which accounts for the meta-environment shift, and through the KL divergence $D(P_{U|Z_{1:N}^M}||Q_U)$. This quantifies the sensitivity of the meta-learner $P_{U|Z_{1:N}^M}$ to the meta-training set $Z_{1:N}^M$ through its divergence with respect to the data-independent hyper-prior Q_U. The second term of (32) captures the generalization gap within the task data from source environment in terms of the average KL divergence $\mathbb{E}_{P_{U|Z_{1:N}^M}}[D(P_{W|U,Z_i^M}||Q_{W|U})]$ between model posterior and prior distributions together with $D(P_{U|Z_{1:N}^M}||Q_U)$, while the last term accounts for the generalization gap within the task data from the target environment. We note that the average KL divergence, $\mathbb{E}_{P_{U|Z_{1:N}^M}}[D(P_{W|U,Z_i^M}||Q_{W|U})]$, quantifies the sensitivity of the base learner $P_{W|Z^M,U}$ to the training set Z^M through its divergence with respect to the data-independent prior $Q_{W|U}$ for a hyperparameter $U \sim P_{U|Z_{1:N}^M}$.

The bound in (32) can be relaxed to obtain the following looser bound that is more amenable to optimization, as we will discuss in the next subsection.

Corollary 4.2: In the setting of Theorem 4.1, the following inequality holds with probability at least $1-\delta$ over $(T_{1:N}, Z_{1:N}^M) \sim P_{T_{1:N}} P_{Z_{1:N}^M|T_{1:N}}$ for $\beta \in (0,1)$,

$$\mathbb{E}_{P_{U|Z_{1:N}^M}}[\mathcal{L}'_g(U)] \leq \mathbb{E}_{P_{U|Z_{1:N}^M}}\left[\mathcal{L}_t(U|Z_{1:N}^M)+\frac{\alpha}{\beta NM}\sum_{i=1}^{\beta N}D(P_{W|Z_i^M,U}||Q_{W|U})\right.$$
$$\left.+\frac{1-\alpha}{(1-\beta)NM}\sum_{i=\beta N+1}^{N}D(P_{W|Z_i^M,U}||Q_{W|U})\right]+\left(\frac{1}{N}+\frac{1}{M}\right)D(P_{U|Z_{1:N}^M}||Q_U)+\Psi, \tag{33}$$

where we have defined the quantity

$$\Psi = \frac{\sigma^2}{2}\left(\frac{\alpha^2}{\beta} + \frac{(1-\alpha)^2}{1-\beta}\right) + \frac{1}{N}\sum_{i=1}^{\beta N} \log \frac{P_T(T_i)}{P'_T(T_i)} + \frac{1}{N}\log\frac{2}{\delta} + \frac{\alpha}{\beta N}\sum_{i=1}^{\beta N}\frac{\delta_{T_i}^2}{2} + \frac{\alpha}{M}\log\frac{4\beta N}{\delta}$$
$$+ \frac{1-\alpha}{(1-\beta)N}\sum_{i=\beta N+1}^{N}\frac{\delta_{T_i}^2}{2} + \frac{1-\alpha}{M}\log\frac{4(1-\beta)N}{\delta}. \tag{34}$$

Proof: To obtain the required bound, we proceed as in the proof of Theorem 4.1. To bound the first difference of (31), we use the upper bound in (91) with $\lambda = -N$. To bound the second difference, we use the upper bound in (95) with $\lambda = -M$. Combining the resultant bounds via the union bound and rearranging results in (33). ■

C. Information Meta-Risk Minimization (IMRM) for Transfer Meta-Learning

For fixed base learner $P_{W|Z^M,U}$ and given prior $Q_{W|U}$ and hyper-prior Q_U distributions, the PAC-Bayesian bound in (33) holds for any meta-learner $P_{U|Z_{1:N}^M}$. Consequently, following the principle of information risk minimization [29], one can design a meta-learner $P_{U|Z_{1:N}^M}$ so as to minimize the upper bound (33) on the transfer meta-generalization loss. As compared to EMRM, this approach accounts for the transfer meta-generalization gap, and can hence outperform EMRM in terms of meta-generalization performance. The same idea was explored in [27] for conventional meta-learning, i.e., for the special case when $P_T = P'_T$.

To proceed, we consider $\beta \in (0,1)$ and denote

$$\mathcal{L}(u, Z_{1:N}^M) = \mathcal{L}_t(u|Z_{1:N}^M) + \frac{\alpha}{\beta NM}\sum_{i=1}^{\beta N} D(P_{W|Z_i^M,U=u}||Q_{W|U=u})$$
$$+ \frac{1-\alpha}{(1-\beta)NM}\sum_{i=\beta N+1}^{N} D(P_{W|Z_i^M,U=u}||Q_{W|U=u}) \tag{35}$$

as the meta-training loss regularized by the average KL divergence $D(P_{W|Z_i^M,U=u}||Q_{W|U=u})$ between the base learner output and the prior distribution $Q_{W|U=u}$ over the base learner input data from source and target environments. The IMRM meta-learner is then defined as any algorithm that solves the optimization problem

$$P_{U|Z_{1:N}^M}^{\text{IMRM}} = \arg \min_{P_{U|Z_{1:N}^M} \in \mathcal{P}(\mathcal{U})} \left(\mathbb{E}_{P_{U|Z_{1:N}^M}}[\mathcal{L}(U, Z_{1:N}^M)] + \left(\frac{1}{N} + \frac{1}{M}\right) D(P_{U|Z_{1:N}^M}||Q_U) \right). \tag{36}$$

For fixed N, M, $Q_U, Q_{W|U}$ and base learner $P_{W|Z^M,U}$, the IMRM meta-learner can be expressed as

$$P^{\text{IMRM}}_{U|Z^M_{1:N}}(u) \propto Q_U(u) \exp\Big(-\frac{NM}{N+M}\mathcal{L}(u, Z^M_{1:N})\Big), \tag{37}$$

where the normalization constant is given by $\mathbb{E}_{Q_U}\Big[\exp\Big(-NM\mathcal{L}(U; Z^M_{1:N})/(N+M)\Big)\Big]$.

For a given meta-training set, EMRM outputs the single value of the hyperparameter $u \in \mathcal{U}$ that minimizes the meta-training loss (9). In contrast, the IMRM meta-learner (37) updates the prior belief Q_U after observing meta-training set, producing a distribution in the hyperparameter space. Given the significance of the meta-learning criterion (36) as an upper bound on the transfer meta-generalization loss, the optimizing distribution (37) captures the impact of the epistemic uncertainty related to the limited availability of the meta-training data. In line with this discussion, it can be seen from (36) that as $M, N \to \infty$ with M/N equal to a constant, the IMRM meta-learner tends to EMRM.

To implement the proposed IMRM meta-learner, we adopt one of the two approaches. The first, referred to as *IMRM-mode*, selects a single hyperparameter centered at the mode of (37) as

$$U^{\text{IMRM}-\text{mode}}(Z^M_{1:N}) = \arg\max_{u\in\mathcal{U}} \; Q_U(u) \exp\Big(-\frac{NM}{N+M}\mathcal{L}(u; Z^M_{1:N})\Big). \tag{38}$$

IMRM-mode is akin to Maximum A Posteriori (MAP) inference in conventional machine learning. Alternatively, we obtain one sample from the IMRM meta-learner (37) for use by the base learner and then average the obtained transfer meta-generalization loss as per definition (12). This can be in practice done by using Monte Carlo methods such as Metropolis-Hastings or Langevin dynamics [43]. As mentioned, this approach, referred to *IMRM-Gibbs*, reduces to the EMRM in the limit as $M, N \to \infty$ when M/N is a constant.

V. Single-Draw Probability Bounds on Transfer Meta-Learning

So far, we have considered the performance of meta-learning procedures defined by a stochastic mapping $P_{U|Z^M_{1:N}}$ on average over distributions $P_{U|Z^M_{1:N}}$. As discussed in the context of IMRM, this implies that the performance metric of interest is to be evaluated by averaging over realizations of the hyperparameter $U \sim P_{U|Z^M_{1:N}}$. It is, however, also of interest to quantify performance guarantees under the assumption that a single draw $U \sim P_{U|Z^M_{1:N}}$ is fixed and used throughout. Similar single-draw bounds have been derived for conventional learning in [9]. With this goal

in mind, in this section, we present novel single-draw probability bounds for transfer meta-learning. The bound takes the following form: With probability at least $1-\delta$, with $\delta \in (0,1)$, over $(T_{1:N}, Z_{1:N}^M, U) \sim P_{T_{1:N}} P_{Z_{1:N}^M|T_{1:N}} P_{U|Z_{1:N}^M}$, the transfer meta-generalization gap satisfies the bound

$$\left|\Delta\mathcal{L}'(U|Z_{1:N}^M)\right| \leq \epsilon. \tag{39}$$

Towards the evaluation of single-draw bounds of this form, we resort again to the decomposition (31) used to derive the PAC-Bayesian bound in Section IV-B. We use the following *mismatched information density*

$$\jmath(U, Z_{1:N}^M) = \log \frac{P_{U|Z_{1:N}^M}(U|Z_{1:N}^M)}{Q_U(U)}, \tag{40}$$

which quantifies the evidence for the hyperparameter U to be generated according to the meta-learner $P_{U|Z_{1:N}^M}$ based on meta-training set, rather than being generated according to the hyper-prior distribution Q_U. Considering Assumption 4.1 on loss functions and Assumption 4.2 on information densities then yield the following single-draw probability bound for transfer meta-learning.

Theorem 5.1: For a fixed base learner $P_{W|Z^M,U}$, let $Q_U \in \mathcal{P}(\mathcal{U})$ be a hyper-prior distribution over the space of hyperparameters and $Q_{W|U=u} \in \mathcal{P}(\mathcal{W})$ be a prior distribution over the space of model parameters for each $u \in \mathcal{U}$ and $\beta \in (0,1)$. Then, under Assumption 4.1 and Assumption 4.2, the following inequality holds uniformly for any meta-learner $P_{U|Z_{1:N}^M}$ with probability at least $1-\delta$, $\delta \in (0,1)$, over $(T_{1:N}, Z_{1:N}^M, U) \sim P_{T_{1:N}} P_{Z_{1:N}^M|T_{1:N}} P_{U|Z_{1:N}^M}$

$$\begin{aligned}\left|\Delta\mathcal{L}'(U|Z_{1:N}^M)\right| \leq & \sqrt{2\sigma^2\left(\frac{\alpha^2}{\beta N} + \frac{(1-\alpha)^2}{(1-\beta)N}\right)\left(\sum_{i=1}^{\beta N} \log \frac{P_T(T_i)}{P'_T(T_i)} + \jmath(U, Z_{1:N}^M) + \log\frac{2}{\delta}\right)} \\ & + \frac{\alpha}{\beta N}\sum_{i=1}^{\beta N}\sqrt{\frac{2\delta_{T_i}^2}{M}\left(D(P_{W|Z_i^M,U}||Q_{W|U}) + \jmath(U, Z_{1:N}^M) + \log\frac{4\beta N}{\delta}\right)} \\ & + \frac{1-\alpha}{(1-\beta)N}\sum_{i=\beta N+1}^{N}\sqrt{\frac{2\delta_{T_i}^2}{M}\left(D(P_{W|Z_i^M,U}||Q_{W|U}) + \jmath(U, Z_{1:N}^M) + \log\frac{4(1-\beta)N}{\delta}\right)}.\end{aligned} \tag{41}$$

Proof: See Appendix H. ■

As in the PAC-Bayesian bound (32), the upper bound in (41) comprises of three contributions: (i) the environment-level generalization gap, which is captured by the meta-environment shift

term $\log(P_T(T_i)/P'_T(T_i))$ and by the mismatched information density $\jmath(U, Z_{1:N}^M)$, with the latter quantifying the sensitivity of the meta-learner $P_{U|Z_{1:N}^M}$ to the meta-training set; (ii) the generalization within the task drawn from source environment, which is accounted for by the KL divergence $D(P_{W|Z_i^M,U}||Q_{W|U})$ quantifying the sensitivity of the base learner $P_{W|Z^M,U}$ to the training set Z^M through its divergence with respect to the prior distribution $Q_{W|U}$, along with the mismatched information density $\jmath(U, Z_{1:N}^M)$, and finally, (iii) the generalization gap within the task data from target environment, which is similarly captured through the KL divergence $D(P_{W|Z_i^M,u}||Q_{W|U})$ and the mismatched information density $\jmath(U, Z_{1:N}^M)$.

The bound in (41) can be specialized to the case of conventional meta-learning as given in the following corollary, which appears also to be a novel result.

Corollary 5.2: Assume that the source and target task distributions coincide, i.e., $P_T = P'_T$, and $\alpha = \beta = 1$. Then, under the setting of Theorem 5.1, the following bound holds with probability at least $1-\delta$, $\delta \in (0,1)$, over $(T_{1:N}, Z_{1:N}^M, U) \sim P_{T_{1:N}} P_{Z_{1:N}^M|T_{1:N}} P_{U|Z_{1:N}^M}$

$$\left|\Delta\mathcal{L}(U|Z_{1:N}^M)\right| \leq \sqrt{\frac{2\sigma^2}{N}\left(\jmath(U, Z_{1:N}^M) + \log\frac{2}{\delta}\right)} + \frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{2\delta_{T_i}^2}{M}\left(D(P_{W|Z_i^M,U}||Q_{W|U}) + \jmath(U, Z_{1:N}^M) + \log\frac{2N}{\delta}\right)}. \quad (42)$$

VI. Example

In this section, we consider the problem of estimating the mean of a Bernoulli process based on a few samples. To this end, we adopt a base learner based on biased regularization and meta-learn the bias as the hyperparameter [44].

A. Setting

The data distribution for each task is given as $P_{Z|T=\tau} \sim \mathrm{Bern}(\tau)$ for a task-specific mean parameter $\tau \in [0,1]$. For meta-training, we sample tasks from the source task distribution $\tau \sim P_T$ given by a beta distribution $\mathrm{Beta}(\tau; a, b)$ with shape parameters $a, b > 0$, while the target task distribution $\tau \sim P'_T$ encountered during meta-testing is $\mathrm{Beta}(\tau; a', b')$ with generally different shape parameters $a', b' > 0$. We recall that the mean of a random variable $\tau \sim \mathrm{Beta}(\tau; a, b)$ is given as $R(a,b) = a/(a+b)$ and the variance is $V(a,b) = ab/((a+b)^2(a+b+1))$. For any task τ, the base learner uses training data, distributed i.i.d. from $\mathrm{Bern}(\tau)$, to determine the model parameter W, which is used as a predictor of a new observation $Z \sim \mathrm{Bern}(\tau)$ at test time. The

loss function $l(w,z)=(w-z)^2$ measures the quadratic error between prediction and the test input z.

The base learner adopts a quadratic regularizer with bias given by a hyperparameter $u\in[0,1]$ [44], and randomizes its output. Accordingly, the base learner computes the empirical average $D_i=\frac{1}{M}\sum_{j=1}^{M}Z_{i,j}^{M}$, over the training set, where $Z_{i,j}^{M}$ denotes the jth data sample in the training set of ith task. Then, it computes the convex combination $R_i(u)=\gamma D_i+(1-\gamma)u$, with the hyperparameter $u\in[0,1]$, where $\gamma\in[0,1]$ is a fixed scalar. Finally, it outputs a random model parameter W with mean $R_i(u)$ by drawing W as

$$P_{W|Z_i^M,U=u}(w)=\mathrm{Beta}(w;cR_i(u),c(1-R_i(u))), \tag{43}$$

where $c>0$ is fixed and it determines the variance $V_i(u):=V(cR_i(u),c(1-R_i(u)))$ of the output of the base learner.

The meta-training loss (9) can be directly computed as

$$\begin{aligned}\mathcal{L}_t(u|Z_{1:N}^M)=&\frac{\alpha}{\beta N}\sum_{i=1}^{\beta N}\Big(V_i(u)+R_i(u)^2-2R_i(u)D_i+\sum_{j=1}^{M}\frac{1}{M}Z_{i,j}^2\Big)\\&+\frac{(1-\alpha)}{(1-\beta)N}\sum_{i=\beta N+1}^{N}\Big(V_i(u)+R_i(u)^2-2R_i(u)D_i+\sum_{j=1}^{M}\frac{1}{M}Z_{i,j}^2\Big),\end{aligned} \tag{44}$$

while the transfer meta-generalization loss (8) evaluates as

$$\begin{aligned}\mathcal{L}_g'(u)=&u(1-\gamma)\left(\frac{1}{c+1}+u(1-\gamma)\frac{c}{c+1}+2\gamma R'\frac{c}{c+1}-2R'\right)+\frac{\gamma R'}{c+1}\\&+\frac{\gamma^2 c}{c+1}\left(\frac{R'}{M}+(V'+R'^2)\left(1-\frac{1}{M}\right)\right)-2\gamma(V'+R'^2)+R',\end{aligned} \tag{45}$$

where $V'=V(a',b')$ is the variance and $R'=R(a',b')$ is the mean of the random variable $\tau\sim P_T'$.

B. Experiments

For the base learner as described above, we analyze the average transfer meta-generalization gap $\mathbb{E}_{P_{Z_{1:N}^M}P_{U|Z_{1:N}^M}}[\Delta\mathcal{L}'(U|Z_{1:N}^M)]$ in (12) under EMRM (10) and IMRM (37), as well as the average excess meta-risk $\mathbb{E}_{P_U}[\mathcal{L}_g'(U)]-\min_{u\in[0,1]}\mathcal{L}_g'(u)$. For IMRM, we consider a prior distribution $Q_{W|U=u}(w)=\mathrm{Beta}(w;cu,c-cu)$ in the space of model parameters and a hyper-prior distribution $Q_U(u)=\mathrm{Beta}(u;1.8,2.5)$ in the space of hyperparameters. The prior distribution $Q_{W|U}$ indicates that, in the absence of data, the base learner should select a model parameter

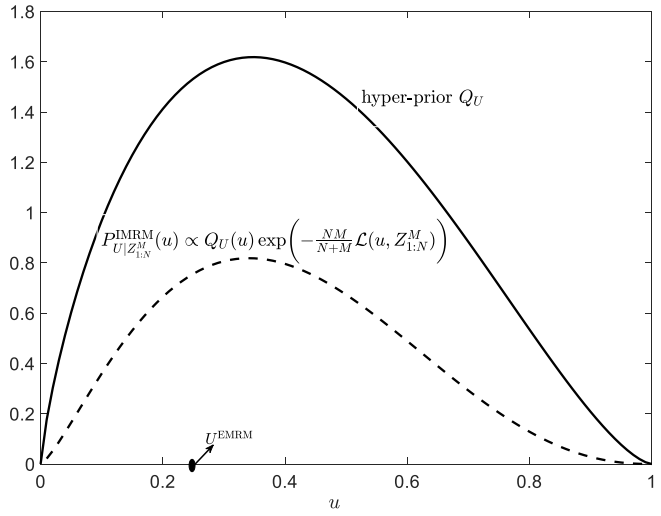


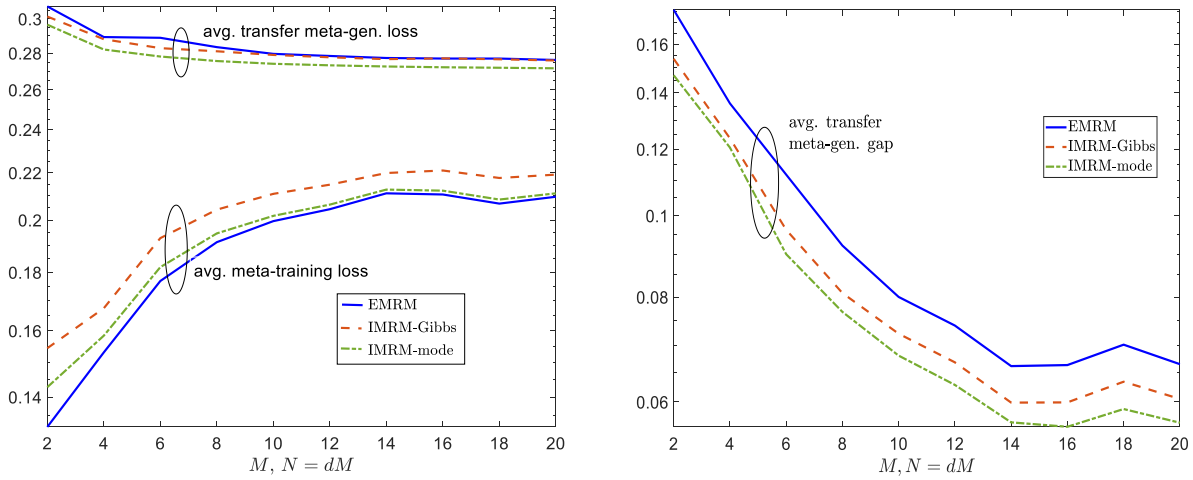
Fig. 4: Hyper-prior distribution Q_U, IMRM hyper-posterior $P^{\mathrm{IMRM}}_{U|Z^M_{1:N}}$ in (37), and EMRM solution (9) for a given meta-training set $Z^M_{1:N}$. ($M = 10$, $N = 8$, $a = 1.5$, $b = 7.5$, $a' = 4$, $b' = 5$, $c = 5$, $\alpha = 0.1$, $\beta = 0.6$, $\gamma = 0.55$)

with mean equal to the hyperparameter u. For the IMRM, we consider both IMRM-mode and IMRM-Gibbs.

To start, in Figure 4, we illustrate the hyper-prior $Q_U(u)$, the IMRM hyper posterior $P^{\mathrm{IMRM}}_{U|Z^M_{1:N}}$ in (37), and the output of EMRM (10). It is observed that, for the given values of $M = 10$ and $N = 8$ and for the given hyper-prior, the IMRM hyper-posterior retains information about the residual uncertainty on the value of the hyperparameter u, which is instead reduced to a point estimate based solely on meta-training data by EMRM.

In Figure 5, we then analyze the performance of EMRM, IMRM-mode and IMRM-Gibbs as a function of increasing values of M and N, for a fixed ratio $M/N = 0.85$, where $a = 1.5, b = 7.5$, $a' = 4$, $b' = 5$, $\gamma = 0.55$, $\alpha = 0.48$, $\beta = 0.48$ and $c = 5$. It can be seen that while EMRM yields, by definition, the smallest meta-training loss, IMRM improves the average transfer meta-generalization loss (Figure $5(a)$) by decreasing the average transfer meta-generalization gap (Figure $5(b)$). This gain is more significant for sufficiently small values of M and N, since, as M and N increases, IMRM tends to EMRM. We also observe that there exists a non-vanishing generalization gap even at high values of M and N. As discussed in Section III-B, this is caused by the meta-environment shift from P_T to P'_T. Finally, IMRM-mode and IMRM-Gibbs are seen to perform similarly, with the former being generally advantageous in this example. This suggest

that the main advantage of IMRM is due to the meta-regularizing effect of the KL term in (36). In the following two experiments, we adopt IMRM-mode.

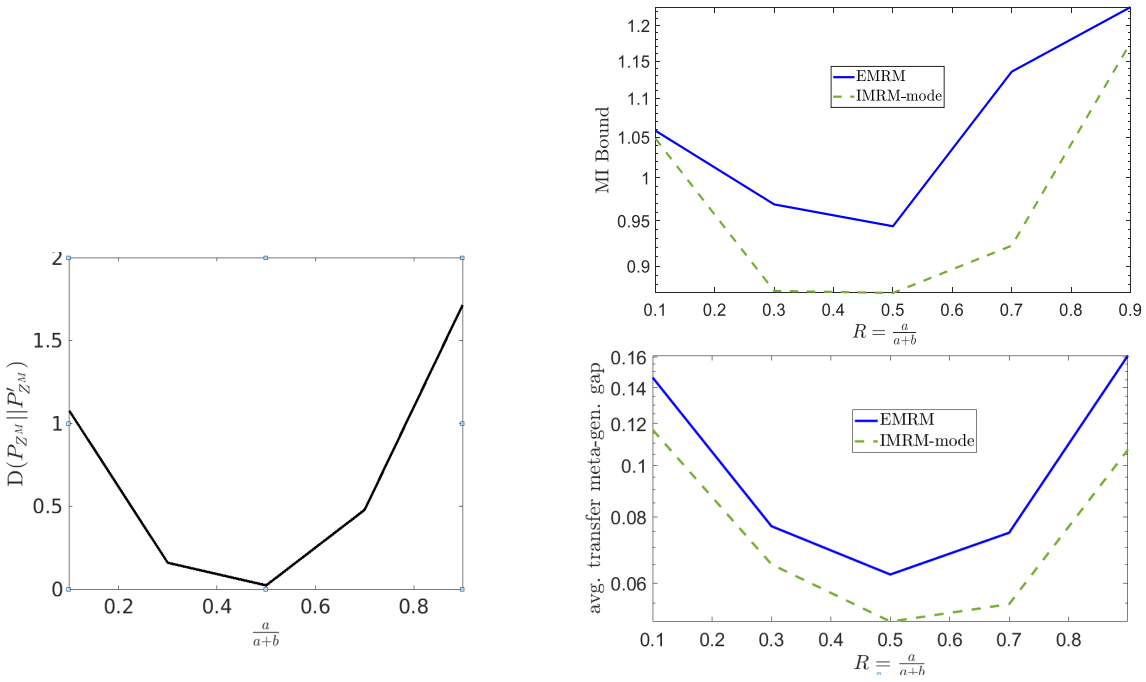


(a) (b)

Fig. 5: Average losses under EMRM, IMRM-mode and IMRM-Gibbs against increasing values of M and $N = dM$ with $d = 1/0.85$ for $a = 1.5, b = 7.5$, $a' = 4$, $b' = 5$, $\gamma = 0.55$, $\alpha = 0.48$, $\beta = 0.48$ and $c = 5$: (a) average transfer meta-generalization loss (8) and average meta-training loss (9), and (b) average transfer meta-generalization gap (12).

Figure 6 studies the impact of the meta-environment shift when the target task distribution P'_T is fixed to $\text{Beta}(a' = 4, b' = 5)$ and the source task distribution P_T is given as $\text{Beta}(a = 9 - b, b = 9(1 - R))$ with a varying mean $R = a/(a + b)$. Other parameters are set as $\gamma = 0.55$, $\alpha = 0.6$, $\beta = 0.6$, $N = 10$, $m = 5$ and $c = 5$. The analysis in Section III-B revealed that the KL divergence $D(P_{Z^M}||P'_{Z^M})$ between the data distributions under source and target environments is a key quantity in bounding the average transfer meta-generalization gap. The numerical results in the figure confirm that average transfer meta-generalization gap (23) for EMRM and IMRM-mode also shows a similar trend as the KL divergence as we vary the degree of meta-environment shift: The gap is small when P_T and P'_T are similar in term of KL divergence, and it increases when the divergence grows.

The average transfer excess meta-risk of EMRM and IMRM-mode are considered in Figure 7 as a function of the parameter α used in the definition (9) of the weighted meta-training loss. The choice of α that minimizes the average transfer excess meta-risk is seen to generally



(a) (b)

Fig. 6: Impact of meta-environment shift when P'_T is fixed to $\mathrm{Beta}(a'=4, b'=5)$ and P_T varies as $\mathrm{Beta}(a=9-b, b=9(1-R))$, $R=a/(a+b)$: (a) the KL divergence between P_{Z^M} and P'_{Z^M}; and (b) (top) MI bound on the average transfer meta-generalization gap (23); (bottom) average transfer meta-generalization gap for EMRM and IMRM-mode ($\gamma = 0.55$, $\alpha = 0.6$, $\beta = 0.6$, $N = 10$, $M = 5$ and $c = 5$).

lie somewhere between the extreme points $\alpha = 0$, which prescribes the use of only target environment data, or $\alpha = 1$, corresponding to the exclusive use of source environment datasets. Furthermore, the analytical bound (28) for EMRM (top figure) is seen to accurately predict the optimal value of α obtained from the actual average transfer excess meta-risk (14) (bottom figure). We note that it would also be interesting to derive similar analytical upper bound on the average transfer excess meta-risk for IMRM, by following the methodologies of papers such as [45], [46].

Finally, in Figure 8, we evaluate the *single-draw* probability bounds obtained in (41) for IMRM-Gibbs. Note that the single-draw performance of EMRM coincides with its average performance since it is deterministic. To illustrate the single-draw scenario, for each meta-training set of N tasks, we generate samples U of the hyperparameter according to $P^{\mathrm{IMRM}}_{U|Z^M_{1:N}}$.

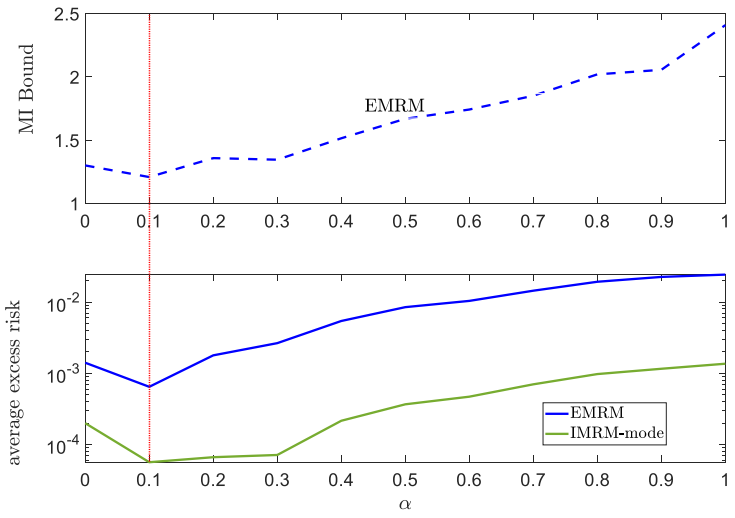


Fig. 7: Average transfer excess meta-risk as a function of the parameter α used in the definition (9) of the weighted meta-training loss: (top) MI-based bound on the average transfer excess meta-risk (28) for EMRM; (bottom) average excess transfer meta-risk for EMRM and IMRM-mode ($a = 1.67, b = 8.3$, $a' = 4.45, b' = 5.55$, $\gamma = 0.55$, $\beta = 0.4$, $N = 23$, $M = 15$ and $c = 5$).

We then compute the transfer meta-generalization gap $\Delta\mathcal{L}'(u|Z_{1:N}^M)$ for each of the generated samples. In the bottom panel of Figure 8, we use a box plot to illustrate the obtained empirical distribution of the transfer meta-generalization gap $\Delta\mathcal{L}'(U|Z_{1:N}^M)$ with $U \sim P_{U|Z_{1:N}^M}^{\text{IMRM}}$ for increasing values of N and fixed $M = 5$. For each value of N, the top of the box represents the 25th percentile ($\delta = 0.25$), the bottom corresponds to the 75th percentile ($\delta = 0.75$) and the centre dash correspond to the median ($\delta = 0.5$) of the distribution of $\Delta\mathcal{L}'(U|Z_{1:N}^M)$. The two lines outside the box are the "whiskers" that indicate the support of the empirical distribution. The information-density based single-draw upper bound (41) is illustrated for comparison in the top panel of Figure 8 for $\delta = 0.25, 0.5$, and 0.75. It can be seen that the bounds exhibit a similar decreasing trend as the empirical transfer meta-generalization gap in the bottom panel.

VII. Conclusions

This paper introduced the problem of transfer meta-learning, in which the meta-learner observes data from tasks belonging to a source task environment, while its performance is evaluated on a new meta-test task drawn from the target task environment. We obtained three forms of

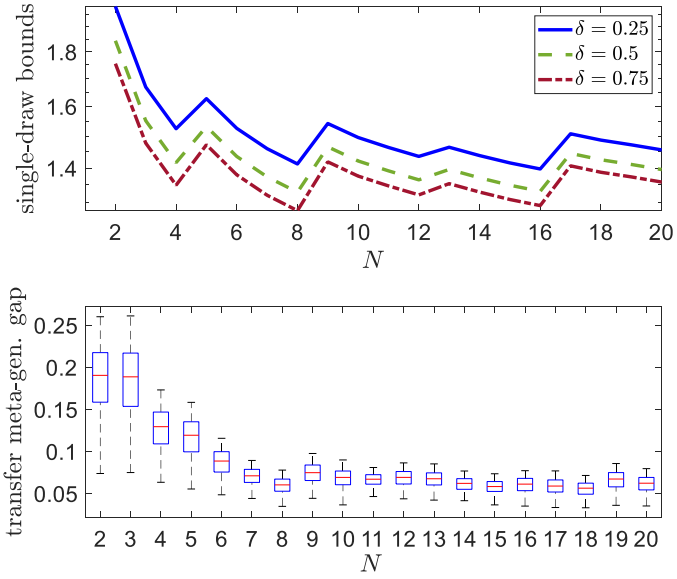


Fig. 8: Analysis of single-draw probability bounds for IMRM-Gibbs as a function of the number of tasks N for $\delta = 0.25, 0.5$ and 0.75. The top panel illustrates the single-draw bound (41) on transfer meta-generalization gap, while the bottom panel shows a box plot of the numerical evaluation of transfer meta-generalization gap. The lower quantile ($\delta = 0.25$) correspond to the top of the box and the upper quantile ($\delta = 0.75$) correspond to the bottom of the box, while the circled dot in the middle of the box indicates the median ($\delta = 0.5$) ($a = 1.5, b = 7.5$, $a' = 4, b' = 5$, $\gamma = 0.55$, $\beta = \alpha = 0.25$, $M = 5$ and $c = 5$).

upper bounds on the transfer meta-generalization gap – bounds on average generalization gap, high-probability PAC-Bayesian bounds and high-probability single-draw bounds. These bounds capture the meta-environment shift between source and target task distributions via the KL divergence between source and target data distributions for the average generalization gap bound, and the log-likelihood ratio between the source and target task distributions for the PAC-Bayesian and single-draw bounds. We note that these metrics can be numerically estimated from finite per-task data sets via various parametric or non-parametric methods [47]. Furthermore, we leveraged the derived PAC-Bayesian bound to propose a new meta-learning algorithm for transfer meta-learning, IMRM, which was shown in experiments to outperform an empirical weighted meta-risk minimization algorithm.

Directions for future work include the development of larger-scale experiments for linear

and non-linear base learners, the application of the bounding methodologies of [13], [14] and the analysis of the excess risk for IMRM by adapting the tools of [45], [46]. It would also be interesting to analyze bounds on transfer meta-generalization gap that capture the meta-environment shift via other statistical divergences like Jensen-Shannon divergences [48].

APPENDIX A
PROOFS OF LEMMA 3.1 AND LEMMA 3.2

Throughout the Appendices, we use the notation $P_{W|\tau}$ to denote the distribution $P_{W|T=\tau}$, $P_{Z|\tau}$ to denote $P_{Z|T=\tau}$ and $P_{W|Z^M,u}$ to denote $P_{W|Z^M,U=u}$. Under Assumption 3.1(a), the following inequality holds for each task $\tau \in \mathcal{T}$,

$$\mathbb{E}_{P_{W|\tau}P_{Z_j|\tau}}\left[\exp\left(\lambda(l(W,Z_j) - \mathbb{E}_{P_{W|\tau}P_{Z_j|\tau}}[l(W,Z)] - \frac{\lambda^2\delta_\tau^2}{2}\right)\right] \leq 1, \tag{46}$$

which in turn implies that

$$\mathbb{E}_{P_{W|\tau}P_{Z_j|\tau}}\left[\mathbb{I}_{\mathcal{E}}\exp\left(\lambda(l(W,Z) - \mathbb{E}_{P_{W|\tau}P_{Z_j|\tau}}[l(W,Z)] - \frac{\lambda^2\delta_\tau^2}{2}\right)\right] \leq 1, \tag{47}$$

where $\mathcal{E} = \text{supp}(P_{W,Z_j|\tau})$. Subsequently, using a change of measure from $P_{W|\tau}P_{Z_j|\tau}$ to $P_{W,Z_j|\tau}$ as in [49, Prop. 17.1] then yield the inequality (18).

Under Assumption 3.1(b), the following inequality holds for $i = 1, \ldots, N$,

$$\mathbb{E}_{P_U P'_{Z_i^M}}\left[\exp\left(\lambda(L_t(U|Z_i^M) - \mathbb{E}_{P_U P'_{Z_i^M}}[L_t(U|Z_i^M)] - \frac{\lambda^2\sigma^2}{2}\right)\right] \leq 1. \tag{48}$$

To get to (19), we note that (48) implies the following inequality for $i = \beta N + 1, \ldots, N$,

$$\mathbb{E}_{P_U P'_{Z_i^M}}\left[\mathbb{I}_{\mathcal{E}_1}\exp\left(\lambda(L_t(U|Z_i^M) - \mathbb{E}_{P_U P'_{Z_i^M}}[L_t(U|Z_i^M)] - \frac{\lambda^2\sigma^2}{2}\right)\right] \leq 1, \tag{49}$$

where $\mathcal{E}_1 = \text{supp}(P_{U|Z_i^M}P'_{Z_i^M})$. Applying change of measure as before from $P'_{Z_i^M}P_U$ to $P'_{Z_i^M}P_{U|Z_i^M}$ then yields inequality (19).

To get to (20), we start from (48), which implies for $i = 1, \ldots, \beta N$

$$\mathbb{E}_{P_U P'_{Z_i^M}}\left[\mathbb{I}_{\mathcal{E}_2}\exp\left(\lambda(L_t(U|Z_i^M) - \mathbb{E}_{P_U P'_{Z_i^M}}[L_t(U|Z_i^M)] - \frac{\lambda^2\sigma^2}{2}\right)\right] \leq 1, \tag{50}$$

where $\mathcal{E}_2 = \text{supp}(P_{Z_i^M})$. Performing change of measure from $P'_{Z_i^M}$ to $P_{Z_i^M}$ then gives that

$$\mathbb{E}_{P_U P_{Z_i^M}}\left[\exp\left(\lambda(L_t(U|Z_i^M) - \mathbb{E}_{P_U P'_{Z_i^M}}[L_t(U|Z_i^M)] - \frac{\lambda^2\sigma^2}{2} - \log\frac{P_{Z_i^M}(Z_i^M)}{P'_{Z_i^M}(Z_i^M)}\right)\right] \leq 1. \tag{51}$$

Applying the change of measure again from $P_{Z_i^M}P_U$ to $P_{Z_i^M}P_{U|Z_i^M}$ then yields (20).

APPENDIX B
PROOF OF THEOREM 3.1

To obtain the required upper bound on $|\mathbb{E}_{P_{Z^M_{1:N}}P_{U|Z^M_{1:N}}}[\Delta\mathcal{L}'(U|Z^M_{1:N})]|$, we leverage the decomposition in (21). Using triangle inequality, it then follows that

$$\begin{aligned}&|\mathbb{E}_{P_{Z^M_{1:N}}P_{U|Z^M_{1:N}}}[\Delta\mathcal{L}'(U|Z^M_{1:N})]|\\&\leq |\mathbb{E}_{P_U}[\mathcal{L}'_g(U)-\mathcal{L}'_{g,t}(U)|+|\mathbb{E}_{P_{Z^M_{1:N}}P_{U|Z^M_{1:N}}}[\mathcal{L}'_{g,t}(U)-\mathcal{L}_t(U|Z^M_{1:N})|.\end{aligned}\tag{52}$$

The idea is to separately bound the two averages in (52). Towards this, we first consider the average difference $|\mathbb{E}_{P_U}[\mathcal{L}'_g(U)-\mathcal{L}'_{g,t}(U)|$ which can be equivalently written as

$$\begin{aligned}&|\mathbb{E}_{P'_T P^M_{Z|T}}\mathbb{E}_{P_U P_{W|Z^M,U}}[L_g(W|T)-L_t(W|Z^M)|\\&\leq \mathbb{E}_{P'_T}|\mathbb{E}_{P^M_{Z|T}}\mathbb{E}_{P_{W|Z^M}}[L_g(W|T)-L_t(W|Z^M)|\qquad(53)\\&\leq \mathbb{E}_{P'_T}\left[\frac{1}{M}\sum_{j=1}^{M}\left|\mathbb{E}_{P_{W|T}P_{Z_j|T}}[l(W,Z_j)]-\mathbb{E}_{P_{W,Z_j|T}}[l(W,Z_j)]\right|\right].\qquad(54)\end{aligned}$$

We now bound the difference $\mathbb{E}_{P_{W|T}P_{Z_j|T}}[l(W,Z_j)]-\mathbb{E}_{P_{W,Z_j|T}}[l(W,Z_j)]$ using (18). For $T=\tau$, applying Jensen's inequality on (18) and taking log on both sides of the resultant inequality gives that

$$\lambda\left(\mathbb{E}_{P_{W,Z_j|T=\tau}}[l(W,Z_j)]-\mathbb{E}_{P_{W|T=\tau}P_{Z_j|T=\tau}}[l(W,Z)]\right)\leq\frac{\lambda^2\delta_\tau^2}{2}+I(W;Z_j|T=\tau).\tag{55}$$

Choosing $\lambda=\sqrt{2I(W;Z_j|T=\tau)}/\delta_\tau$ then yields that

$$[|\mathbb{E}_{P_{W,Z_j|T=\tau}}[l(W,Z_j)]-\mathbb{E}_{P_{W|T=\tau}P_{Z_j|T=\tau}}[l(W,Z_j)]|\leq\sqrt{2\delta_\tau^2 I(W;Z_j|T=\tau)}.\tag{56}$$

Substituting back in (54), averaging over T, then yields the following upper bound

$$\mathbb{E}_{P'_T}|\mathbb{E}_{P^M_{Z|T}}\mathbb{E}_{P_{W|Z^M}}[L_g(W|T)-L_t(W|Z^M)]|\leq\mathbb{E}_{P'_T}\left[\frac{1}{M}\sum_{j=1}^{M}\sqrt{2\delta_T^2 I(W;Z_j|T=\tau)}\right].\tag{57}$$

We now bound the second average difference in (52) using the the exponential inequalities (19)–(20). Towards this, we denote by $P_{Z^M_{1:\beta N}}$ the marginal of the joint distribution $\prod_{i=1}^{\beta N}P_{T_i}P^M_{Z|T_i}$ and by $P'_{Z^M_{\beta N+1:N}}$ the marginal of the joint distribution $\prod_{i=\beta N+1}^{N}P'_{T_i}P^M_{Z|T_i}$. We will also use

$$\mathcal{L}_t(u|Z^M_{1:\beta N})=\frac{1}{\beta N}\sum_{i=1}^{\beta N}L_t(u|Z^M_i)$$

for the the meta-training loss on task data from source environment and

$$\mathcal{L}_t(u|Z^M_{\beta N+1:N}) = \frac{1}{(1-\beta)N}\sum_{i=\beta N+1}^{N} L_t(u|Z_i^M)$$

for the meta-training loss on task data from target environment. Then, the second average difference in (52) can be equivalently written as

$$\begin{aligned}
&|\mathbb{E}_{P_{Z^M_{1:N},U}}[\mathcal{L}'_{g,t}(U) - \mathcal{L}_t(U|Z^M_{1:N})| \\
&= \left|\mathbb{E}_{P_{Z^M_{1:\beta N}}P'_{Z^M_{\beta N+1:N}}P_{U|Z^M_{1:N}}}\left[\alpha\left(\mathcal{L}'_{g,t}(U) - \mathcal{L}_t(U|Z^M_{1:\beta N})\right) + (1-\alpha)\left(\mathcal{L}'_{g,t}(U) - \mathcal{L}_t(U|Z^M_{\beta N+1:N})\right)\right]\right| \\
&\leq \alpha|\mathbb{E}_{P_{Z^M_{1:\beta N}}P_{U|Z^M_{1:\beta N}}}[\mathcal{L}'_{g,t}(U) - \mathcal{L}_t(U|Z^M_{1:\beta N})]| \\
&+ (1-\alpha)|\mathbb{E}_{P'_{Z^M_{\beta N+1:N}}P_{U|Z^M_{\beta N+1:N}}}[\mathcal{L}'_{g,t}(U) - \mathcal{L}'_t(U|Z^M_{\beta N+1:N})]| \\
&= \alpha\left|\frac{1}{\beta N}\sum_{i=1}^{\beta N}\left(\mathbb{E}_{P_U P'_{Z_i^M}}[L_t(U|Z_i^M)] - \mathbb{E}_{P_{Z_i^M}P_{U|Z_i^M}}[L_t(U|Z_i^M)]\right)\right| \\
&+ (1-\alpha)\left|\frac{1}{(1-\beta)N}\sum_{i=\beta N+1}^{N}\left(\mathbb{E}_{P_U P'_{Z_i^M}}[L_t(U|Z_i^M)] - \mathbb{E}_{P'_{Z_i^M}P_{U|Z_i^M}}[L_t(U|Z_i^M)]\right)\right|. \qquad (58)
\end{aligned}$$

We now proceed to use the exponential inequalities in (19) and (20) to bound the two terms in (58). To bound the first difference, we use (20). Applying Jensen's inequality and taking log on both sides of the resulting inequality yields

$$\lambda\left(\mathbb{E}_{P_{Z_i^M}P_{U|Z_i^M}}[L_t(U|Z_i^M)] - \mathbb{E}_{P_U P'_{Z_i^M}}[L_t(U|Z_i^M)]\right) \leq \frac{\lambda^2\sigma^2}{2} + D(P_{Z^M}||P'_{Z^M}) + I(U;Z_i^M). \qquad (59)$$

Further, choosing $\lambda = \sqrt{2(D(P_{Z^M}||P'_{Z^M}) + I(U;Z_i^M))}/\sigma$ then gives that

$$|\mathbb{E}_{P_{Z_i^M}P_{U|Z_i^M}}[L_t(U|Z_i^M)] - \mathbb{E}_{P_U P'_{Z_i^M}}[L_t(U|Z_i^M)]| \leq \sqrt{2\sigma^2\left(D(P_{Z^M}||P'_{Z^M}) + I(U;Z_i^M)\right)}. \qquad (60)$$

In a similar way, the second difference in (58) can be bounded by using (19). Applying Jensen's inequality, taking log on both sides, and finally choosing $\lambda = \sqrt{2I(U;Z_i^M)}/\sigma$ then yields

$$|\mathbb{E}_{P_U P'_{Z_i^M}}[L_t(U|Z^M)] - \mathbb{E}_{P'_{Z_i^M}P_{U|Z_i^M}}[L_t(U|Z_i^M)]| \leq \sqrt{2\sigma^2 I(U;Z_i^M)} \qquad (61)$$

Combining (60) and (61) in (58) and using it in (52) together with (57) gives the upper bound in (23).

APPENDIX C
PROOF OF COROLLARY 3.4

The bound (26) follows by specializing the bound (23) to the setting considered here. Towards this, we first note that the meta-training set $Z_{1:N}^M = Z^{\bar{M}} = (Z_1, \dots Z_{\bar{M}})$ with its ith sub-set Z_i^M corresponding to the data sample Z_i, where $Z_i \sim P_{Z|\tau}$ for $i = 1, \dots, \beta\bar{M}$ and $Z_i \sim P_{Z|\tau'}$ for $i = \beta\bar{M}+1, \dots, \bar{M}$. Thus, there are $\beta NM = \beta\bar{M}$ data samples from the source task environment and $(1-\beta)\bar{M}$ samples from the target task environment. Using $P_{W|Z^M,U} = \delta(W-U)$ and $U = W$, we then have $\mathcal{L}'_g(u) = L_g(w|\tau')$, and $\mathcal{L}_t(u|Z_{1:N}^M) = L_t(w|Z^{\bar{M}})$ with $L_t(u|Z_i^M) = l(w, Z_i)$. Consequently, we have $I(U; Z_i^M) = I(W; Z_i)$ and the KL divergence $D(P_{Z^M}||P'_{Z^M}) = D(P_{Z|\tau}||P_{Z|\tau'})$. It can also be verified that $P_{W,Z_j|\tau'} = P_{W|\tau'}P_{Z_j|\tau'} = P_W P_{Z_j|\tau'}$ whereby the MI $I(W; Z_j|\tau') = 0$ in (23). Further, since $L_t(u|z^M) = l(w, z)$, Assumption 3.1 then implies that $\sigma^2 = \delta_{\tau'}^2$. Using all these expressions in (23) yields the bound (26).

APPENDIX D
EXPONENTIAL INEQUALITIES BASED ON ASSUMPTION 3.3

Lemma D.1: Under Assumption 3.3 and Assumption 3.2, the following inequality holds for $j = 1, \dots, M$,

$$\mathbb{E}_{P_{W,Z_j|\tau}}\left[\exp\left(\lambda(l(W,Z_j) - \mathbb{E}_{P_{Z_j|\tau}}[l(W,Z_j)] - \imath(W,Z_j|T=\tau) - \frac{\lambda^2\delta_\tau^2}{2}\right)\right] \leq 1, \tag{62}$$

for all $\lambda \in \mathbb{R}$ and for each task $\tau \in \mathcal{T}$. Moreover, we have the following inequality for $i = 1, \dots, \beta N$

$$\mathbb{E}_{P_{Z_i^M}P_{U|Z_i^M}}\left[\exp\left(\lambda(L_t(U|Z_i^M) - \mathbb{E}_{P'_{Z_i^M}}[L_t(U|Z_i^M)] - \log\frac{P_{Z_i^M}(Z_i^M)}{P'_{Z_i^M}(Z_i^M)} - \imath(U,Z_i^M) - \frac{\lambda^2\sigma^2}{2}\right)\right] \leq 1, \tag{63}$$

and for $i = \beta N + 1, \dots, N$, we have

$$\mathbb{E}_{P'_{Z_i^M}P_{U|Z_i^M}}\left[\exp\left(\lambda(L_t(U|Z_i^M) - \mathbb{E}_{P'_{Z_i^M}}[L_t(U|Z_i^M)] - \imath(U,Z_i^M) - \frac{\lambda^2\sigma^2}{2}\right)\right] \leq 1, \tag{64}$$

which holds for all $\lambda \in \mathbb{R}$.

Proof: Under Assumption 3.3, the following inequality holds for each task $\tau \in \mathcal{T}$ and for all $w \in \mathcal{W}$ and $\lambda \in \mathbb{R}$,

$$\mathbb{E}_{P_{Z_j|\tau}}\left[\exp\left(\lambda(l(w,Z_j) - \mathbb{E}_{P_{Z_j|\tau}}[l(w,Z)] - \frac{\lambda^2\delta_\tau^2}{2}\right)\right] \leq 1. \tag{65}$$

Now, averaging both sides with respect to $W \sim P_{W|\tau}$, where $P_{W|\tau}$ is obtained by marginalizing $P_{W|Z^M,U}P_U P_{Z^M|T=\tau}$, we get that

$$\mathbb{E}_{P_{W|\tau}P_{Z_j|\tau}}\left[\exp\left(\lambda(l(W,Z_j) - \mathbb{E}_{P_{Z_j|\tau}}[l(W,Z)] - \frac{\lambda^2\delta_\tau^2}{2}\right)\right] \leq 1. \tag{66}$$

Performing change of measure from $P_{Z_j|\tau}P_{W|\tau}$ to $P_{W,Z_j|\tau}$ similar to Appendix A gets us to the exponential inequality in (62).

Similarly, for obtaining environment-level exponential inequalities, we have from Assumption 3.3 the following inequality

$$\mathbb{E}_{P'_{Z_i^M}}\left[\exp\left(\lambda(L_t(u|Z_i^M) - \mathbb{E}_{P'_{Z_i^M}}[L_t(u|Z_i^M)] - \frac{\lambda^2\sigma^2}{2}\right)\right] \leq 1, \tag{67}$$

for $i = 1, \ldots, N$, which holds for all $u \in \mathcal{U}$ and $\lambda \in \mathbb{R}$. Now, to get to (63), average both sides with respect to $U \sim P_U$, and change measure from $P'_{Z_i^M}$ to $P_{Z_i^M}$. This results in the following for $i = 1, \ldots, \beta N$

$$\mathbb{E}_{P_{Z_i^M}P_U}\left[\exp\left(\lambda(L_t(U|Z_i^M) - \mathbb{E}_{P'_{Z_i^M}}[L_t(U|Z_i^M)] - \log\frac{P_{Z_i^M}(Z_i^M)}{P'_{Z_i^M}(Z_i^M)} - \frac{\lambda^2\sigma^2}{2}\right)\right] \leq 1. \tag{68}$$

Performing a second change of measure from $P_{Z_i^M}P_U$ to $P_{Z_i^M}P_{U|Z_i^M}$ then yields the exponential inequality in (63). For $i = \beta N + 1, \ldots, N$, we obtain (64) from (67) by first averaging over P_U, then performing a change of measure from $P'_{Z_i^M}P_U$ to $P'_{Z_i^M}P_{U|Z_i^M}$. ■

To see how the exponential inequalities in Lemma D.1 yields the upper bound in Theorem 3.1, we proceed as in the proof of Theorem 3.1 in Appendix B. To bound the difference in expectation in (54), we use the exponential inequality (62). Note that applying Jensen's inequality on (62) results in the inequality (55). Similarly, to bound the enviroment-level generalization gap in (58), we use the exponential inequalities (63) and (64) and apply Jensen's inequality. In particular, applying Jensen's inequality on (63) leads to (59). The required bound is then obtained by proceeding as in Appendix B.

APPENDIX E
PROOF OF THEOREM 3.5

For obtaining an upper bound on the average transfer meta-excess risk, we bound the average transfer generalization gap, the first difference in (27), by (23).

We now bound the second difference in (27). This can be equivalently written as

$$\mathbb{E}_{P_{Z^M_{1:N}}}[\mathcal{L}_t(u^*|Z^M_{1:N}) - \mathcal{L}'_g(u^*)] \tag{69}$$

$$= \mathbb{E}_{P_{Z^M_{1:N}}}[\mathcal{L}_t(u^*|Z^M_{1:N}) - \mathcal{L}'_{g,t}(u^*) + \mathcal{L}'_{g,t}(u^*) - \mathcal{L}'_g(u^*)] \tag{70}$$

$$= \alpha\mathbb{E}_{P_{Z^M_{1:\beta N}}}[\mathcal{L}_t(u^*|Z^M_{1:\beta N}) - \mathcal{L}'_{g,t}(u^*)] + (1-\alpha)\mathbb{E}_{P'_{Z^M_{\beta N+1:N}}}[\mathcal{L}_t(u^*|Z^M_{\beta N+1:N}) - \mathcal{L}'_{g,t}(u^*)]$$

$$+ \mathcal{L}'_{g,t}(u^*) - \mathcal{L}'_g(u^*)$$

$$= \alpha\mathbb{E}_{P_{Z^M_{1:\beta N}}}[\mathcal{L}_t(u^*|Z^M_{1:\beta N}) - \mathcal{L}'_{g,t}(u^*)] + \mathcal{L}'_{g,t}(u^*) - \mathcal{L}'_g(u^*) \tag{71}$$

where the last equality follows since $\mathbb{E}_{P'_{Z^M_{\beta N+1:N}}}[\mathcal{L}_t(u^*|Z^M_{\beta N+1:N})] = \mathcal{L}'_{g,t}(u^*)$. We now separately bound the two differences in (71).

To bound the first difference in (71), note that

$$\mathbb{E}_{P_{Z^M_{1:\beta N}}}[\mathcal{L}_t(u^*|Z^M_{1:\beta N}) - \mathcal{L}'_{g,t}(u^*)] = \mathbb{E}_{P_{Z^M}}[L_t(u^*|Z^M)] - \mathbb{E}_{P'_{Z^M}}[L_t(u^*|Z^M)].$$

To bound this term, we resort to the inequality (67) which is a consequence of Assumption 3.3 (note that we can ignore the subscript i in the current context), and fix $u = u^*$. Applying change of measure from P'_{Z^M} to P_{Z^M} then yields the following inequality,

$$\mathbb{E}_{P_{Z^M}}\left[\exp\left(\lambda(L_t(u^*|Z^M) - \mathbb{E}_{P'_{Z^M}}[L_t(u^*|Z^M)] - \log\frac{P_{Z^M}(Z^M)}{P'_{Z^M}(Z^M)} - \frac{\lambda^2\sigma^2}{2}\right)\right] \leq 1, \tag{72}$$

which holds for all $\lambda \in \mathbb{R}$. Applying Jensen's inequality and choosing $\lambda = \sqrt{2D(P_{Z^M}||P'_{Z^M})}/\sigma$ then gives that

$$\mathbb{E}_{P_{Z^M}}[L_t(u^*|Z^M)] - \mathbb{E}_{P'_{Z^M}}[L_t(u^*|Z^M)] \leq \sqrt{2\sigma^2 D(P_{Z^M}||P'_{Z^M})}. \tag{73}$$

We now bound the second difference in (71). Towards this, note that the following set of relations hold,

$$\mathcal{L}'_{g,t}(u^*) - \mathcal{L}'_g(u^*) = \mathbb{E}_{P'_T}\mathbb{E}_{P^M_{Z|T}}\mathbb{E}_{P_{W|Z^M,u^*}}[L_t(W|Z^M) - L_g(W|T)]$$

$$= \mathbb{E}_{P'_T}\left[\frac{1}{M}\sum_{j=1}^{M}\left(\mathbb{E}_{P_{W,Z_j|u^*,T=\tau}}[l(W,Z_j)] - \mathbb{E}_{P_{W|u^*,T=\tau}P_{Z_j|T=\tau}}[l(W,Z_j)]\right)\right]. \tag{74}$$

To bound the difference $\mathbb{E}_{P_{W,Z_j|u^*,T=\tau}}[l(W,Z_j)] - \mathbb{E}_{P_{W|u^*,T=\tau}P_{Z_j|T=\tau}}[l(W,Z_j)]$, we slightly modify the exponential inequality (62) in Lemma D.1. Towards this, we average the inequality (65) with respect to $W \sim P_{W|\tau,u^*}$, where $P_{W|\tau,u^*}$ is the marginal of the joint $P_{W|Z^M,u^*}P_{Z^M|\tau}$, and

subsequently perform a change of measure from $P_{Z_j|\tau}P_{W|\tau,u^*}$ to $P_{W,Z_j|\tau,u^*}$. This results in the following modified form of (62)

$$\mathbb{E}_{P_{W,Z_j|\tau,u^*}}\left[\exp\left(\lambda(l(W,Z_j)-\mathbb{E}_{P_{Z_j|\tau}}[l(W,Z_j)]-\imath(W,Z_j|T=\tau,u^*)-\frac{\lambda^2\delta_\tau^2}{2}\right)\right]\leq 1. \tag{75}$$

Now, applying Jensen's inequality, and choosing $\lambda=\sqrt{2I(W;Z_j|T=\tau,u^*)}/\delta_\tau$ gives that

$$\mathbb{E}_{P_{W,Z_j|\tau,u^*}}[l(W,Z_j)]-\mathbb{E}_{P_{W|\tau,u^*}P_{Z_j|\tau}}[l(W,Z_j)]\leq\sqrt{2\delta_\tau^2 I(W;Z_j|T=\tau,u^*)}. \tag{76}$$

Substituting this in (74), and using the resulting inequality together with (73) in (71) yields the required bound.

APPENDIX F
EXPONENTIAL INEQUALITIES FOR PAC-BAYESIAN AND SINGLE-DRAW PROBABILITY BOUNDS

We now present two exponential inequalities that are crucial to the derivation of high-probability PAC-Bayesian and high-probability single-draw bounds. Towards this, we first define the following *mismatched information densities*

$$\jmath(U,Z_{1:N}^M)=\log\frac{P_{U|Z_{1:N}^M}(U|Z_{1:N}^M)}{Q_U(U)},\quad \jmath(W,Z^M|U)=\log\frac{P_{W|Z^M,U}(W|Z^M,U)}{Q_{W|U}(W|U)} \tag{77}$$

where $Q_U\in\mathcal{P}(\mathcal{U})$ represents an arbitrary data-independent hyper-prior over the space of hyperparameters, and $Q_{W|U=u}\in\mathcal{P}(\mathcal{W})$ represents a class of arbitrary data-independent priors over the space of model parameters for each $u\in\mathcal{U}$. The mismatched information density $\jmath(U,Z_{1:N}^M)$ quantifies the evidence for the hyperparameter U to be generated according to the meta-learner $P_{U|Z_{1:N}^M}$ based on meta-training set, rather than being generated according to the hyper-prior distribution Q_U. Similarly, the density $\jmath(W,Z^M|U)$ quantifies the evidence of the model parameter W being generated by the base learner $P_{W|Z^M,U}$ based on the training set Z^M, rather than being generated according to the prior.

We denote $Z_{1:N/i}^M:=(Z_1^M,\ldots,Z_{i-1}^M,Z_{i+1}^M,\ldots,Z_N^M)$, for $i=1,\ldots,N$, to be the meta-training set without the ith subset and is distributed according to $P_{Z_{1:N/i}^M}$ which is obtained by marginalizing $P_{Z_{1:N}^M}$.

Lemma F.1: Under Assumption 4.1(a) and Assumption 4.2, the following exponential inequality holds for the ith sub-set, $Z_i^M \sim P_{Z|T=T_i}^M$, of the meta-training set $Z_{1:N}^M = (Z_i^M, Z_{1:N/i}^M)$ for $i = 1, \ldots, N$,

$$\mathbb{E}_{P_{Z_{1:N/i}^M}} \mathbb{E}_{P_{Z|T=T_i}^M P_{U|Z_{1:N}^M} P_{W|Z_i^M,U}} \Bigg[\exp\Bigg(\lambda(L_t(W|Z_i^M) - L_g(W|T_i)) - \frac{\lambda^2 \delta_{T_i}^2}{2M} - \jmath(W, Z_i^M|U) - \jmath(U, Z_{1:N}^M) \Bigg) \Bigg] \leq 1. \tag{78}$$

Proof: From Assumption 4.1(a), we have that for task $T = T_i$, $L_t(w|Z_i^M)$ is the average of M independent $\delta_{T_i}^2$-sub-Gaussian random variables $l(w, Z_i)$. It is then easy to see that $L_t(w|Z_i^M)$ is $\delta_{T_i}^2/M$-sub-Gaussian under $Z_i^M \sim P_{Z|T_i}^M$ for all $w \in \mathcal{W}$. This can be equivalently expressed as

$$\mathbb{E}_{P_{Z|T=T_i}^M} \left[\exp\left(\lambda(L_t(w|Z_i^M) - L_g(w|T_i)) - \frac{\lambda^2 \delta_{T_i}^2}{2M} \right) \right] \leq 1 \tag{79}$$

which holds for all $w \in \mathcal{W}$ and $\lambda \in \mathbb{R}$. Averaging both sides with respect to $Z_{1:N/i}^M$ gives that

$$\mathbb{E}_{P_{Z_{1:N/i}^M}} \mathbb{E}_{P_{Z|T=T_i}^M} \left[\exp\left(\lambda(L_t(w|Z_i^M) - L_g(w|T_i)) - \frac{\lambda^2 \delta_{T_i}^2}{2M} \right) \right] \leq 1 \tag{80}$$

for all $w \in \mathcal{W}$. To get to the inequality (78), we consider (80) as a function of both model parameter w and hyperparameter u. Subsequently, average both sides of inequality (80) with respect to $Q_{W,U} = Q_U Q_{W|U} \in \mathcal{P}(\mathcal{W} \times \mathcal{U})$. We now follow the approach of [49, Prop. 17.1] and apply a change of measure as detailed below. Towards this, we first note that average over $Q_{W,U}$ on (80) implies the following inequality

$$\mathbb{E}_{P_{Z_{1:N/i}^M}} \mathbb{E}_{P_{Z|T=T_i}^M} \mathbb{E}_{Q_{W,U}} \left[\mathbb{I}_{\mathcal{E}(Z_i^M, Z_{1:N/i}^M)} \exp\left(\lambda(L_t(W|Z_i^M) - L_g(W|T_i)) - \frac{\lambda^2 \delta_{T_i}^2}{2M} \right) \right] \leq 1, \tag{81}$$

where $\mathcal{E}(z_i^M, z_{1:N/i}^M) = \text{supp}(P_{W,U|z_i^M, z_{1:N/i}^M})$ and $P_{W,U|Z_i^M, Z_{1:N/i}^M} = P_{U|Z_{1:N}^M} P_{W|U,Z_i^M}$. It is then easy to see that for $Z_i^M = z_i^M$, $Z_{1:N/i}^M = z_{1:N/i}^M$, the following relation holds

$$\mathbb{E}_{Q_{W,U}} \left[\mathbb{I}_{\mathcal{E}(z_i^M, z_{1:N/i}^M)} \exp\left(\lambda(L_t(W|z_i^M) - L_g(W|T_i)) - \frac{\lambda^2 \delta_{T_i}^2}{2M} \right) \right] \tag{82}$$

$$= \mathbb{E}_{P_{W,U|z_i^M, z_{1:N/i}^M}} \left[\exp\left(\lambda(L_t(W|z_i^M) - L_g(W|T_i)) - \frac{\lambda^2 \delta_{T_i}^2}{2M} - \log \frac{P_{W,U|z_i^M, z_{1:N/i}^M}(W,U)}{Q_{W,U}(W,U)} \right) \right]. \tag{83}$$

Using this in (81) and averaging over $Z_i^M, Z_{1:N/i}^M$ then yields inequality (78) with

$$\log \frac{P_{W,U|Z_i^M, Z_{1:N/i}^M}(W,U|Z_i^M, Z_{1:N/i}^M)}{Q_{W,U}(W,U)} = \jmath(W, Z_i^M|U) + \jmath(U, Z_{1:N}^M).$$

■

Inequality (78) relates the per-task training loss to the per-task generalization loss and the information densities $\jmath(W, Z_i^M|U)$, $\jmath(U, Z_{1:N}^M)$. We will use this to bound the contribution of within-task generalization gap to transfer meta-generalization gap. Assumption 4.1(b) then provides the following exponential inequality on the difference $\mathcal{L}_{t,g}(u|Z_{1:N}^M, T_{1:N}) - \mathcal{L}'_g(u)$.

Lemma F.2: Under Assumption 4.1(b) and Assumption 4.2, the following exponential inequality holds

$$\mathbb{E}_{P_{T_{1:N}} P_{Z_{1:N}^M|T_{1:N}}} \mathbb{E}_{P_{U|Z_{1:N}^M}} \Bigg[\exp\Bigg(\lambda \Bigg(\mathcal{L}_{t,g}(U|T_{1:N}, Z_{1:N}^M) - \mathcal{L}'_g(U) \Bigg) - \frac{\lambda^2 \alpha^2 \sigma^2}{2\beta N} - \sum_{i=1}^{\beta N} \log \frac{P_T(T_i)}{P'_T(T_i)} - \frac{\lambda^2 (1-\alpha)^2 \sigma^2}{2(1-\beta)N} - \jmath(U, Z_{1:N}^M) \Bigg) \Bigg] \leq 1. \tag{84}$$

Proof: In the following, we denote $T_{1:\beta N} := (T_1, \ldots, T_{\beta N})$, $T_{\beta N+1:N} := (T_{\beta N+1}, \ldots, T_N)$, the empirical average per-task test loss of the source environment data set as

$$\mathcal{L}_{t,g}(u|T_{1:\beta N}, Z_{1:\beta N}^M) = \frac{1}{\beta N} \sum_{i=1}^{\beta N} L_g(u|Z_i^M, T_i),$$

and the empirical average per-task test loss of the target environment data set as

$$\mathcal{L}_{t,g}(u|T_{\beta N+1:N}, Z_{\beta N+1:N}^M) = \frac{1}{(1-\beta)N} \sum_{i=\beta N+1}^{N} L_g(u|Z_i^M, T_i).$$

From Assumption 4.1(b), we get that $\mathcal{L}_{t,g}(u|T_{1:\beta N}, Z_{1:\beta N}^M)$ is the average of i.i.d. σ^2-sub-Gaussian random variables under $(T_i, Z_i^M) \sim P'_{T_i} P_{Z|T_i}^M$. Consequently, it is $\sigma^2/\beta N$-sub-Gaussian when $(T_{1:\beta N}, Z_{1:\beta N}^M) \sim P'_{T_{1:\beta N}} P_{Z_{1:\beta N}^M|T_{1:\beta N}}$ for all $u \in \mathcal{U}$. Note here that we use $P'_{T_{1:\beta N}} P_{Z_{1:\beta N}^M|T_{1:\beta N}}$ to denote the product distribution $\prod_{i=1}^{\beta N} P'_{T_i} P_{Z|T_i}^M$. Similarly, $\mathcal{L}_{t,g}(u|T_{\beta N+1:N}, Z_{\beta N+1:N}^M)$ is $\sigma^2/(1-\beta)N$-sub-Gaussian under $(T_{\beta N+1:N}, Z_{\beta N+1:N}^M) \sim P'_{T_{\beta N+1:N}} P_{Z_{\beta N+1:N}^M|T_{\beta N+1:N}}$ for all $u \in \mathcal{U}$. Here, $P'_{T_{\beta N+1:N}} P_{Z_{\beta N+1:N}^M|T_{\beta N+1:N}}$ denotes the product distribution $\prod_{i=\beta N+1}^{N} P'_{T_i} P_{Z|T_i}^M$. Denoting $P'_{T_{1:N}} = \prod_{i=1}^{N} P'_{T_i}$, the following set of relations then follow from the sub-Gaussianity assumptions discussed above, and holds for all $u \in \mathcal{U}$ and $\lambda \in \mathbb{R}$:

$$\mathbb{E}_{P'_{T_{1:N}} P_{Z_{1:N}^M|T_{1:N}}} \Bigg[\exp\Bigg(\lambda (\mathcal{L}_{t,g}(u|T_{1:N}, Z_{1:N}^M) - \mathcal{L}'_g(u)) \Bigg) \Bigg] \tag{85}$$

$$= \mathbb{E}_{P'_{T_{1:\beta N}} P_{Z_{1:\beta N}^M|T_{1:\beta N}}} \Bigg[\exp\Bigg(\lambda\alpha \Bigg(\mathcal{L}_{t,g}(u|T_{1:\beta N}, Z_{1:\beta N}^M) - \mathcal{L}'_g(u) \Bigg) \Bigg) \Bigg] \times \mathbb{E}_{P'_{T_{\beta N+1:N}} P_{Z_{\beta N+1:N}^M|T_{\beta N+1:N}}} \Bigg[\exp\Bigg(\lambda(1-\alpha) \Bigg(\mathcal{L}_{t,g}(u|T_{\beta N+1:N}, Z_{\beta N+1:N}^M) - \mathcal{L}'_g(u) \Bigg) \Bigg) \Bigg]$$

$$\leq \exp\Bigg(\frac{\lambda^2 \alpha^2 \sigma^2}{2\beta N} \Bigg) \exp\Bigg(\frac{\lambda^2 (1-\alpha)^2 \sigma^2}{2(1-\beta)N} \Bigg). \tag{86}$$

This in turn implies that

$$\mathbb{E}_{P'_{T_{\beta N+1:N}}P_{Z^M_{\beta N+1:N}|T_{\beta N+1:N}}}\mathbb{E}_{P'_{T_{1:\beta N}}P_{Z^M_{1:\beta N}|T_{1:\beta N}}}\Bigg[\mathbb{I}_{\mathcal{E}}\exp\Bigg(\lambda(\mathcal{L}_{t,g}(u|T_{1:N},Z^M_{1:N})-\mathcal{L}'_g(u))-\frac{\lambda^2\alpha^2\sigma^2}{2\beta N} - \frac{\lambda^2(1-\alpha)^2\sigma^2}{2(1-\beta)N}\Bigg)\Bigg]\leq 1 \tag{87}$$

where $\mathcal{E}=\operatorname{supp}(P_{T_{1:\beta N}}P_{Z^M_{1:\beta N}|T_{1:\beta N}})$. Applying change of measure from $P'_{T_{1:\beta N}}P_{Z^M_{1:\beta N}|T_{1:\beta N}}$ to $P_{T_{1:\beta N}}P_{Z^M_{1:\beta N}|T_{1:\beta N}}$, then yields

$$\mathbb{E}_{P_{T_{1:\beta N}}P'_{T_{\beta N+1:N}}P_{Z^M_{1:N}|T_{1:N}}}\Bigg[\exp\Bigg(\lambda\Bigg(\mathcal{L}_{t,g}(u|T_{1:N},Z^M_{1:N})-\mathcal{L}'_g(u)\Bigg)-\frac{\lambda^2\alpha^2\sigma^2}{2\beta N} - \log\frac{P_{T_{1:\beta N}}(T_{1:\beta N})}{P'_{T_{1:\beta N}}(T_{1:\beta N})}-\frac{\lambda^2(1-\alpha)^2\sigma^2}{2(1-\beta)N}\Bigg)\Bigg]\leq 1, \tag{88}$$

which holds for all $u\in\mathcal{U}$. Average both sides of the inequality with respect to $Q_U\in\mathcal{P}(\mathcal{U})$. The resultant inequality implies the following

$$\mathbb{E}_{P_{T_{1:\beta N}}P'_{T_{\beta N+1:N}}P_{Z^M_{1:N}|T_{1:N}}}\mathbb{E}_{Q_U}\Bigg[\mathbb{I}_{\mathcal{E}(Z^M_{1:N})}\exp\Bigg(\lambda\Bigg(\mathcal{L}_{t,g}(U|T_{1:N},Z^M_{1:N})-\mathcal{L}'_g(u)\Bigg)-\frac{\lambda^2\alpha^2\sigma^2}{2\beta N} - \log\frac{P_{T_{1:\beta N}}(T_{1:\beta N})}{P'_{T_{1:\beta N}}(T_{1:\beta N})}-\frac{\lambda^2(1-\alpha)^2\sigma^2}{2(1-\beta)N}\Bigg)\Bigg]\leq 1, \tag{89}$$

where $\mathcal{E}(z^M_{1:N})=\operatorname{supp}(P_{U|Z^M_{1:N}=z^M_{1:N}})$. Applying change of measure from Q_U to $P_{U|Z^M_{1:N}}$ together with $\log\Big(P_{T_{1:\beta N}}(T_{1:\beta N})/P'_{T_{1:\beta N}}(T_{1:\beta N})\Big)=\sum_{i=1}^{\beta N}\log(P_T(T_i)/P'_T(T_i))$ then gives the required inequality (84). ■

The inequality (84) relates the difference between weighted average per-task test loss and transfer meta-generalization loss, $\mathcal{L}_{t,g}(U|T_{1:N},Z^M_{1:N})-\mathcal{L}'_g(U)$, to the mismatched information density $\jmath(U,Z^M_{1:N})$ and the log-likelihood ratio $\log(P_T(T_i)/P'_T(T_i))$, that captures the meta-environment shift in task distributions.

APPENDIX G
PROOF OF THEOREM 4.1

To obtain the required PAC-Bayesian bound, we use the decomposition (31). The idea is to separately bound the two differences in (31) in high probability over $(T_{1:N},Z^M_{1:N})$, and subsequently combine the bounds via union bound.

To start, we bound the first difference in (31). Towards this, we resort to the exponential inequality (84). Applying Jensen's inequality with respect to just $P_{U|Z_{1:N}^M}$ on (84) results in

$$\mathbb{E}_{P_{T_{1:N}}P_{Z_{1:N}^M|T_{1:N}}}\Bigg[\exp\Bigg(\lambda\mathbb{E}_{P_{U|Z_{1:N}^M}}\Big[\mathcal{L}_{t,g}(U|T_{1:N},Z_{1:N}^M)-\mathcal{L}_g'(U)\Big]-\frac{\lambda^2\alpha^2\sigma^2}{2\beta N} \\ -\sum_{i=1}^{\beta N}\log\frac{P_T(T_i)}{P_T'(T_i)}-\frac{\lambda^2(1-\alpha)^2\sigma^2}{2(1-\beta)N}-D(P_{U|Z_{1:N}^M}||Q_U)\Bigg)\Bigg]\leq 1. \tag{90}$$

Take $V=\exp\big(\lambda\mathbb{E}_{P_{U|Z_{1:N}^M}}\big[\mathcal{L}_{t,g}(U|T_{1:N},Z_{1:N}^M)-\mathcal{L}_g'(U)\big]-\lambda^2\alpha^2\sigma^2/(2\beta N)-\sum_{i=1}^{\beta N}\log(P_T(T_i)/P_T'(T_i))-\lambda^2(1-\alpha)^2\sigma^2/(2(1-\beta)N)-D(P_{U|Z_{1:N}^M}||Q_U)\big)$. Applying Markov's inequality of the form $\mathbb{P}[V\geq\frac{1}{\delta_0}]\leq\delta_0\mathbb{E}[V]\leq\delta_0$ then gives that with probability at least $1-\delta_0$ over $(Z_{1:N}^M,T_{1:N})\sim P_{T_{1:N}}P_{Z_{1:N}^M|T_{1:N}}$ we have $V\leq\frac{1}{\delta_0}$. Taking logarithm on both sides of the inequality then results in

$$\lambda\Bigg(\mathbb{E}_{P_{U|Z_{1:N}^M}}\Big[\mathcal{L}_{t,g}(U|T_{1:N},Z_{1:N}^M)-\mathcal{L}_g'(U)\Big]\Bigg)\leq\frac{\lambda^2\alpha^2\sigma^2}{2\beta N}+\frac{\lambda^2(1-\alpha)^2\sigma^2}{2(1-\beta)N} \\ +\sum_{i=1}^{\beta N}\log\frac{P_T(T_i)}{P_T'(T_i)}+D(P_{U|Z_{1:N}^M}||Q_U)+\log\frac{1}{\delta_0} \tag{91}$$

which when optimized over λ with the choice

$$\lambda=\sqrt{\frac{\sum_{i=1}^{\beta N}\log\frac{P_T(T_i)}{P_T'(T_i)}+D(P_{U|Z_{1:N}^M}||Q_U)+\log\frac{1}{\delta_0}}{\frac{\alpha^2\sigma^2}{2\beta N}+\frac{(1-\alpha)^2\sigma^2}{2(1-\beta)N}}}$$

yields

$$\Bigg|\mathbb{E}_{P_{U|Z_{1:N}^M}}\Big[\mathcal{L}_{t,g}(u|T_{1:N},Z_{1:N}^M)-\mathcal{L}_g'(u)\Big]\Bigg| \\ \leq\sqrt{2\sigma^2\Bigg(\frac{\alpha^2}{\beta N}+\frac{(1-\alpha)^2}{(1-\beta)N}\Bigg)\Bigg(\sum_{i=1}^{\beta N}\log\frac{P_T(T_i)}{P_T'(T_i)}+D(P_{U|Z_{1:N}^M}||Q_U)+\log\frac{1}{\delta_0}\Bigg)}. \tag{92}$$

We now bound the second difference in (31), which can be equivalently written as

$$\mathbb{E}_{P_{U|Z_{1:N}^M}}\Big[\mathcal{L}_{t,g}(U|Z_{1:N}^M,T_{1:N})-\mathcal{L}_t(U|Z_{1:N}^M)\Big] \\ =\mathbb{E}_{P_{U|Z_{1:N}^M}}\Bigg[\frac{\alpha}{\beta N}\sum_{i=1}^{\beta N}(L_g(U|Z_i^M,T_i)-L_t(U|Z_i^M))+\frac{1-\alpha}{(1-\beta)N}\sum_{i=\beta N+1}^{N}(L_g(U|Z_i^M,T_i)-L_t(U|Z_i^M))\Bigg]. \tag{93}$$

The idea then is to bound each of the terms $\mathbb{E}_{P_{U|Z_{1:N}^M}}[L_g(U|Z_i^M,T_i)-L_t(U|Z_i^M)]$ separately with probability at least $(1-\delta_i)$ over $(Z_{1:N/i}^M,T_i,Z_i^M)\sim P_{Z_{1:N/i}^M}P_{T_i}P_{Z|T_i}^M$ for $i=1,\ldots,\beta N$ and over $(Z_{1:N/i}^M,T_i,Z_i^M)\sim P_{Z_{1:N/i}^M}P_{T_i}'P_{Z|T_i}^M$ for $i=\beta N+1,\ldots,N$. Towards this, we resort to the

exponential inequality (78) and apply Jensen's inequality with respect to $P_{U|Z_{1:N}^M} P_{W|Z_i^M,U}$. This results in

$$\mathbb{E}_{P_{Z_{1:N/i}^M}} \mathbb{E}_{P_{Z|T_i}^M} \Bigg[\underbrace{\exp\Bigg(\lambda \mathbb{E}_{P_{U|Z_{1:N}^M}} [L_t(U|Z_i^M) - L_g(U|Z_i^M, T_i)] - \frac{\lambda^2 \delta_{T_i}^2}{2M} - D(P_{W,U|Z_{1:N}^M, Z_i^M} || Q_{W,U}) \Bigg)}_{V} \Bigg]$$

$$\leq 1, \tag{94}$$

which holds for $i = 1, \ldots, N$. Note that the above inequality holds even after averaging both sides of the inequality with respect to P_{T_i} (or P'_{T_i}). Applying Markov's inequality of the form $\mathbb{P}[V \geq \frac{1}{\delta_i}] \leq \beta_0 \mathbb{E}[V] \leq \delta_i$ then gives that with probability at least $1 - \delta_i$ over $(Z_{1:N/i}^M, T_i, Z_i^M) \sim P_{Z_{1:N/i}^M} P_{T_i} P_{Z|T_i}^M$ for $i = 1, \ldots, \beta N$ and over $(Z_{1:N/i}^M, T_i, Z_i^M) \sim P_{Z_{1:N/i}^M} P'_{T_i} P_{Z|T_i}^M$ for $i = \beta N + 1, \ldots, N$

$$\lambda \Bigg(\mathbb{E}_{P_{U|Z_{1:N}^M}} [L_t(U|Z_i^M) - L_g(U|Z_i^M, T_i)] \Bigg)$$
$$\leq \frac{\lambda^2 \delta_{T_i}^2}{2M} + D(P_{U|Z_{1:N}^M} || Q_U) + \mathbb{E}_{P_{U|Z_{1:N}^M}} [D(P_{W|U,Z_i^M} || Q_{W|U})] + \log \frac{1}{\delta_i}. \tag{95}$$

Now, choosing

$$\lambda = \sqrt{\frac{D(P_{U|Z_{1:N}^M} || Q_U) + \mathbb{E}_{P_{U|Z_{1:N}^M}} [D(P_{W|U,Z_i^M} || Q_{W|U})] + \log \frac{1}{\delta_i}}{\frac{\delta_{T_i}^2}{2M}}}$$

then results in

$$\Bigg| \mathbb{E}_{P_{U|Z_{1:N}^M}} [L_t(U|Z_i^M) - L_g(U|Z_i^M, T_i)] \Bigg|$$
$$= \sqrt{\frac{2\delta_{T_i}^2}{M} \Bigg(D(P_{U|Z_{1:N}^M} || Q_U) + \mathbb{E}_{P_{U|Z_{1:N}^M}} [D(P_{W|U,Z_i^M} || Q_{W|U})] + \log \frac{1}{\delta_i} \Bigg)}. \tag{96}$$

Choosing $\delta_0 = \frac{\delta}{2}$ and $\delta_i = \frac{\delta}{4\beta N}$ for $i = 1, \ldots, \beta N$ and $\delta_i = \frac{\delta}{4(1-\beta)N}$ for $i = \beta N + 1, \ldots, N$, and combining the bounds (92) and (96) in (93) via union bound then yields the bound (32).

APPENDIX H
PROOF OF THEOREM 5.1

To obtain the required single-draw bound, we use the decomposition (31). We start by bounding the first difference in (31) without the expectation over meta-training algorithm. Towards this, we resort to the exponential inequality (84). Take

$$V = \exp\Bigg(\lambda \big(\mathcal{L}_{t,g}(U|T_{1:N}, Z_{1:N}^M) - \mathcal{L}'_g(U) \big) - \frac{\lambda^2 \alpha^2 \sigma^2}{2\beta N} - \sum_{i=1}^{\beta N} \log \frac{P_T(T_i)}{P'_T(T_i)} - \frac{\lambda^2 (1-\alpha)^2 \sigma^2}{2(1-\beta)N} - \jmath(U, Z_{1:N}^M) \Bigg).$$

Applying Markov's inequality of the form $\mathbb{P}[V \geq \frac{1}{\delta_0}] \leq \delta_0 \mathbb{E}[V] \leq \delta_0$ then gives that with probability at least $1-\delta_0$ over $(T_{1:N}, Z_{1:N}^M, U) \sim P_{T_{1:N}} P_{Z_{1:N}^M|T_{1:N}} P_{U|Z_{1:N}^M}$ we have $V \leq \frac{1}{\delta_0}$. Taking logarithm on both sides of the inequality then results in

$$\lambda\bigg(\mathcal{L}_{t,g}(U|T_{1:N}, Z_{1:N}^M) - \mathcal{L}'_g(U)\bigg) \leq \frac{\lambda^2\alpha^2\sigma^2}{2\beta N} + \frac{\lambda^2(1-\alpha)^2\sigma^2}{2(1-\beta)N} + \sum_{i=1}^{\beta N} \log \frac{P_T(T_i)}{P'_T(T_i)} + \jmath(U, Z_{1:N}^M) + \log \frac{1}{\delta_0} \tag{97}$$

which when optimized over λ yields

$$\bigg|\mathcal{L}_{t,g}(u|T_{1:N}, Z_{1:N}^M) - \mathcal{L}'_g(u)\bigg| \leq \sqrt{2\sigma^2\bigg(\frac{\alpha^2}{\beta N} + \frac{(1-\alpha)^2}{(1-\beta)N}\bigg)\bigg(\sum_{i=1}^{\beta N} \log \frac{P_T(T_i)}{P'_T(T_i)} + \jmath(U, Z_{1:N}^M) + \log \frac{1}{\delta_0}\bigg)}. \tag{98}$$

We now bound the second difference in (31), which can be equivalently written as

$$\mathcal{L}_{t,g}(U|Z_{1:N}^M, T_{1:N}) - \mathcal{L}_t(U|Z_{1:N}^M) = \frac{\alpha}{\beta N}\sum_{i=1}^{\beta N}(L_g(U|Z_i^M, T_i) - L_t(U|Z_i^M)) + \frac{1-\alpha}{(1-\beta)N}\sum_{i=\beta N+1}^{N}(L_g(U|Z_i^M, T_i) - L_t(U|Z_i^M)). \tag{99}$$

We now bound each of the terms $L_g(U|Z_i^M, T_i) - L_t(U|Z_i^M)$ separately with probability at least $(1-\delta_i)$ over $(Z_{1:N/i}^M, T_i, Z_i^M, U) \sim P_{Z_{1:N/i}^M} P_{T_i} P_{Z|T_i}^M P_{U|Z_{1:N}^M}$ for $i = 1, \ldots, \beta N$ and over $(Z_{1:N/i}^M, T_i, Z_i^M, U) \sim P_{Z_{1:N/i}^M} P'_{T_i} P_{Z|T_i}^M P_{U|Z_{1:N}^M}$ for $i = \beta N+1, \ldots, N$. Towards this, we resort to the exponential inequality (78) and apply Jensen's inequality with respect to $P_{W|Z_i^M,U}$. This results in

$$\mathbb{E}_{P_{Z_{1:N/i}^M}} \mathbb{E}_{P_{Z|T_i}^M P_{U|Z_{1:N}^M}} \Bigg[\underbrace{\exp\bigg(\lambda(L_t(U|Z_i^M) - L_g(U|Z_i^M, T_i)] - \frac{\lambda^2\delta_{T_i}^2}{2M} - D(P_{W|U,Z_i^M}||Q_{W|U}) - \jmath(U, Z_{1:N}^M)\bigg)}_{V}\Bigg] \leq 1, \tag{100}$$

which holds for $i = 1, \ldots, N$. Note that the above inequality holds even after averaging both sides of the inequality with respect to P_{T_i} (or P'_{T_i}). Applying Markov's inequality of the form $\mathbb{P}[V \geq \frac{1}{\delta_i}] \leq \beta_0 \mathbb{E}[V] \leq \delta_i$ then gives that with probability at least $(1-\delta_i)$ over $(Z_{1:N/i}^M, T_i, Z_i^M, U) \sim P_{Z_{1:N/i}^M} P_{T_i} P_{Z|T_i}^M P_{U|Z_{1:N}^M}$ for $i = 1, \ldots, \beta N$ and over $(Z_{1:N/i}^M, T_i, Z_i^M, U) \sim P_{Z_{1:N/i}^M} P'_{T_i} P_{Z|T_i}^M P_{U|Z_{1:N}^M}$ for $i = \beta N+1, \ldots, N$, the following inequality holds,

$$\lambda(L_t(U|Z_i^M) - L_g(U|Z_i^M, T_i)) \leq \frac{\lambda^2\delta_{T_i}^2}{2M} + D(P_{W|U,Z_i^M}||Q_{W|U}) + \jmath(U, Z_{1:N}^M) + \log \frac{1}{\delta_i}. \tag{101}$$

Optimizing over λ then results in

$$\begin{aligned}&\left|L_t(U|Z_i^M) - L_g(U|Z_i^M, T_i)\right| \\ &\leq \sqrt{\frac{2\delta_{T_i}^2}{M}\left(D(P_{W|U,Z_i^M}||Q_{W|U}) + \jmath(U, Z_{1:N}^M) + \log\frac{1}{\delta_i}\right)}. \end{aligned} \tag{102}$$

Choosing $\delta_0 = \frac{\delta}{2}$ and $\delta_i = \frac{\delta}{4\beta N}$ for $i = 1, \ldots, \beta N$ and $\delta_i = \frac{\delta}{4(1-\beta)N}$ for $i = \beta N + 1, \ldots, N$, and combining the bounds (98) and (102) via union bound then yields the bound (41).

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, "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.

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

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

[5] L. Collins, A. Mokhtari, and S. Shakkottai, "Task-Robust Model-Agnostic Meta-Learning," *arXiv preprint 2002.04766*, 2020.

[6] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of Representations for Domain Adaptation," in *Advances in Neural Information Processing Systems*, 2007, pp. 137–144.

[7] S. Bickel, M. Brückner, and T. Scheffer, "Discriminative Learning for Differing Training and Test Distributions," in *Proceedings of the ICML*, 2007, pp. 81–88.

[8] J. Blitzer, R. McDonald, and F. Pereira, "Domain Adaptation with Structural Correspondence Learning," in *Proceedings of the 2006 conference on empirical methods in natural language processing*, 2006, pp. 120–128.

[9] F. Hellström and G. Durisi, "Generalization Bounds via Information Density and Conditional Information Density," *arXiv preprint arXiv:2005.08044*, 2020.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] T. Steinke and L. Zakynthinou, "Reasoning About Generalization via Conditional Mutual Information," vol. 125, pp. 3437–3452, 09–12 Jul 2020.

[15] X. Wu, J. H. Manton, U. Aickelin, and J. Zhu, "Information-Theoretic Analysis for Transfer Learning," *arXiv preprint arXiv:2005.08697*, 2020.

[16] S. T. Jose and O. Simeone, "Information-Theoretic Generalization Bounds for Meta-Learning and Applications," *arXiv preprint arXiv:2005.04372*, 2020.

[17] 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.

[18] 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.

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

[20] G. K. Dziugaite, K. Hsu, W. Gharbieh, and D. M. Roy, "On the Role of Data in PAC-Bayes Bounds," 2020.

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

[22] D. A. McAllester, "PAC-Bayesian Stochastic Model Selection," *Machine Learning*, vol. 51, no. 1, pp. 5–21, 2003.

[23] A. Maurer, "A Note on the PAC-Bayesian Theorem," *arXiv preprint cs/0411099*, 2004.

[24] 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.

[25] 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.

[26] 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.

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

[28] P. Germain, A. Habrard, F. Laviolette, and E. Morvant, "PAC-Bayes and Domain Adaptation," *arXiv preprint arXiv:1707.05712*, 2017.

[29] T. Zhang, "Information-Theoretic Upper and Lower Bounds for Statistical Estimation," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1307–1321, 2006.

[30] 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.

[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] A. R. Esposito, M. Gastpar, and I. Issa, "Generalization Error Bounds via Rényi-, f-Divergences and Maximal leakage," *arXiv preprint arXiv:1912.01439*, 2019.

[33] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman, "Learning Bounds for Domain Adaptation," in *Advances in Neural Information Processing Systems*, 2008, pp. 129–136.

[34] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Domain Adaptation: Learning Bounds and Algorithms," *arXiv preprint arXiv:0902.3430*, 2009.

[35] C. Zhang, L. Zhang, and J. Ye, "Generalization Bounds for Domain Adaptation," in *Advances in Neural Information Processing Systems*, 2012, pp. 3320–3328.

[36] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A Theory of Learning from Different Domains," *Machine Learning*, vol. 79, no. 1-2, pp. 151–175, 2010.

[37] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Multiple Source Adaptation and the Rényi Divergence," *arXiv preprint arXiv:1205.2628*, 2012.

[38] P. Germain, A. Habrard, F. Laviolette, and E. Morvant, "A PAC-Bayesian Approach for Domain Adaptation with Specialization to Linear Classifiers," in *International Conference on Machine Learning*, 2013, pp. 738–746.

[39] J. Hoffman, M. Mohri, and N. Zhang, "Algorithms and Theory for Multiple-Source Adaptation," in *Advances in Neural Information Processing Systems*, 2018, pp. 8246–8256.

[40] I. Redko, A. Habrard, and M. Sebban, "Theoretical Analysis of Domain Adaptation with Optimal Transport," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2017, pp. 737–753.

[41] 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.

[42] B. Guedj, "A Primer on PAC-Bayesian Learning," *arXiv preprint arXiv:1901.05353*, 2019.

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

[44] G. Denevi, M. Pontil, and C. Ciliberto, "The Advantage of Conditional Meta-Learning for Biased Regularization and Fine-Tuning," *arXiv preprint arXiv:2008.10857*, 2020.

[45] M. Raginsky, A. Rakhlin, and M. Telgarsky, "Non-Convex Learning via Stochastic Gradient Langevin Dynamics: A Nonasymptotic Analysis," *arXiv preprint arXiv:1702.03849*, 2017.

[46] I. Kuzborskij, N. Cesa-Bianchi, and C. Szepesvári, "Distribution-Dependent Analysis of Gibbs-ERM Principle," *arXiv preprint arXiv:1902.01846*, 2019.

[47] M. Sugiyama, T. Suzuki, and T. Kanamori, *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012.

[48] S. T. Jose and O. Simeone, "Information-Theoretic Bounds on Transfer Generalization Gap Based on Jensen-Shannon Divergence," *arXiv preprint 2010.09484*, 2020.

[49] Y. Polyanskiy and Y. Wu, "Lecture Notes on Information Theory," *Lecture Notes for ECE563 (UIUC) and*, vol. 6, no. 2012-2016, p. 7, 2014.